응용환경 적응을 위한 온톨로지 매핑 방법론에 관한 연구
Adaptive ontology mapping methodology for an application area

안성준[a], 김우주[b], 박상언[c]

[a] *Department of Information Industrial Engineering, College of Engineering, Yonsei University*
134 Shinchon-Dong, Seodaemun-Gu, Seoul, 120-749, Korea
Tel: +82-2-2123-7754, E-mail: sungjun@yonsei.ac.kr

[b] *Department of Information Industrial Engineering, College of Engineering, Yonsei University*
134 Shinchon-Dong, Seodaemun-Gu, Seoul, 120-749, Korea
Tel: +82-2-2123-5716, Fax: +82-2-2260-8824, E-mail: wkim@yonsei.ac.kr

[c] Division of Business Administration, Kyonggi University
San 94-6 Yiui-Dong, Paldal-Gu, Suwon, Kyonggi 442-760, South Korea
Tel : +82-31-249-9459, Fax:+82-31-249-9401 supark@kgu.ac.kr

Abstract

온톨로지 매핑 기술은 시맨틱 웹을 비롯한 여러 분야에서 중요한 기술 중 하나이다. 온톨로지 매핑은 두 개의 온톨로지를 입력으로 받고, 이를 몇 개의 매개변수로 구성된 특정 알고리즘을 이용하여 두 온톨로지 간의 매칭 관계를 알아내고 이를 표현하는 절차를 말한다. 온톨로지 매핑을 이용하여 대용량 온톨로지의 통합이나, 지능화된 통합 검색을 구현할 수 있고, 여러 응용프로그램이 하나의 도메인을 공유하는 등 여러가지 방안으로 사용할 수 있다. 일반적으로 온톨로지 매핑의 성능을 판단하는데 있어서 매핑 결과를 측정하는 방법론의 측정 값을 주로 고려해왔다. 본 연구에서는 매핑을 수행할 때 두 개의 파라미터를 사용하였는데 하나는 알파이고 하나는 Threshold이다. 이것은 매핑의 정확성을 판단하는데 많은 영향을 미친다. 앞서 언급했듯이 매핑결과에 대한 측정값을 중요하게 여기기 때문에 많은 매핑관련 연구에서 알고리즘이 좋은 측정값을 도출할 수 있도록 파라미터를 조절하는 것에 초점을 맞춰왔다. 본 연구에서 측정방법에 따른 높은 측정결과를 지향하는 것이 아닌 온톨로지의 성격과 매핑결과의 사용처에 따라 파라미터를 적절히 변화시켜야 한다는 점에 주목하고, 주어진 환경과 매핑의 사용처에 알맞게 파라미터를 조정하는 방법론을 제안하고자 한다.

Keywords: 시맨틱 웹, 온톨로지 매핑, Semantic Web, Ontology Mapping,

1. 서론

매핑기술은 시맨틱 웹을 비롯한 여러 분야에서 중요한 기술 중 하나이다. 본 연구에서 다루고자 하는 것은 시맨틱 웹 에서의 온톨로지 매핑이다. 온톨로지란 특정 도메인에 대한 내용을 형식적 혹은 공식적인언어로 표현한 것을 말하는데, 어플리케이션은 이 온톨로지를 이용하여 도메인에 대한 정보를 얻거나, 이를 바탕으로 추론을 한다. 온톨로지 매핑은 둘 이상의 온톨로지에 기술되어있는 도메인에 대한 정보들을 비교하여 온톨로지간의 유사성을 측정하여 이 결과를 나타내는 것을 말한다. 온톨로지 매핑을 통하여 공통적인 도메인을 다루고 있지만 이질적인 구조로 구성되어 있는 온톨로지들을 통합하므로 자료의 통합관리를 이룬다거나, 에이전트가 매핑결과를 이용하여 지능적인 통합검색을 수행한다거나, 다른 시스템간의 상호운용성을 이룰 수 있는 등 여러가지 방안으로 사용할 수 있다. 시맨틱 웹의 기본적인 연구 방향이 시스템간 받아들인 정보의 의미를 파악하여 처리하는 것의 자동화를 지향한다는 것을 감안할 때 앞서 언급했듯이 시맨틱 웹에서 매핑은 중요한 기술로서 생각할 수 있다.

온톨로지 매핑의 결과를 평가하기 위한 방법으로 여러 방법이 고안되어 왔다. 매핑결과를 평가하기 위한 대표적인 척도로는 정확도(Precision)과 재현율(Recall)이 있다.[4] 정확도는 매핑 결과 중 맞은 결과의 정도를 뜻하고, 재현율은 연재 매핑을 해야할 것들 중 실제로 매핑에 성공한 정도를 뜻한다. 여러 매핑결과 평가 방법론에서 사용하는 성능측정 방법 중 대표적으로 F-Measure를 꼽을 수 있다.[1] 이 측정방법에서는 정확도와 재현율을 가공하여 온톨로지 매핑의 결과를 평가한다. F-Measure의 경우는 다음의 식으로 구할 수 있다.

$$F = \frac{2 \cdot \mathrm{Pr}ecision \cdot \mathrm{Re}call}{\mathrm{Pr}ecision \cdot \mathrm{Re}call} \quad (1)$$

일반적인 연구에서 온톨로지매핑 시 매핑의 결과에 대한 높은 측정 결과에 집중해왔다 그렇지만 이러한

접근 방법은 상황에 따라서 제대로된 매핑결과를 얻지 못하는 경우가 있을 수 있다. 예를 들어 전자상거래 환경에서 상이한 쇼핑 사이트의 상품들을 통합하여 관리 또는 검색하기 위한 목적으로 상품 카테고리에 대한 온톨로지 매핑이 이루어질경우, 매핑의 정확성을 지나치게 강조하다보면 매핑의 재현율에 영향을 받아 매핑이 이루어져야할 대상도 매핑에서 제외될 수 있는 경우가 그것이다. 때문에 본 연구에서는 온톨로지 매핑시 높은 측정 수치를 목표로 매핑을 진행시키는 것이 아니라 온톨로지 매핑을 실행하는 환경에 따라 적합한 매핑을 실행해야 한다는 점에 초점을 맞추어 연구를 진행하고자하고, 이에 대한 방법을 정확도와 재현율의 비율을 변화시키는 방식으로 접근하였다. 일반적으로 정확도와 재현율은 반비례 관계에 있다. 즉 정확한 매핑결과를 위하여 정확도를 높일 경우, 재현율은 떨어진다 반대로 많은 결과를 얻기 위해 재현율을 높일 경우 정확도가 떨어지게 된다. 앞서 예를 든 전자 상거래 온톨로지의 경우 정확도를 조금 희생하는 대신 재현율을 높이는 방향으로 매핑을 적용하는 것이 좋다고 볼 수 있다. 반대로 전자상거래의 경우가 아니라 온톨로지 매핑을 통하여 정확한 통합 자료를 원하는 경우는 재현율보다는 정확도를 더욱 고려하여 매핑을 수행하는 것이 좋다고 볼 수 있다.

본 연구에서는 정확도와 재현율 중 특정 척도를 중요시 여겨 온톨로지 매핑을 수행할 때 다른 척도의 손실률을 최소로하는 방안의 제시를 위해 온톨로지 매핑을 반복하는 실험을 통해 정확도와 재현율의 관계에 대해서 알아보고자 한다. 실험에 사용할 온톨로지 매핑 방법론은 김우주외[5]가 제안한 온톨로지 매핑 방법론을 이용하여 진행할 것이다. 본 논문의 구성은 다음과 같다 2장에서 온톨로지 매핑에 대한 관련 연구를 설명하고, 3장에서 실험에 사용된 온톨로지 매핑 방법론과 온톨로지 매핑 알고리즘에 대하여 설명할 것이다. 4장에서는 실험의 결과를 분석하고 마지막 5장에서 결론과 향후 연구에 대하여 다룰 것이다.

2. 관련연구

온톨로지 매핑과 관련된 연구분야는 다음과 같이 Mapping Discovery, Declarative formal representations of mappings, Reasoning with mappings로 나눌 수 있다. [2] Mapping Discovery는 주어진 두 온톨로지에 대해 서로 유사성을 보이는 클래스나 속성을 찾는 방법을 연구하는 분야이고, Declarative formal representation of mappings는 추론(Reasoning)을 위해 두 온토롤지간의 유사성을 어떤 식으로 표현(Representation)할 것인가에 대한 연구이다. Reasoning with mappings은 매핑이 이루어진 후 그것을 이용하는 것을 뜻한다. 이 본 연구에서 주로 다룰 내용은 온톨로지간의 유사성을 알아보는 Mapping Discovery에 속한다고 볼 수 있다. 지금까지 많은 수의 온톨로지 매핑 방법론에 대한 연구가 이루어져 왔다. 매핑을하는 방법에따라 크게 공통적인 온톨로지를 사용한 후 각 사용처에 따라 온톨로지를 확장하여 추후 매핑을 실시할 때 보다 효율적인 매핑을 추구하는 방법인 공통온톨로지를 사용하는 방법과, 공통적인 온톨로지를 사용할 수 없는 상황의 경우 경험적, 혹은 학습을 통한 정보를 이용해 매핑을 하는 휴리스틱적, 기계학습적 방법으로 나눌 수 있다. 매핑은 대상이 되는 두 온톨로지에 존재하는 여러 클래스들에 대해서 이루어지는데, 매핑이 이루어지는 범위에 따라서 온톨로지의 클래스와 서브클래스들만 매핑을 수행하는 것과 해당 클래스들의 인스턴스들까지 매핑의 대상으로 삼는 방법이 존재한다. 매핑의 정확성을 본다면, 인스턴스까지 고려하는 것이 더 정확하다고 볼 수 있다. 그렇지만 이 방법의 경우 매핑 시간이 온톨로지에 포함되어있는 인스턴스의 양에 따라서 기하급수적으로 늘어날 수 있다는 단점 역시 존재한다. 이상과 같은 이유로 본 연구에서는 온톨로지 매핑에 인스턴스를 사용하지 않고, 온톨로지 스키마만으로 매핑을 수행하는 스키마 기반 매칭(Schema-Based Matching)[3]을 이용한 매핑 방법론에 초점을 맞추고자 한다.

스키마 기반 매칭 에선 입력 받은 온톨로지를 매핑하는 대상에 따라 Element-level과 Structure-level로 나눌 수 있다. Element-level은 온톨로지내의 클래스의 이름만을 분석하는 것이고, Structure-level 매핑은 온톨로지내의 클래스들의 서브, 슈퍼클래스 관계를 분석하여 매핑을 수행하는 것이다. Element-level과 Structure-level은 매핑을 하기 위해 입력된 정보를 해석하는 방식에 따라 Syntactic, External, Semantic으로 다시 나눌 수 있는데, Syntactic은 단순한 String비교나, 단어비교 등으로 클래스간의 유사성을 판별하는 것이고 External은 클래스간의 유사성을 판단할 때 WordNet과 같은 외부 자원을 참고하여 클래스간의 유사성을 판단하는 것이고, Semantic은 매핑을 하기 위해 입력된 온톨로지를 해석할 때 formal semantic을 이용하는 것이다. 앞서 언급한 Element-level, Structure-level과 Syntactic, External, Semantic과 같은 정보 해석방법을 사용함에 따라 각기 String-based, Linguistic based, Alignment reuse, Graph based, Taxonomy based 등등으로 나눌 수 있다. 최근의 매핑 방법론은 위에서 언급한 String-based, Linguistic based, Alignment reuse 등등 여러가지 방법론을 혼합하여 사용하고 있다. [3] 우리의 매핑 방법은 소스 온톨로지와. 목표 온톨로지에 있는 클래스들의 서브, 슈퍼 클래스구조를 이루는 경로를 매핑하는 방법으로 기본적으로 클래스들의 이름을 비교하여 매핑을 수행한다 매핑 수행시 동의어, 유사어에로 기술된 클래스이름에 대비하여 WordNet을 사용하여 클래스이름의 의미를 확장하여 이를 가지고 매핑을 수행한다.[9] 이러한 방법은 앞서 언급한 매핑

방법론 분류 절차에 의하면 Structure-level의 taxonomy-based이나, repository of structures 와 유사한 방법이다.

3. 온톨로지 매핑 방법론

3.1 온톨로지 매핑 개요

일반적으로 이루어지고 있는 매핑 프로세스는 다음과 같다. 매핑은 두 개 이상의 온톨로지
를 입력으로 받아들여 이를 특정 알고리즘을 이용하여 각 온톨로지에 있는 클래스들의 유사성을 평가한다. 매핑 방법론에 따라서 파라미터가 유사성을 판단하는데 이용될 수도 있다. 매핑을 하려는 두 개의 온톨로지 중 매핑을할 온톨로지를 소스 온톨로지라고 하고 매핑의 대상이 되는 온톨로지를 목표 온톨로지라고 한다. 유사성을 판단하는 파라미터에는 여러 가지가 있는데 대표적으로는 가중치(weight)와, Threshold, 개념별 분류 어휘집(Thesaurus)등이 존재한다. 예를 들어 알고리즘을 통해 각 온톨로지에 있는 a와 a'의 유사정도가 0.65라고 하고 이 방법론이 사용하고 있는 파라미터 중 Threshold역할을 하는 파라미터의 값이 0.63이고, 이것보다 높아야 유사성이 있다고 판단한다라고 하면, 두 클래스는 같은 것으로 결정되는 것이다.[3]
위와 같이 특정 프로세스에 의해서 수행된 매핑된 결과는 매핑 평가방법론에 의해서 평가할 수 있다. 평가 방법론에 따라서 다르지만, 앞서 언급했듯이 많은 매핑 측정 방법에서 정확도와 재현율은 측정시 중요한 척도로서 작용한다. 이는 위에 소개한 일반적인 매핑 프로세스내의 파라미터를 이용하여 조정할 수 있다. 때문에 파라미터의 조정은 매핑 방법론의 방향을 결정하는 중요한 요소 중 하나이다. 일반적으로 많은 연구에서는 매핑 방법 측정 결과가 높게 나올 수 있도록 파라미터를 조정하여 정확도와 재현율을 설정하지만 본 연구에서 적용하는 매핑 방법론에선 각 환경에 맞는 파라미터 조정을 통해 온톨로지 매핑을 이루고자 한다. 환경에 적합한 온톨로지 매핑를 앞서 언급했던 전자상거래 환경과 F-measure를 사용해서 예를 들자면 다음과 같다. 두 쇼핑몰의 온톨로지를 매핑하여 통합검색을 이루고자 온톨로지 매핑을 하는데 각기 다른 두 가지의 온톨로지 매핑 방법론이 사용되었다고 가정하고 매핑을 해야하는 대상 10개 중 매핑 방법론 A에 의한 결과 15개 중 5개가 맞았고 매핑 방법론 B에 의한 결과 4개 중 3개가 맞았다고 할 때, 식(1)에 의한 온톨로지 A와 B에 대한 F-measure값은 각각 0.4, 약 0.428로 나타난다. 이러한 결과가 나왔을 때 F-measure에 의한 결과는 온톨로지B가 좋지만, 매핑의 목적에는 온톨로지A가 더 부합한다고 볼 수 있다. 앞서 제시한 예는 정확도를 포기하고 재현율을 높이는 경우만을 살펴봤지만, 상황에 따라 물론 반대의 경우도 있을 것이다. 그럴 때 역시 우리는 파라미터조절로 인해 환경에 최적화 된 온톨로지 매핑을 수행할 수 있다고 생각한다. 이처럼 측정의 결과가 우수한 매핑의 사용이 무조건 좋은 것이 아니라, 온톨로지 매핑의 사용처에 따라 적절한 파라미터 조정은 온톨로지 매핑의 결과를 효과적으로 사용할 수 있게 해준다. 본 논문의 3.2절에서 매핑 알고리즘과, 정확도와 재현율이 매핑 알고리즘에 어떤 식으로 적용되는지에 대해서 다루고, 4장에서는 반복실험을 통한 나타난 정확도와 재현율의 관계에 대해서 다룰 것이다.

3.2 매핑 프로세스 및 알고리즘

본 논문에서 제안하고자하는 매핑 프로세스는 크게 세 단계로 나뉜다.
첫 번째 단계는 소스 온톨로지내 매핑을 할 경로에 대한 정확한 의미 파악이다. 이 단계를 거치는 이유는 소스, 목표 온톨로지에 동일한 의미로 존재하고 있는 클래스일지라도 서로 다른 명칭으로 사용되었을 경우나 동일한 명칭으로 사용되었으나 의미가 다를 경우를 정확히 구별하여 매핑을 수행하기 위함이다. 예를 들어 Car와 Vehicle은 같은 의미의 클래스이지만, 다른 명칭으로 사용된 것이고, 공책 Notebook과 휴대용 컴퓨터 Notebook은 다른 의미를 같은 명칭으로 사용한 것이다.
두 번째 단계는 선정된 확장의미의 어휘로 매핑을 수행하는 단계이다. 앞서 언급했듯이 매핑은 소스와 목표 온톨로지의 경로에 대해 이후어진다. 매핑을 하는 두 경로를 대상으로 동일 단어의 출현 빈도와, 출현 순서를 비교한다.
세 번째 단계는 매핑 값 산출 및 평가 단계이다. 두 번째 단계에 대한 결과에 두 개의 파라미터를 이용하여 매핑 값을 산출하고 유사성을 평가한다. 이 단계를 거치면서 유사하다는 판정이 나오게 된다.
이제 본 연구를 위해 매핑을 수행했던 것을 예로 들어 앞서 언급한 3단계를 알아보겠다. 매핑할 소스 온톨로지는 ODP 구조를 온톨로지로 만든 것이고 목표 온톨로지는 Amazon의 상품분류체계를 온톨로지화 시킨 것을 사용하였다.

3.2.1. 어휘의 의미 파악 및 확장

이 단계는 알고리즘의 첫 번째 단계이다. 이 단계를 통하여 매핑할 소스 온톨로지내 클래스에 대한 정확한 의미 파악이 이루어진다. 이 단계는 동일한 의미로 존재하고 있는 클래스들에 대한 매핑을 수행하기 위해서나, 동음이의어로 사용되어 매핑시 적합한 대상이 아닌 클래스를 매핑하는 것을 방지하기 위함이다. 이 단계는 영어 어휘 데이터베이스 WordNet을 이용하여 이루어진다. WordNet은 단어에 대한 Hyponyms, Hypernyms, Synoyms, Attribute 등을 저장한 DataBase이다.[10] 매핑대상이될 소스 온톨로지에 있는 클래스의

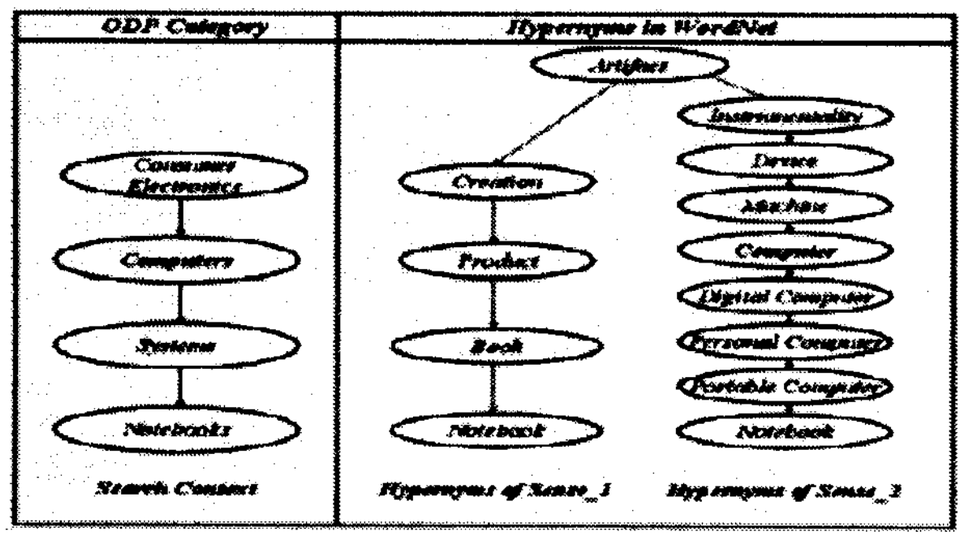
이름에 대한 동의어를 WordNet을 이용하여 찾은 후 동의어 집합을 소스 온톨로지의 클래스가 속하여있는 경로와 비교하는 과정을 통해서 의미를 확장한다. 이러한 일련의 과정을 *pathproximity*라 하고, *pathproximity*는 *CS*, *hypernymprozimity*로 이루어진다. 이를 식으로 표현하면 다음과 같다.

$$CS(x,p) = \{h \mid h \in synsets(x) and h \in hypernyms(p)\}$$
$$where\_x\_is\_an\_uppercategpry\_of\_the\_producthierarchy \quad (2)$$

위 식에서 x는 매핑하려는 클래스의 상위 혹은 하위 클래스의 이름이고, p는 매핑하려는 클래스 이름을 WordNet을 이용하여 구한 sense들을 의미한다.
예를 들어 매핑하고자 하는 온톨로지 클래스의 이름이 notebooks 이고, 소스 온톨로지의 경로가 Consumer Electronics > Computer > Systems > Notebooks 라고 한다면 Notebook의 sense들은 p에 들어가게 되고, 경로명 Consumer Electronics, Computer, Systems 는 x에 들어가게 된다.
비교는 소스 온톨로지의 경로 개수만큼 이루어지고 클래스이름 하나와 sense 전체를 대상으로 이루어진다. 만약 sense가 하나 이상이라면, sense의 수만큼 비교를 하게 된다. CS를 이용하여 *hypernymproximity*를 구할 수 있는데 그 식은 다음과 같다.

$$hypernymproximity(x,p) = \left\{ \frac{1}{Min\_dis(cs(x,p), base)} \right\} \quad (3)$$

식에서 알 수 있듯이 hypernymproximity는 x가 존재하는 p 중 sense의 최하위 노드와 x와 일치하는 노드사이의 거리이다. 노드사이의 거리를 구할 때 최소값의 역수를 취하는 이유는 일치하는 단어가 중복해서 나타날 때를 대비해서이다.
CS와 *hypernymproximity*를 이용하여 *Pathproximity*를 구할 수 있는데, 이는 다음과 같은 식으로 나타낸다. 다음은 *hypernymproximity*과 *Pathproximity*의 예이다.



$$hyperproximity(Systems, path2) = 0$$
$$hyperproximity(Computers, path2) = 1/4$$
$$hyperproximity(Consumer\_Electronics, path2) = 0$$

※ *Sense1*에 대한 *hyperproximity* 값은 모두 *0*.

$$pathPtoximity(p) = \frac{x \in upper\_categories(base)}{n}$$

$$pathproximity(sense\_2) = \left(\frac{0 + \frac{1}{4} + 0}{4} \right) = 0.0625 \quad (4)$$

*Pathproximity*를 사용하므로 Notebook의 sense 중 두 번째 sense가 선택 되었다 때문에 WordNet에서 두 번째 sense의 의미인 {Notebook, Notebook Computer}의 의미로 사용된 목표 온톨로지의 경로가 매핑대상으로 선정되었다.이때 같은 클래스구조를 포함하고 있는 경로가 여러 개 있을 수 있다. 이러한 경우에는, 다른 것보다 범위가 좁은 경로들은 중복이 발생할 수 있기 때문에 이를 제외하고 매핑을 수행한다.

3.2.2. 온톨로지 내 경로의 유사성 판단

이 단계는 매핑하고자 하는 소스와 목표 온톨로지의 경로 들에 대해 동일한 문자열이 얼마만큼, 또 동일 순서로 위치했는가를 비교해보는 것이다. 즉 소스 온톨로지의 경로구조와 가장 비슷한 구조를 찾는 단계라고 할 수 있다. 이 단계에서는 *Co-Occurrence*와 *Order Consistency* 를 이용하여 유사한 경로를 찾는다.
*Co-Occurrence*는 *TermMatch*, *NodeMatch*, *MaxSim*으로 이루어지며, 그 설명은 다음과 같다.
*TermMatch*는 주어진 두 개의 term에 대하여 문자열 비교를 하여 문자열간 중복여부를 체크 한다. 다음은 *TermMatch*의 수식이다.

$$TermMatch(term_1, term_2)$$
$$= \begin{cases} \frac{strlen(term_1)}{strlen(term_2)} & if\ term1\ is\ substring\ of\ term2 \\ \frac{strlen(term_2)}{strlen(term_1)} & if\ term2\ is\ substring\ of\ term1 \\ -1 & otherwise \end{cases} \quad (5)$$

strlen 은 term의 길이를 뜻한다. 예를 들어 term1, term2에 각각 Electronics와 Consumer Electronics가 들어가면 *TermMatch*의 결과는 11/19라는 결과가 나오게 된다.

*NodeMatch*는 소스 온톨로지에서 매핑 대상이 되는 경로의 클래스 이름과, 목표 온톨로지에서 매핑대상이 되는 경로의 클래스 이름간에 *TermMatch*를 하는 것이다. 정확한 매핑을 위해 소스 온톨로지에 있는 클래스 이름은 WordNet을 이용한 Sense를 이용하며 이 중 가장 큰 값을 취한다. 이를 식으로 나타내면 다음과 같다.

$$NodeMatch(cat, node) = \max_{cterm \in TermSet^{ST}(cat)} TermMatch(cterm, nterm) \quad (6)$$

MaxSim은 소스 온톨로지 경로 중 하나의 클래스 이름과 목표 온톨로지 경로의 모든 클래스 이름과 TermMatch를 하는 것이다. MaxSim의 수식은 다음과 같다.

$$MaxSim = (cterm, cpath) = \max_{pterm \in cpath} NodeMatch(cterm, pterm) \quad (7)$$

Co-Occurrence는 MaxSim을 이용하여 구할 수 있다. 다음은 Co-Occurrence의 수식이다.

$$Co-Occurrence(source, target) = \left(\frac{\sum_{n(source)} member(n(source), source)}{n(target)} \right) \left(\frac{\sum_{n(target)} member(n(target), target)}{n(source)} \right) \quad (8)$$

*Order Consistency*는 매핑하려는 두 경로에 유사한 문자열로 이루어진 클래스가 유사한 순서로 있는지 체크하는 것이다. *Order Consistency*는 *Common, Prelset, Consistent* 로 이루어진다.

*Common*은 소스 온톨로지의 경로와 목표 온톨로지의 경로에 유사한 문자열이 포함되어 있는지를 확인하는 것으로 *Order Consistency*의 *TermMatch*와 일부 유사하다. 유사 여부 판별은 별도의 Threshold를 주어 시행한다. 이를 바탕으로 *Prelset*을 만든다. *Prelset*은 *Precedence Relationship*을 뜻하며 *Common*의 결과와 경로1, 혹은 경로2에 대한 순서를 binary relationship으로 만든 것이다.

*Common*과 *Prelset*의 간단한 예는 다음과 같다.

Path1 (A, B, C, D)
Path2(A', B', C', E) (A와 A'는 유사 문자열.)

Common (path1, path2) = {[A, A'], [B, B'], [C, C']} (9)

Prelset (common(), path2) = ([A'B'], [B', C'], [A', C']) (10)

*Prelset*의 결과는 *Consistent*를 구하는데 사용된다. *Prelset*으로 구한 결과들 중 하나와 경로1 혹은 경로2의 순서가 나머지 하나에서도 마찬가지로 지켜지는지 알아보는 과정으로 결과는 0 또는 1로 나온다. *Order Consistency*는 *prelset*의 결과만큼 *Consistent*를 보는 것이다. 다음은 *Consistent*와 *Order Consistency*의 식이다.

$$Consistent((t_p, t_s), path_1) = \begin{cases} 1 & \text{If } t_p \text{ precedes } t_s \text{ is Category Path} \\ 0 & \text{otherwise} \end{cases} \quad (11)$$

Where $(t_p, t_s) \in prelset(common(\text{sc. Path1}), \text{sc})$ sc is search context.

$$order\_consistency(Catergoty\_path, Search\_Context) = \frac{\sum_{pr \in prelset(ccts, Category\_path)} consistent(pr, Catergory\_path)}{\binom{ns(ccts)}{2}} \quad (12)$$

Where ccts = *common*(Category Path, Search Context)

3.2.3. 최종 유사성 판단

Co-Occurrence, *Order Consistency* 을 통한 것은 다음을 이용하여 유사성을 판별한다.

If α *(Co-Occurrence) + (1-*α*) Order Consistency* $\geq$ *t (Threshold)*
Then, path1 is similar to path2

유사성을 판단하는데 있어서 앞서 언급한 파라미터가 사용된다. 위 식을 보면 α, t 두 개의 파라미터를 사용하고 있는 것을 볼 수 있다. α는 *Co-occurrence*와 *Order consistency*의 상대적 가중치이고, t는 Threshold를 나타내는 파라미터이다. α를 높일 경우, *Co-Occurrence*를 비중 있게 다루고 상대적으로 *Order Consistency*를 중요하게 고려하지 않는다. α를 낮출 경우 반대의 경우가 나타나게 된다. α와 *Co-Occurrence*, *Order Consistency*를 이용하여 구한 값은 t를 이용하여 유사성이 판정되는데 t를 높이면 매핑의 정확도를 중요시 여기는 것이고, t를 낮추는 것

은 일부 유사해도 두 경로가 같다고 판정할 수 있도록 하는 것이다. 이 α와 t의 비율을 조정하는 방법으로 정확도와 재현율의 비율을 설정할 수 있다. 4장에서는 α와 t값을 변화시켜가면서 정확도와 재현율의 변화와 둘 사이의 관계에 대해서 소개할 것이다.

4. 실험 및 결과

4.1 실험 개요

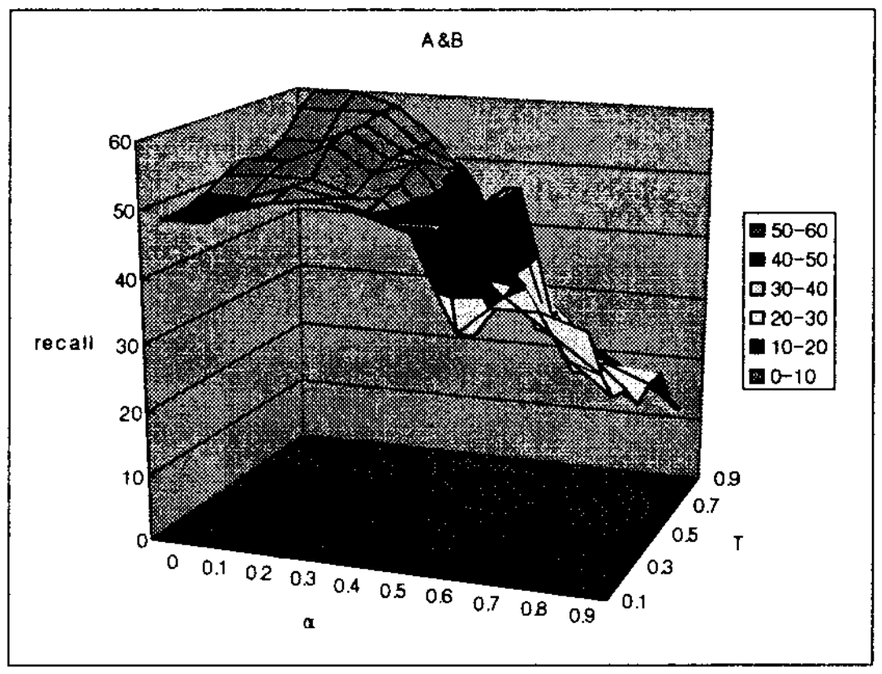
본 연구에서는 온톨로지 매핑을 수행할 때, α와 t 두 개의 파라미터를 변화시키므로 이에 따른 정확도와 재현율의 관계를 알아보기 위해서 실험을 수행하였다. 여기서 알아본 정확도와 재현율의 관계를 이용하면, 정확도와 재현율의 최적 비율을 결정하여 특정 환경에 맞는 온톨로지의 매핑 방법의 방향을 설정하는데 도움을 줄 것이라 생각된다.

본 연구의 실험에 사용된 온톨로지는 Amazon.com과 Buy.com, ODP 의 상품분류체계를 온톨로지화 시킨 것으로 이루어 졌다. 실험을 위해 매핑할 온톨로지 경로 데이터를 선정하였는데, 위한 데이터 선정 방법으로는 위 3가지 온톨로지 전체의 데이터를 사용한 것이 아니라 카테고리 정보의 일부분을 샘플 데이터화 시켜서 사용하였다 실제로 온톨로지를 매핑시킬 때는 데이터 수집 방법에 대한 다음 두 가지 경우를 생각해 볼 수 있다. 첫 번째로 매핑하고자 하는 두 온톨로지가 굉장히 클 경우 소스 온톨로지의 일부분만 매핑 대상으로 취해, 이를 목표 온톨로지와 매핑 시켜보는 것이다. 이러한 방식의 매핑의 경우 두 온톨로지의 구조가 어느 정도 유사성을 가지고 있어 소스 온톨로지의 일부분만을 이용한 매핑 결과를 전체에 적용하여도 매핑의 결과가 크게 달라지지 않는다고 판단될 때 사용할 수 있다. 두 번째 방법은 두 온톨로지 전체에 대한 매핑을 수행하고 매핑을 수행한 한 쌍의 소스, 목표 온톨로지에 대한 최적 파라미터 값의 결과를 별도의 작업 없이 다른 온톨로지 매핑에서도 이용하는 방법이다. 이러한 방법은 각 사이트가 가지고 있는 온톨로지의 표현방식은 상이하지만 클래스의 본질적인 의미가 같다고 판단될 때 사용할 수 있다.

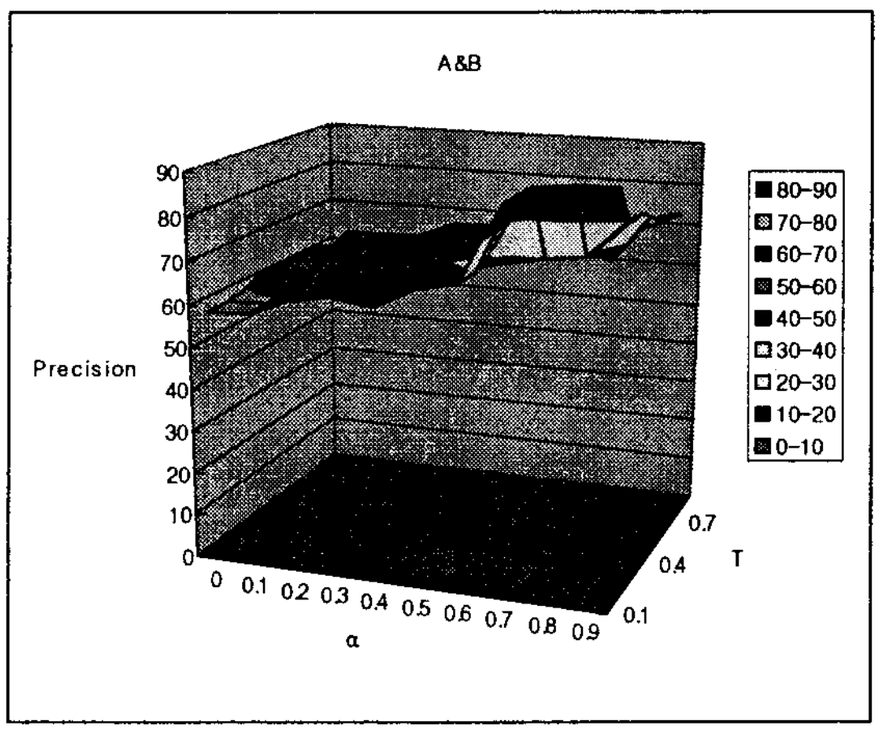
실험은 3개의 온톨로지 중 소스, 목표 온톨로지를 바꾸어 가며 총 3쌍의 온톨로지에 대해 6번의 실험을 수행하였다.

4.2. 실험 결과 및 분석

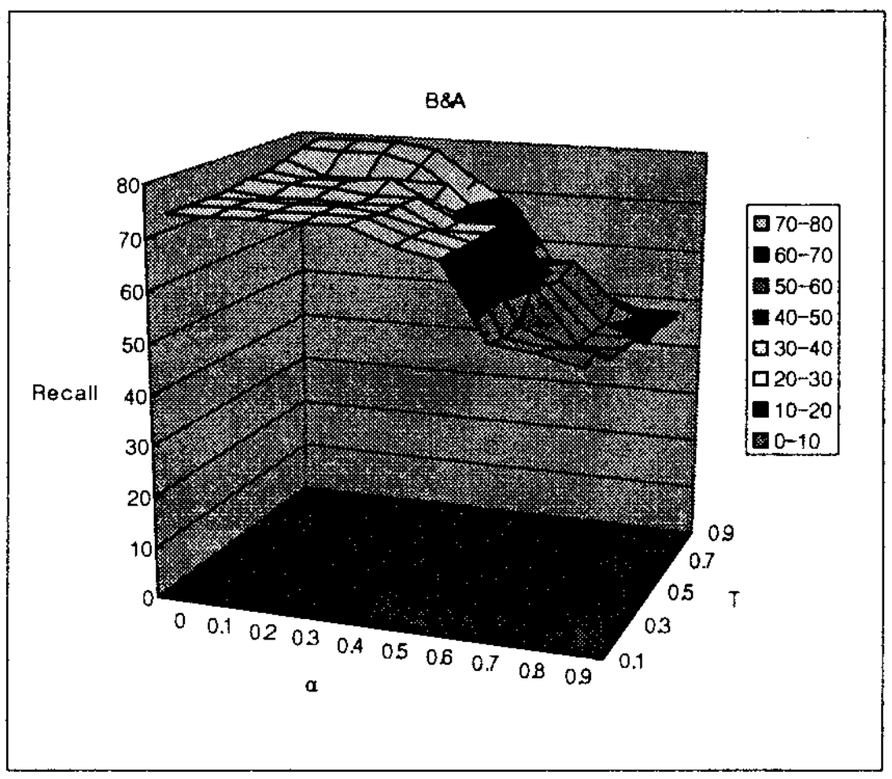
6번의 온톨로지 매핑을 하는 동안 α와 t를 0부터 1까지 10 구간으로 나누어 변화시키면서 각 파라미터의 변화에 대한 정확도와 재현율을 구하였다. 6번의 실험을 통하여 한번의 실험당 정확도, 재현율에 대한 그래프가 하나씩 나와야 하기 때문에 총 12개의 그래프가 결과로 나오지만 실험결과 중 신뢰하기 힘들다고 판단되는 결과들이 있어 이를 제외한 결과의 일부만으로 분석을 수행하였다. 다음은 온톨로지 매핑 결과의 일부이다.



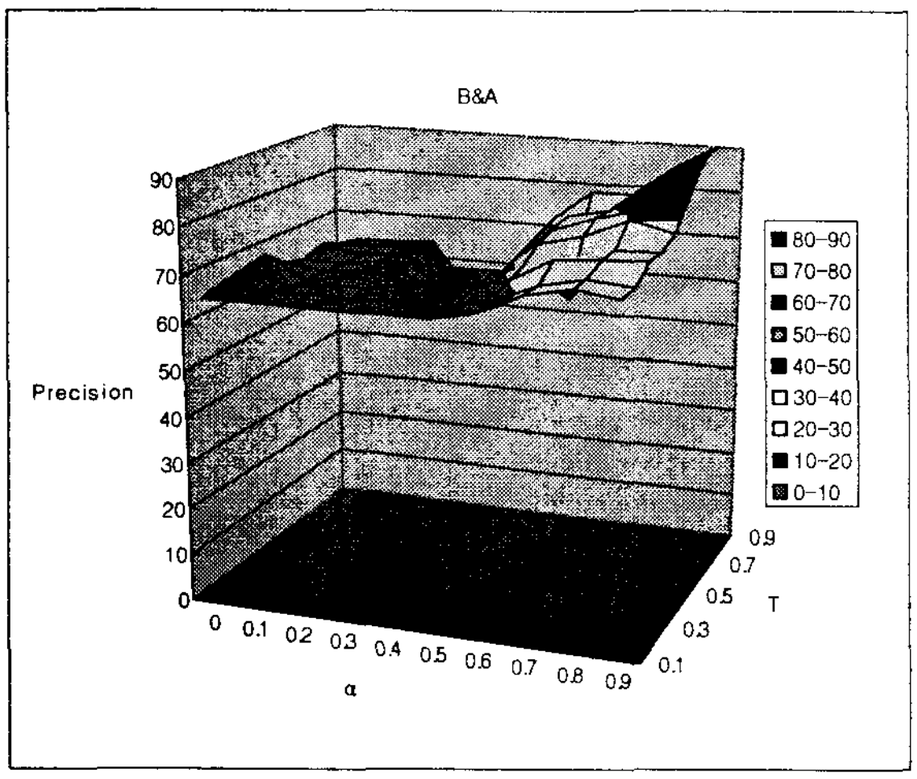
Amazon Buy.com Recall



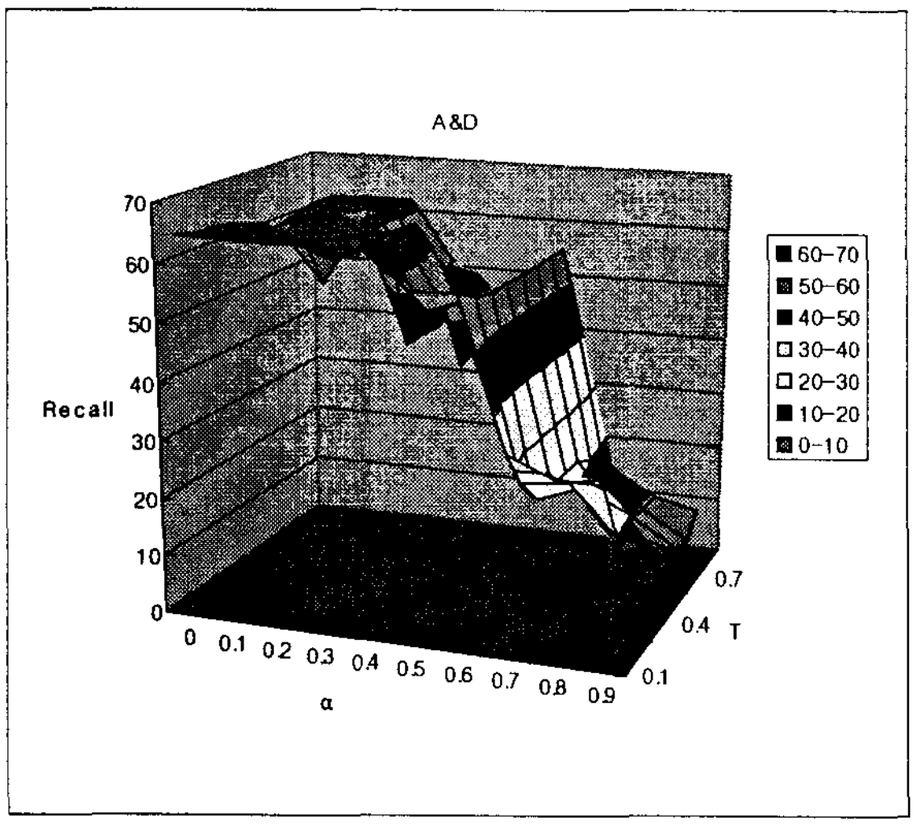
Amazon Buy.com Precision



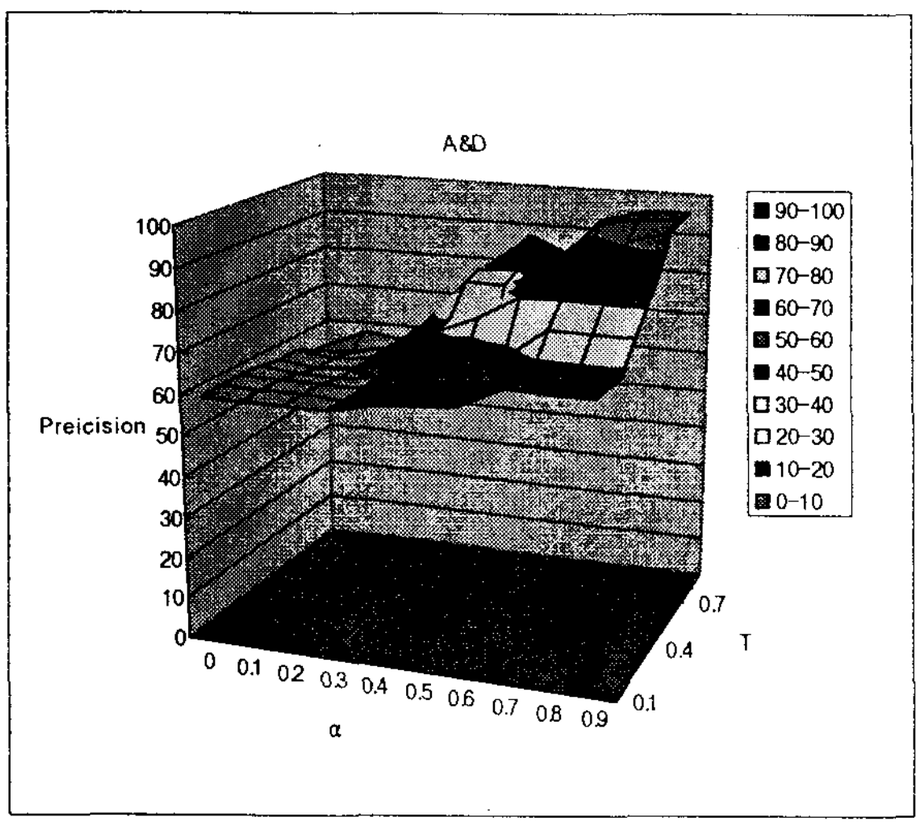
Buy.com Amazon Recall



Buy.com Amazon Precision



Amazon Dmoz Recall



Amazon Dmoz Precision

그래프를 보면 동일한 t값이 주어진 상태에서 α값이 적을수록 재현율이 높게 나오는 것을 발견할 수 있다. 뿐만 아니라 몇몇의 그래프에서는 α값의 변화에 따라 재현율이 급격하게 줄어드는 지점이 있음을 볼 수 있다. 이는 α에 큰 값을 줄수록 *Co-Occurrence*의 비중이 높아져 매핑을 까다롭게 적용하여 매핑으로 얻을 수 있는 결과가 적어짐을 의미하며 특정 값 이상의 α에 대해서는 재현율이 아주 큰 영향을 받는 것으로 생각할 수 있다. 반대로 동일 t에서 α값을 높일수록 정확도가 높게 나타남을 발견할 수 있으나 α값을 줄여 재현율을 얻는 만큼의 급격한 변화를 보이지는 않았다. 이것은 α가 미치는 영향이 정확도보다는 재현율에 더 크다고 볼 수 있다.

t의 경우 동일 α값 에서 재현율에 주는 영향이 α보다 적은 것으로 나타났다. 이는 α가 t보다는 재현율에 더 큰 영향을 주는 것으로 생각할 수 있다. 각 그래프의 정확도를 보면 각 그래프 비슷한 범위의 α값과 t값의 지점에서 높게 나옴을 볼 수 있다. 그렇지만 모든 그래프의 동일 지점에 대한 재현율을 살펴보면 모두 적은 재현율을 보이는 것으로 보아, 해당 환경에서 재현율과 정확도는 반비례 관계에 있음을 볼 수 있다. 전반적으로 α의 값이 재현율이나 정확도에 많은 영향을 끼치는 것으로 나타났는데, 이는 상품 카테고리를 클래스화 한 온톨로지라는 특성상 표현방법은 다를지 몰라도 서로 유사한 상품 분류 기준 하에서 상품 카테고리를 생성했기 때문에 모든 매핑 결과에서 *Order-Consistency*가 높게 나와, 결과적으로 *Co-Occurence*에 비해서 *Order-Consistency*가 매핑결과에 영향을 적게 미친 결과로도 생각할 수 있다.

5. 결론

본 연구에서 온톨로지 매핑 결과의 측정 방법에 적합한 매핑 방법론이 아닌 온톨로지 매핑이 이루어지는 환경에 적합한 온톨로지 매핑 방법론을 제안하였고, 이를 정확도와 재현율을 변화시키는 방법으로 접근하고자 했다. 정확도와 재현율의 변화로 인해 특정환경에 적합한 온톨로지 매핑을 이루기 위한 시도 중 하나로 온톨로지 매핑에 사용되는 알고리즘 중 정확도와 재현율에 영향을 미치는 파라미터를 조정하여 정확도와 재현율의 관계를 실험을 통하여 확인하였다. 실험을 통해 나타난 정확도와 재현율의 관계는 환경에 맞는 온톨로지 매핑을 위한 최적의 정확도, 재현율 비를 구하는데 이용될 수 있을 것이라 생각된다.

그렇지만 실험 결과를 통한 결과에 대한 실제적인 검증이 어렵고, 자료의 신뢰성이 부족한 몇몇 실험 결과 때문에 실험에서 나타난 특정 α에서 재현율이 급격히 줄어드는 것에 대한 자세한 관찰이나 여러 가지 상황에서의 다양한 분석을 하지 못한 점, *Co-Occurrence*와 *Order-Consistency*가 매핑의 결과에 미치는 영향의 정도들을 충분히 밝히지 못한 점

등은 본 연구의 한계점으로 정리할 수 있다. 추후 정확도와 재현율의 변화에 따른 적용한 온톨로지 매핑결과에 대한 실제적인 검증을 통하여 정확도와 재현율의 조정이 특정환경에 적합한 온톨로지 매핑을 이루는데 기여한다는 것을 증명하는 연구가 필요하다고 판단된다. 뿐만 아니라 전자상거래 환경의 온톨로지만이 아닌, 실제 사용하고 있는 온톨로지를 대상으로 한 매핑을 통하여 연구의 보편성을 확보해야 할 것이라 생각한다.

References

[1] Amazon.com
http://www.amazon.com

[2] Fausto Giunchiglia, Pavel Shvaiko, and Mikalai Yatskevich(2005): Semantic Schema Matching," *proceedings of the 13th International Conference on Cooperative Information Systems"*

[3] Natalya F.Noy (2004) : Semantic Intergration : A Survey Of Ontology-Based Approaches, *"SIGMOD Record"*

[4] OWL Seb Ontology Language Reference avail. from http://www.w3.org/TR/owl-ref/

[5] OWL Web Ontology Language Guide avail. form http://www.w3.org/TR/owl-guide/

[6] RDF Primer avail. from
http://www.w3.org/TR/rdf-primer/

[7] Pavel Shvaiko and Jerome Euzenat (2005) : A survey of Schema-based Matching Approaches, *"Journal on Data Semantics (JoDS), IV*

[8] Sangun Park, Wooju Kim, Sunghawn Lee, and Siri Bang(2006) : An Ontology Mapping Algotitm between Hetegogeneous Product Classification Taxonomies."*proceedings of the iiwas"*.

[9] Wooju Kim, Dae Woo Choi, and Sangun Park(2005) : Agent Based Intelligent Search Framework for Product Information Using Ontology Mapping, "*Journal of Intelligent Information Systems*".

[10] 김우주, 방시리, 박상언 (2006) : An Optimized Methodology of Ontology Driven Mapping for Product Search, "*지능 정보시스템 학회 논문집*"

[11] 최남혁 (2006) : 이질적인 쇼핑몰 환경을 위한 온톨로지 기반 상품 매핑 방법론, 연세대학교, 석사학위 논문

[12] 최대우 (2004): 에이전트와 쇼핑몰을 위한 의미 웹 서비스 기반 지능형 상품 정보 검색 프레임웍, 전북대학교, 박사학위논문