최적 온톨로지 매핑 방법론에 관한 연구.
Study for optimal ontology mapping methodology

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

시멘틱 웹에서의 온톨로지는 특정 영역의 설명을 위해 공유할 개념화된 명세란 정의로 널리 알려져 있으며, 시멘틱웹의 중요한 요소기술이다. 온톨로지는 특정 도메인에 대한 정보를 기술하는데, 이러한 온톨로지를 매핑할 경우 많은 양의 정보를 통합관리하거나, 상호호환성을 이룰 수 있다. 여러 온톨로지 매핑 방법론의 성능을 평가하는 수단 중 f-measure란 것이 있는다. f-measure의 값은 정확도(precision)과 응답률(recall)에 의해서 결정된다. 정확도와 응답률이 변화함에 따라 f-measure 값도 자연히 변하기 때문에, 높은 f-measure 값을 구하기 위해서는 정확도와 응답률의 밸런스를 조정할 필요가 있다. 본 논문에서는 높은 f-measure값을 얻을 수 있는 정확도와 재현률을 구하는 방법을 휴리스틱적 방법을 통하여 알아보고자 한다.

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

1. 서론

온톨로지는 "특정 영역의 설명을 위해 공유할 개념화된 명세"로서 정보교환용으로 합의된 어휘를 만들기 위하여 특정 언어로 정의되는 명확한 개념들의 기술과 그들간의 관계를 의미한다.[17] 온톨로지는 지식 공학, 자연어 처리, 협력적 정보시스템, 지능형 정보 통합등과 같은 분야에서 응용할 수 있으며, EAI, CRM, KM, BPM, ERP, EP 등의 여러 상업적 분야에서 도메인에 대한 정보의 공유와 재사용을 목적으로 활용이 가능하다. 이러한 온톨로지는 시멘틱 웹의 주요 응용 분야이다. 시멘틱 웹은 지금까지 정보를 표현하고 배치하는데 주로 초점을 맞춘 웹과는 다름 개념으로, "웹 상의 정보에 의미를 부여하므로 컴퓨터 역시 사람들과 마찬가지로 정보에 대한 해석이 가능한 웹"[13]을 뜻한다. 이를 이루기 위해서는 컴퓨터가 이해할 수 있도록 정보를 컴퓨터가 해석하기 쉽도록 의미를 부여하고 이를 계층화 하는 것이 필요한데, 이는 특정 영역의 설명을 위해 공유할 개념화된 명세로 정의되는 온톨로지를 이용하여 이룰 수 있다. 본 논문에서 다루고자 하는 것은 온톨로지관련 분야에서 온톨로지 매핑에 대한 것을 다루고자 한다. 온톨로지 매핑이란 일반적으로 비교 대상으로 사용된 두 온톨로지 간의 동일성을 판단한 뒤, 그 가운데 비슷한 개념의 노드들을 서로 연결시키는 작업을 의미한다. 온톨로지 매핑은 온토롤지 합병(Ontology Merging)을 위해 반드시 거쳐야 할 단계이며, 온톨로지 합병을 통하여 여러 곳에 분산 배치되어 있는 자료들을 통합관리 하거나, 여러 온톨로지를 통합하여 하나의 큰 온톨로지로 만든 후 시멘틱 웹 기술을 이용하여 온톨로지 내 자료들에 대한 통합 검색을 이룰 수 있다. 본 논문은 온톨로지를 효과적이고 효율적으로 매핑하고자 하는 방법론의 적용 방법을 실험을 통해 알아보고자 하는데 그 목적이 있다. 온톨로지 매핑방법론의 성능을 평가하는데 있어서 본 논문에서 사용하는 방법은 정확도(precision), 재현률(recall)을 이용하여 구하는 f-measure를 사용하였다.[2] 앞서 언급했던 '효과적인 온톨로지 매핑'은 온톨로지 매핑 결과에 대한 f-measure 값을 높이는 방법론을 뜻하고, '효율적인 온톨로지 매핑 방법론은 이러한 f-measure 값을 얻을 수 있는 precision과 recall을 구하는데 있어서 보다 적은 계산으로 구하는 것을 뜻한다. 이와 관련하여 2장에서는 온톨로지 매핑, f-measure의 값을 결정짓는 정확도와 재현률에대한 내용과 본 논문의 실험에 사용된 온톨로지 매핑 방법론에 대하여 알아볼 것이고, 3장에서는 온톨로지 매핑 방법론을 이용한 실험에 대해서 알아볼 것이다. 4장에서는 휴리스틱적 온톨로지 매핑 방법론을 이용하여 효과적이고 효율적인 온톨로지 매핑

방법론 구현 방안에 대하여 알아보고 마지막 5장에서는 결론과 향후 연구에 대해서 다룰 것이다.

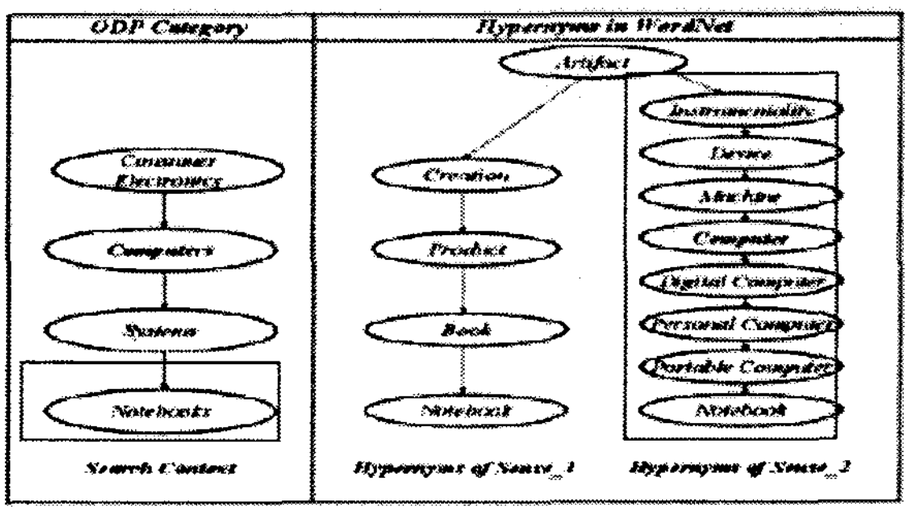
2. 관련연구

2.1 온톨로지 매핑

온톨로지 매핑을 다루기 앞서 온토롤지 매핑과 비슷한 개념인 정보 통합에 대하여 알아보고자 한다. 정보통합이란, 서로 이질적인 프로그래밍 언어나 포맷으로 이루어진 수많은 데이터를 통합하여 요약된 형태로 제공하는 기술이라 정의 내릴 수 있다. [18] 이러한 정보통합은 정보기술의 발전으로 인한 웹 상의 정보가 폭발적으로 증가하여 사용자가 이러한 상황에 오히려 피곤을 느끼는 정보피곤증후군까지 낳고 있는 정보의 홍수에 대한 대안으로써 여러 곳에 산재해 있는 필요이상의 정보를 한곳에 모아 통합관리하고자 하는 필요에 의하여 대두되었다. 시멘틱 웹 환경 하에서 정보통합이란, 일반적으로 온톨로지 통합을 의미한다. 이러한 정보통합은 합병, 정렬, 변환, 명료화로 분류할 수 있다. 합병(Merging)은 2개의 온톨로지를 결합하여 새로운 온톨로지를 생성하는 방식을 의미하며, 합병을 통하여 새롭게 생성된 온톨로지는 앞서 두 온톨로지의 특성을 모두 지니고 있어야 한다. 정렬(Alignment)은 두 온톨로지를 비교하여 비슷하거나 동일한 노드를 가려내는 방법으로, 합병을 위해 선행되어야 하는 과정이라고 할 수 있다. 변환(Transformation)은 온톨로지를 매핑하는 과정에서 효과적인 매핑 작업을 위해 특정 노드를 변형시키는 것을 말한다. 마지막으로 명료화(Articulation)은 두 온톨로지 노드들의 일부만 속성만을 매핑 하는 것을 말한다. 현재는 온톨로지 매핑에 있어서 통합과, 정렬이 동일한 의미로 사용되고 있으며, 가장 큰 패러다임을 이루고 있다. 온톨로지 매핑과 관련된 연구 분야는 Mapping Discovery, Declarative formal representations of mappings, Reasoning with mappings로 나눌 수 있는데, [1] Declarative formal representations of mappings는 매핑의 결과를 어떤 식의 공통적인 표현 방법으로 보여줄 것인가를 연구하는 분야이고, Reasoning with mappings는 추론을 통해 온톨로지 매핑을 이끌어 내려는 분야이다. 마지막으로 Mapping Discovery는 주어진 두 개의 온톨로지에 대하여 유사성이나 일치여부를 판별하는 작업을 말하고 본 논문의 접근방향도 Mapping Discovery와 같다. 두 온톨로지의 유사성과 일치 여부를 판단하는 방법에도 공통 온톨로지를 사용하거나, 휴리스틱적 기계학습적 방법으로 온톨로지 매핑을 수행하는 방법으로 나눌 수 있다. 공통 온톨로지는 매핑 대상이 되는 두 온톨로지를 공통의 온톨로지에 매핑하는 방법으로 빠른 매핑을 이룰 수 있지만, 공통 온톨로지 자체가 주관적으로 만들어 질 수 있고, 온토롤지 갱신시 이를 적용해 주어야 한다는 단점이 있다. 기계학습적 휴리스틱적 방법은 공통 온톨로지를 사용할 수 없는 경우 취하는 방법으로, 본 논문이 지향하는 방법이기도 하다. 일반적으로 온톨로지 매핑은 온톨로지를 이루고 있는 여러 노드들에 대하여 이루어 지는데, 온토롤지의 스키마만을 가지고 매핑을 수행하는 스키마기반 매핑[3] 과, 노드의 인스턴스까지 고려하는 방법이 있다. 노드의 인스턴스까지 고려하면 매핑의 정확성은 늘어나겠지만, 경우에 따라 계산시간이 기하급수적으로 늘어날 수 있다는 단점이 있다. 본 논문에서 적용하고자 하는 방법은 온톨로지 스키마만을 대상으로 하는 매핑방법론 이다. 스키마 기반 매핑 방법론은 다시 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-level는 각각 정보해석의 방향에 따라 여러 가지 갈래로 나뉜다. [3] 본 논문에서 적용하는 방법은 Structure-level의 taxonomy-based이나, repository of structures 이다.

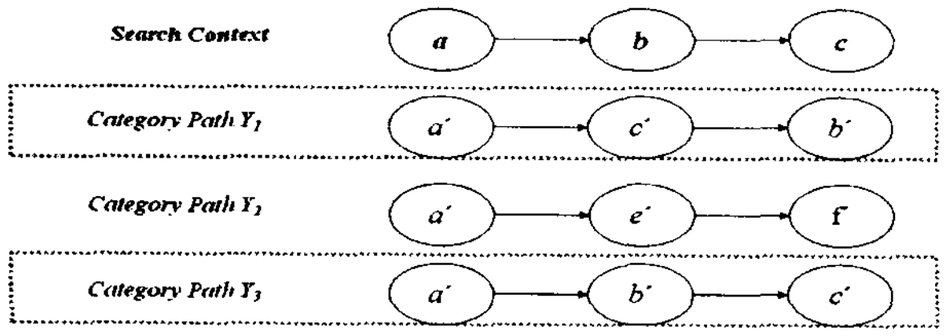
2.2 온톨로지 매핑 알고리즘

본 논문에서 사용한 온톨로지 매핑 방법론은 크게 네 단계로 나뉜다. 첫 번째 단계는 소스 온톨로지 매핑 대상에 대한 정확한 의미 파악이다. 의미 소스, 타겟 온톨로지간 노드가 동의어에 의하여 표현되었을 경우 이를 매핑하기 위한 수단으로, WordNet과 Stemming을 이용하여 매핑대상의 정확한 의미파악을 수행한다. 소스에 있는 노드가 속한 Path와 노드의 WordNet결과를 비교하여 정확한 의미를 파악하게 된다. 다음은 그 간단한 예이다.

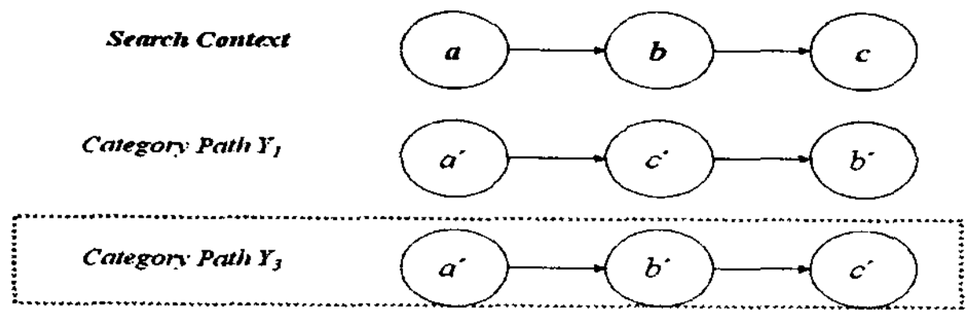


왼편은 Notebooks라는 노드가 속한 온톨로지 Path이고 우측은 Notebook을 Stemming한 결과인 Notebook의 WordNet결과들 이다. 노드가 Path와 우측 사각형안의 WordNet 결과와 유사하므로 우측의 동의어를 사용하게 된다. 두 번째 단계는 소스와 매핑을 할 타겟 온톨로지의 노드들 중 Candidate Path를 결정하는 단계이다. 이는, 앞 단계를 거친 소스 노드를 참고하여 이를 포함하고 있는

모든 타겟의 Path를 Candidate Path로 설정하고, 이에 대하여 서로간의 매핑을 수행한다. 이 단계는 계산시간의 절약과, 보다 정확한 온톨로지 매핑을 위한 단계이다. 세 번째는 소스와 타겟 온톨로지의 유사성 판단 단계이다. 먼저 파악할 것이 소스와 타겟 Path간에 얼마나 많이 동일한 노드 이름을 가지고 있나를 판단한다. 이를 그림으로 표현하면 다음과 같다.



그림에서 보듯이 소스에 있는 {a,b,c}와 같은 이름을 많이 가지고 있는 타겟 path1{a',c',b'}과 path3{a', b',c'}에 대해서 소스 path와의 유사성을 판단하게 된다. 그 다음은 소스와 타겟간의 일관성을 판단하는 것이다. 이것은 얼마나 많은 동일 노드가 동일 순서로 배치되었는가를 판단한다.



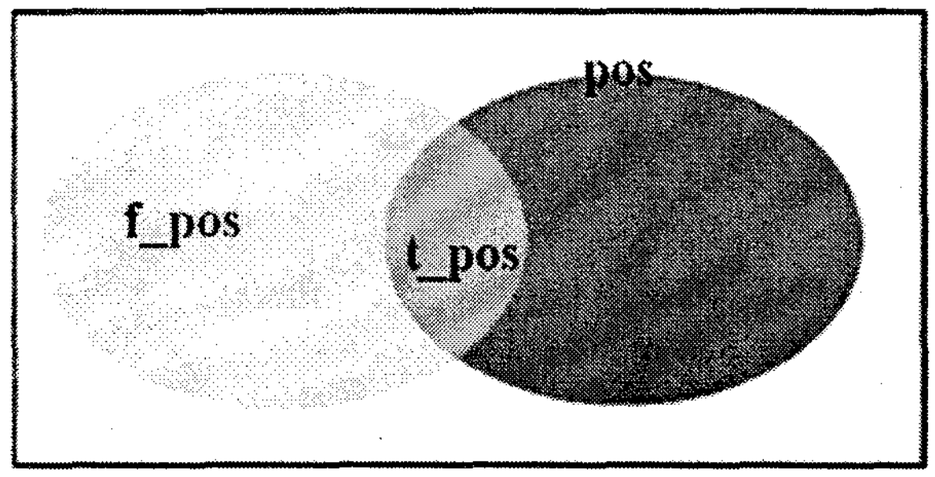
그림에서 보듯이 소스에 있는 {a,b,c}와 같은 노드를 가지고 있고, 동일한 순서를 보이는 path3{a', b',c'}에 대해서 소스 path와의 유사성을 판단하게 된다. 마지막 네 번째 단계는 소스와 타겟의 유사성에 대한 최종 판단이다. 이는 앞에서 구한 동일 노드의 출현빈도 값과, 일관성 유지 값에 α, threshold 값을 적용시켜 산출한다.

$$\text{If } \alpha \cdot \text{Co Occurence(Source, Target)}$$
$$+(\alpha\text{-}1) \cdot \text{Order Consistency(Source, Target)} \geq \text{threshold}$$

위의 조건이 만족할 경우 소스와 타겟은 유사한 것으로 평가한다.

2.3 정확도와 재현률

정확도와 재현률은 f-measure값을 결정짓는 요소이다. 정확도는 매핑을 얼마나 정확하게 했는지를 알아보는 요소이고, 재현률은 매핑을 수행했을 때 얼마나 많은 매핑 결과가 나타나는지 알아보는 요소이다.



위의 그림에서 pos는 매핑을 해야 하는 적합한 data의 set을 말하고, f_pos는 매핑결과 유사하다고 나타났으나, 실제로는 유사하지 않은 것 즉, pos가 아닌 것을 말한다. 마지막으로 t_pos는 매핑 결과 유사하다는 결과를 나타냈고 또 실제로도 pos와 유사한 것을 말한다. 이를 식으로 나타내면 다음과 같다.

$$\text{Precision} = \frac{t\_pos}{t\_pos + f\_pos}$$

$$\text{Recall} = \frac{t\_pos}{pos}$$

3. 실험

3.1 실험의 개요

앞 장에서 설명한 온톨로지 매핑 방법론을 이용하여 온톨로지 매핑을 수행할 때, α, threshold 의 변화에 따라서 정확도와 재현률은 변화한다.[16] 본 실험은 정확도와 재현률이 변화함에 따라 같이 변화하게 될 f-measure 값의 변화를 알아봄으로 휴리스틱적인 방법으로 최적화된 온톨로지 매핑을 이루는 α, threshold 값을 구하는데 참고 자료를 얻기 위한 기초실험이다. 본 실험에 사용된 온톨로지는 OAEI에서 온톨로지 매핑 방법론을 시험하기 위해 제공하는 온톨로지를 이용하였다. 실험에 사용한 온톨로지는 서지학(Bibilography)과 관련한 소스 1개와 소스와 약간 다른 타겟 4개로 이루어져 있다.

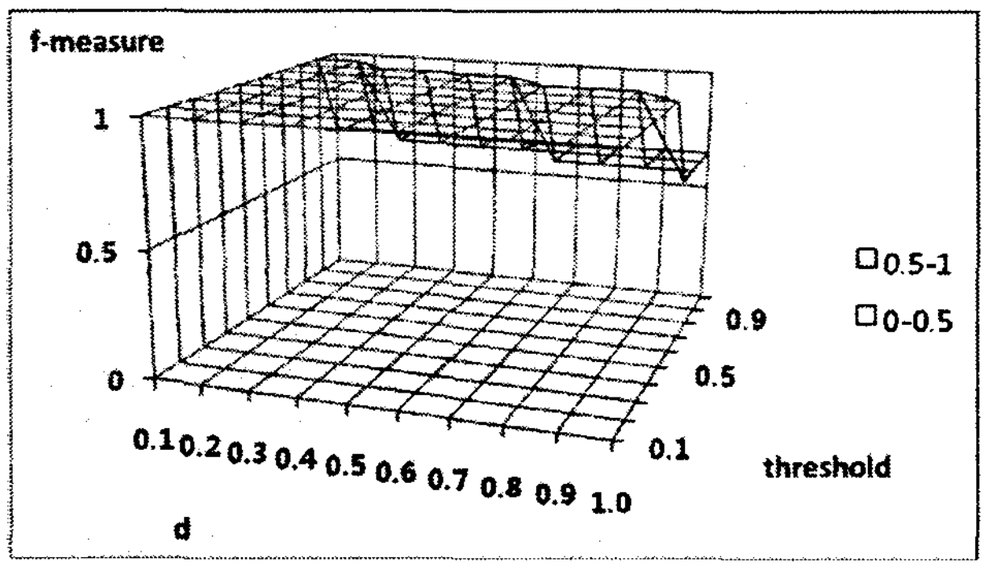
표1 - 타겟 온톨로지 설명

온톨로지	특징
222	소스에 비해 간단한 하위 구조
223	하위 구조 추가
238	하위 구조 추가
231	소스와 같은 구조
102	소스와 다른 구조

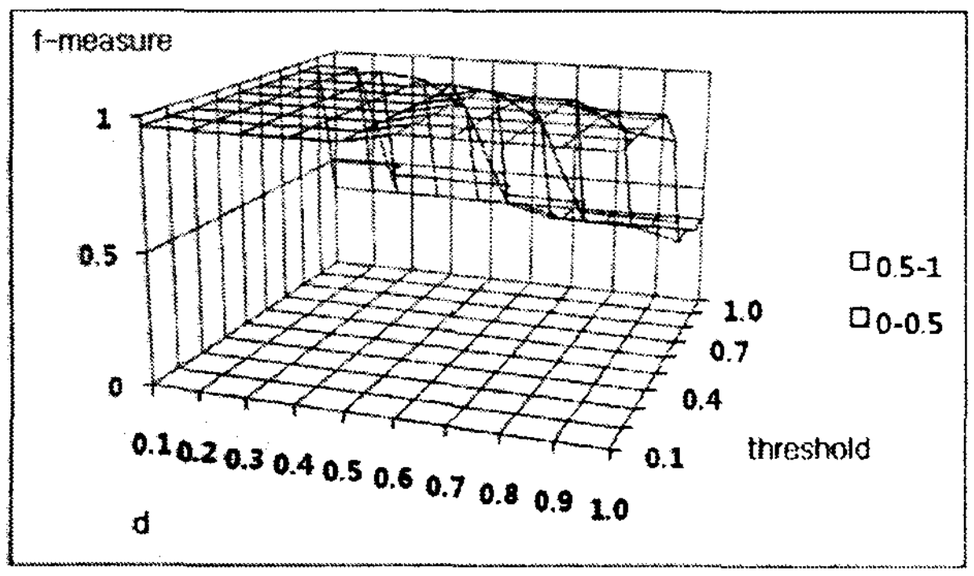
실험은 α, threshold를 각각 0.1~1.0까지 0.1씩 총 100회 변화시켜가면서 변화에 따른 f-measure 값의 변화를 측정하였다.

3.2 실험결과

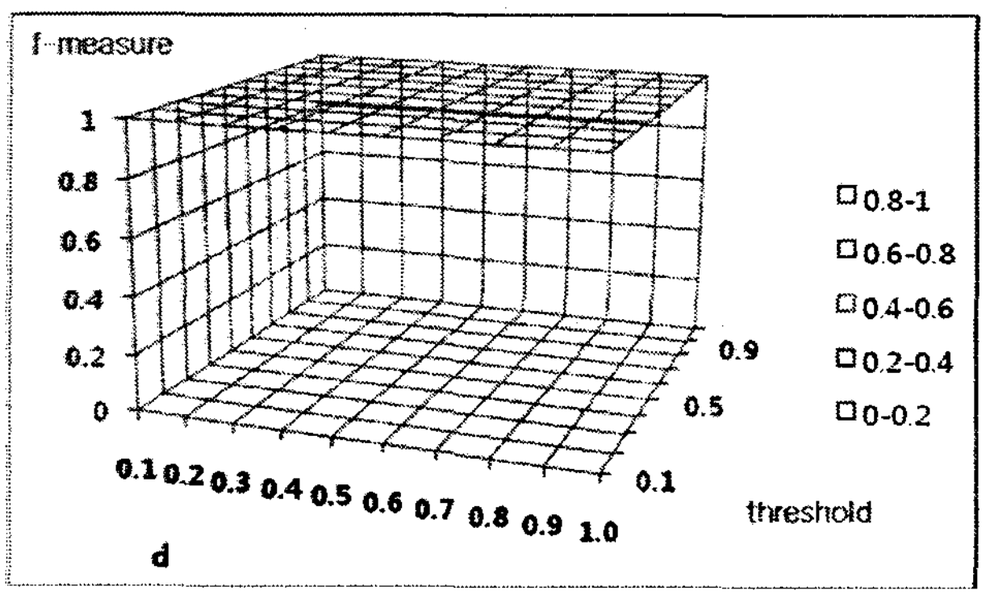
다음은 실험결과의 일부이다.



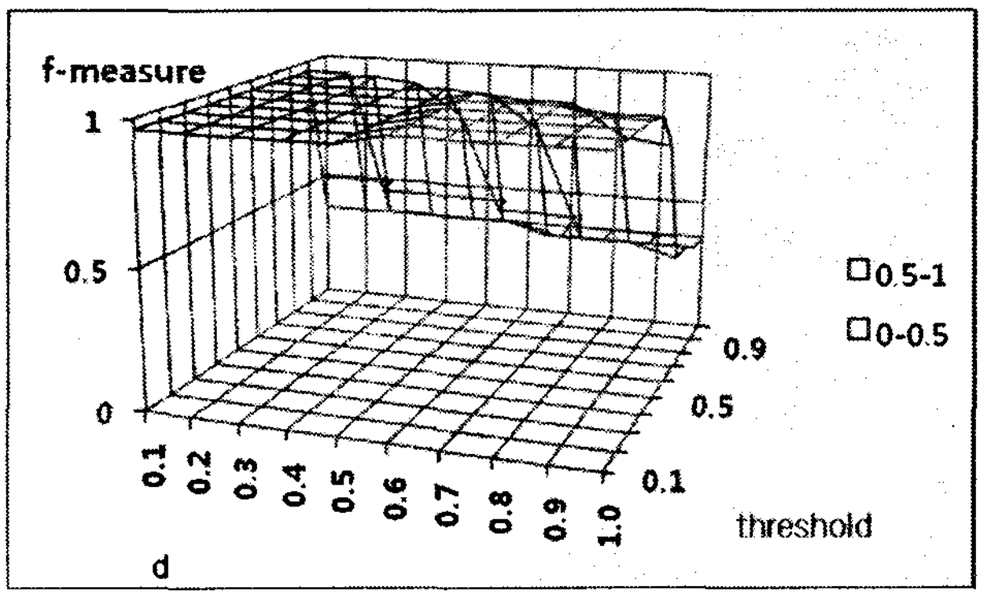
source: 101, target: 222.



source: 101, target: 223.



source: 101, target: 231



source: 101, target: 238

source:101, target102 = 결과 없음.

실험의 결과 α, threshold에 따라 정확도와 응답률의 변화에 따른 f-measure도 같이 변화함을 볼 수 있었다. 완전히 일치하는 231 온톨로지를 제외하고, 가장 높은 f-measure의 값은 각 온톨로지마다 각기 달랐다. 일반적으로 threshold를 05 이하로 고정시킨 상태에서 α를 0.6 이상 움직일 때 가장 높은 f-measure 값을 구할 수 있었는데, 이는 소스 온톨로지에 있는 노드가 타겟에 모두 비슷한 구조로 포함되어있는 OAEI 온톨로지의 특징 때문이라 생각할 수 있다. 소스와 타겟이 일치하는 101과 231 온톨로지의 경우 모든 α, threshold의 값에 대하여 완벽하게 일치함을 보여졌다. 본 실험을 통하여 높은 f-measure를 구하는데 있어서 특정 값의 정확도와 응답률이 필요한 것이 아닌, 값들의 적당한 조정을 통하여 높은 f-measure를 구할 수 있는 것을 알 수 있었다.

4. 휴리스틱적 온톨로지 매핑 방법론

본 장에서는 앞서 실험을 바탕으로 나타난 가장 높은 f-measure값을 구하는 α, threshold를 휴리스틱적으로 찾는 방법을 제안하고자 한다. 앞서 실험에서 최적의 f-measure값을 구할 수 있는 α, threshold를 찾는데 각 온톨로지 마다 100회의 방법이 필요했다. 본 방법론의 목표는 연구에서 사용되었던 100회보다 적은 계산으로 최적의 f-measure, 혹은 최적과 가까운 f-measure값을 구할 수 있는 α, threshold를 얻음을 목표로 한다. 방법론은 다음과 같이 정리할 수 있다.

1. 10X 10의 메트릭스를 만든다, 각 행과 열은 0.1씩 α, threshold가 변화함을 뜻한다.

t/α	0.1	0.2	0.3	…	0.9	1.0
0.1						
0.2						
0.3						
…						
0.9						
1.0						

2. 임의의 한 구간을 결정하고 해당 구간에서의 f-measure 값을 구한다.

3. 선택한 구간의 주위 (최대 8방향) 중 다른 방향을 임의로 정하고 해당 구단의 α, threshold를 이용하여 f-measure 값을 구한 후, 비교한다.

t/α	0.1	0.2	0.3	…	0.9	1.0
0.1						
0.2						
0.3			0.5			
…		0.65				
0.9						
1.0						

4. 3에서 선택한 구간의 f-measure값이 2의 f-measure값보다 크면 해당 구간으로 이동하고, 같거나, 작으면 다른 구간을 선택한다.

t/α	0.1	0.2	0.3	…	0.9	1.0
0.1						
0.2						
0.3			0.5			
…		0.65				
0.9						
1.0						

* 3에서 선택한 구간이 더 클 경우.

t/α	0.1	0.2	0.3	…	0.9	1.0
0.1						
0.2		(2)	(3)	(4)		
0.3		(1)	0.5	(5)		
…		0.4	(6)	(6)		
0.9						
1.0						

* 3에서 선택한 구간이 같거나, 작을 경우 (1)~(6)까지의 다른 구간을 선택하여 비교한다.

5. 선택한 구간의 f-measure 값이 주위 모든 방향의 f-measure 값보다 크거나 같을 때 까지 계속한다.

최대 8개인 주위 구간의 f-measure 값보다 현재의 구간의 f-measure값이 크거나 같을 때 탐색을 계속한다. 해당 방법의 의의는 100회의 계산보다 적은 계산을 통하여 최적의 α, threshold 조합을 구하는 것이므로, 전자의 선택이 옳다고 판단된다.

5. 결론 및 향후 연구과제

본 논문에서는 온톨로지 매핑 방법론에서 사용되는 파라미터인 α, threshold가 변화함에 따라 정확도와 재현률이 같이 변하고, 이에 따라 자연히 변하는 f-measure 중 가장 높은 f-measure 값을 주는 파라미터의 조합을 휴리스틱적인 방법으로 찾는 방안을 제안하였다. f-measure는 정확도와 재현률의 조합으로 이루어지는데, 어느 한쪽을 일방적으로 높임으로 f-measure값 역시 같이 높아지지는 않기 때문이다. 이를 위해 최적 α, threshold 조합을 찾기 위한 연구의 기초자료로 사용하기 위해서 수행한 실험을 통해 α, threshold 각각 100번 변화시켜 가면서 f-measure의 변화를 알아보았고, 휴리스틱적 방법은 이의 계산보다 적게 계산함을 목표로 하였다. 휴리스틱적 방법론을 적용하는데 있어서 중요시되는 것이 이 계산 횟수라고 할 수 있다. 반복된 계산으로 최적의 α, threshold 조합을 얻었다 하더라도, 100번의 계산 횟수가 넘거나, 이와 가까운 횟수의 계산이 수행된다면 효율적인 방법이라 할 수 없기 때문이다. 앞 장에서는 주위 모든 f-measure 값보다 횟수가 클 때까지 탐색을 계속한다고 정하였지만, 계산횟수에 대한 문제는 기초실험에서와 같이 100번에 걸쳐서 α, threshold를 변화시켜가며 높은 f-measure값을 얻을 수 있는 최적의 조합을 구하고 같은 온톨로지에 대하여 휴리스틱적인 방법론을 적용하여 탐색 횟수를 변화시켜가며 매핑을 수행하는 실험을 통해 합리적 수준의 탐색횟수를 위한 연구가 필요하다고 판단된다.

References

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

[2] Herlocker, J, "Understanding and Improving Automated Collaborative Filtering Systems," *Ph.D. Thesis, Computer Science Dept, University of Minnesota, 2000*

[3] Magnini B., Speranza M., Girardi C (2004) : A Semantic-based Approach to Interoperability of Classification Hierarchies: Evaluation of Linguistic Techniques, *Proceedings of COLING-2004*

[4] Mikalai Yatskevich, Fausto Giunchiglia and Paolo Avesani(2006) : A Large Scale Dataset for the Evaluation of Matching Systems

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

[6] OAEI (Ontology Alignment Evaluation Initiative) http://oaei.ontologymatching.org/2007/

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

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

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

[10] Paolo Avesani, Fausto Giunchiglia and Mikalai Yatskevich (2005) : A Large Scale Taxonomy Mapping Evaluation, *"In Proceedings of International Semantic Web Conference(ISWC)"*

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

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

[13] Tim Berners-Lee, James Hendler, Ora Lassila.(2001): "The Semantic Web" *Scientific American. 2001*

[14] 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*".

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

[16] 안성준, 김우주, 박상언 (2007) : 응용환경 적응을 위한 온톨로지 매핑 방법론에 관한 연구 "*지능 정보 시스템 학회 춘계 학술대회*"

[17] 웹 온톨로지 개발 지침 연구 (2004) : 정보사회 진흥원 연구보고서

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

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