분산 환경에서 쿼리 변환을 위한 온톨로지 매핑 결합

정재은
영남대학교 컴퓨터공학과
j2jung@intelligent.pe.kr

Ontology Mapping Composition for Query Transformation on Distributed Environment

Jason J. Jung
Dept. of Computer Engineering, Yeungnam University

요 약

온톨로지 기반 분산 정보시스템 환경에서는 시스템들 간의 정보 공유을 위해 전문가에 의한 명시적 온톨로지 매핑(Explicit Mapping)을 통해 의미적 이질성(Semantic Heterogeneity)을 해결하고 있다. 하지만, 온톨로지 매핑의 고비용성 때문에 모든 정보시스템들 간의 매핑이 이루어지기 힘들다. 따라서, 본 논문에서는 이미 존재하는 온톨로지 매핑 정보들의 재사용을 통해 존재하지 않은 온톨로지 매핑 정보를 묵시적으로 예측하고자 한다. 본 논문에서는 이와 같은 분산 환경에서의 쿼리 전송을 통한 지식 검색에 있어서의 온톨로지 매핑을 기반으로 한 적절한 쿼리 변환 방법론을 소개하고자 한다.

1. 서론

다양한 분야의 정보시스템들은 도메인 온톨로지의 구축을 통해 시스템 내의 정보 및 지식의 관리와 유통에 효율성을 높여왔다[1,2]. 보다 중요하게, 분산 환경에서는 각 시스템들 간의 정보 공유을 기반으로 한 상호운용성(Interoperability)를 필요로 한다. 하지만, 시스템 간의 의미적 이질성(Semantic heterogeneity) 문제를 해결하기 위해서는 각 분야 전문가들에 의해 해당 온톨로지들 간의 매핑(Mapping)[1] 정보를 충분조건으로 하고 있다.

이와 같은 온톨로지 간의 매핑 과정은 단순히 해당 분야의 전문성 부족 뿐만 아니라, 온톨로지의 복잡한 내부 구조(방대한 클래스의 수, 클래스들 간의 관계와 같은) 인식의 어려움들 때문에 매우 고비용(expensive)의 작업임에 틀림없다. 이와 같은 고비용성 문제를 해결하기 위하여, 주어진 임의의 두 온톨로지 간에 자동화된 매핑 알고리즘에 대한 다양한 연구가 진행되고 있다[5]. (보다 자세한 내용은 http://www.ontologymatching.org/ 참조)

그와 같은 온톨로지 간의 명시적이고 직접적인 매핑(Explicit and direct mapping) 알고리즘에 반해, 본 논문에서는 기존에 이미 존재하는 온톨로지 매핑 정보를 재사용(Reuse)함으로써 정보시스템들 간의 상호운용성을 지원하고자 한다. 다시 말해, 정보시스템 S_i와 S_j의 두 온톨로지 O_i, O_j 간의 직접적인 매핑 $M(O_i, O_j)$을 계산하는 대신, 기존에 존재하는 매핑 $M(O_i, O_k)$, $M(O_k, O_j)$의 적절한 결합(Composition)을 통해 S_i와 S_j간의 정보공유를 가능케하고자 한다.

이를 위하여, 온톨로지와 온톨로지 매핑의 형식화를 위한 몇 가지 정의들을 언급해야하며, 특히, 우리는 매핑을 신뢰성을 정량화하기 위해 새로운 척도(Measurement)를 소개한다.

특히, 본 논문에서는 이와 같은 매핑 결합 알고리즘을 분산 환경에서의 에이전트 시스템과 같이 자동화된 시스템들에 쿼리 기반의 지식 검색 시스템에 접목시키고자 한다. 즉, 이질적인 에이전트 시스템들 간의 자동으로 생성되고 전송되어지는 쿼리가, 원하는 수준의 검색 효율을 유지하기 위해서는 적당한 의미적 변환 방법이 필요하다.

다음 2장에서는 온톨로지 기반의 분산 정보시스

[1] 본 논문에서 말하는 온톨로지 매핑은 Alignment와 Matching과 동일한 의미로 간주될 수 있다.

템을 구성하는 요소들을 정의하고 의미 유사도(Semantic similarity) 기반의 온톨로지 매핑 알고리즘을 소개한다. 3장에서는 온톨로지 매핑의 결합에 대해 설명하고자 하며, 4장에서는 몇 가지 매핑 결합에서의 고려되어져야 할 이슈들을 다룬다. 마지막으로 5장에서는 본 논문의 결론을 도출하고, 향후 연구 방향을 소개한다.

2. 온톨로지 기반의 분산 정보시스템

본 논문에서 다루고 있는 온톨로지 기반의 분산 정보시스템은 정보시스템들간의 지식을 서로 공유하기 위함을 목적으로 한다. 즉, 이와 같은 정보시스템 S_A은 i) 온톨로지 O_A와 ii) 지식 공유를 위한 정보시스템의 온톨로지와의 매핑 정보 로 모델링되어 진다. (그림 1 참조)

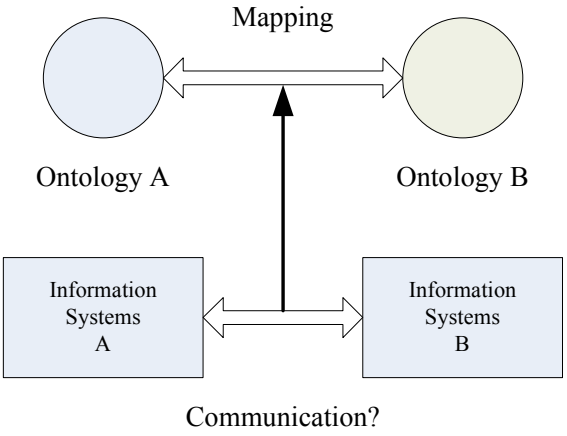

그림 1. 온톨로지 매핑과 정보시스템의 정보 공유

Definition 1 (온톨로지) 온톨로지는 크게 클래스 집합 C와 프로퍼티 집합 P로 이루어져 있으며, 각각의 요소들은 특정 단어(Term)로 레이블 되어 있다.

이와 같은 온톨로지들은 임의의 매핑 알고리즘에 의해 온톨로지 요소들 간의 특정 관계가 있음을 알 수 있으며, 이와 같은 요소들 간의 관계를 Correspondences라 부른다.

Definition 2 (Correspondence) 두 온톨로지 간의 Correspondences는 <e_1, e_2, R>와 같은 Triple 형식으로 표현되는데, e_1와 e_2은 각각 온톨로지 O_1과 O_2의 요소들이다. 특히, R은 두 요소들간의 관계를 표현하는 것으로서, equivalence(≡), subsumption (⊇, ⊆), disjunction (⊥) 등이 가능하다.

이와 더불어, 각 Correspondence는 해당 온톨로지 요소들 간의 매핑이 얼마나 신뢰할 수 있는지를 정량화하기 위해 새로운 Measurement를 추가하게 된다.

Definition 3 (Confidence) 각각의 Correspondence에는 Confidence 값을 할당하여 신뢰(정확성)의 정도를 알수있도록 한다. 따라서 <e_1, e_2, R>는 <e_1, e_2, R, CF>로 확장하게 된다.

두 온톨로지 O_1, O_2 요소들 간의 Correspondences 집합을 본 논문에서는 온톨로지 매핑(Ontology mapping)이라 부르고, $M(O_1, O_2)$로 표현 한다. 특히, 본 논문에서와 같이 Alignment API[4]를 이용한 의미 유사도(Semantic similarity) 기반의 온톨로지 매핑 알고리즘을 사용하고 있다고 한다면, Confidence 값 역시 자동으로 구할 수 있다.

이 온톨로지 매핑 알고리즘의 기본적으로 매핑되어질 온톨로지들의 요소들간의 Lexical similarity의 총합을 최대로 하는 상태를 찾는다[4,6]. 간단한 예로서 [그림 2]에서와 같이 온톨로지 O_1와 O_2 간의 매핑을 통해 다음과 같은 매핑 정보를 얻을 수 있다.

< e_2,	e_1,	R,	CF >
< Full_Prof,	Full professor,	≡,	**.64** >
< Prof,	Professor,	≡,	.48 >
< person,	people	≡,	.33 >
< Secretary,	Researcher,	≡,	.3 >

이와 같은 정의들을 바탕으로 온톨로지 기반의 분산 정보시스템을 아래와 같이 표현하고자 한다.

Definition 4 (분산 정보시스템) 분산 정보시스템 G에 참여하고 있는 각 정보시스템 S_i은 로컬 온톨로지 O_i를 가기고 있으며, 정보공유를 필요로 하는 정보시스템의 온톨로지와 매핑 정보 $M(O_i, O_j)$를 저장하고 있다. 따라서, G는 다음과 같다.

$$G = \{M(O_i, O_j) \mid L(S_i, S_j)\}$$

여기서 L은 정보시스템들간의 매핑 정보 유무를 나타내기 위한 행렬이다.

Example 1. 분산 정보시스템 G에 참여한 정보시스템들간에 쿼리 기반의 정보 공유가 이루어진다고 가정하자. [그림 2]에서 S_2에서 운용가능한 쿼리 q가 S_2에서 사용되기 위해서는 $M(O_1, O_2)$을 참조함으로써 q'으로 Transformation 되어 사용될 수 있다.

하지만, 이와 같은 분산 정보시스템에서 문제는 모든 온톨로지 간의 일대일 매핑 정보를 구하기 불

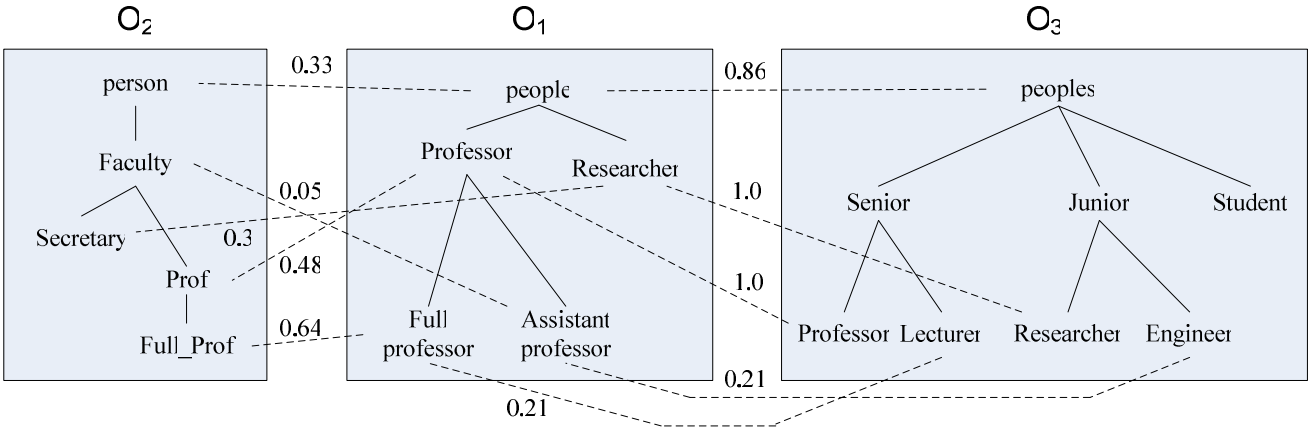

그림 2. 의미 유사도를 이용한 온톨로지 매핑[6]과 분산 정보 시스템 환경

가능하여 매핑 정보가 존재하지 않는 시스템들간의 상호운용성이 힘들다는 것이다.

3. 쿼리 변환을 위한 매핑 결합

본 연구에서는 정보 시스템들 간의 매핑 정보의 재사용(결합)을 통해 존재하지 않은 매핑 정보를 간접적으로 생성하고자 한다. 즉, [그림 2]에서 S_2와 S_3 간의 통신을 위해서 다음과 같은 간접적인 온톨로지 매핑을 예측하여 사용한다.

$$\tilde{M}(O_2,O_3) = Compose(M(O_2,O_1),M(O_1,O_3))$$

이와 같은 매핑 결합을 통해, S_2시스템의 에이전트는 간접적으로 생성된 쿼리를 변환하여 S_3시스템에 저장된 정보를 검색할 수 있다.

하지만 이와 같은 쿼리 변환을 위한 매핑 결합에 있어서 심각한 문제는 결합을 위한 경로가 하나 이상일 수 있다는 것이다.

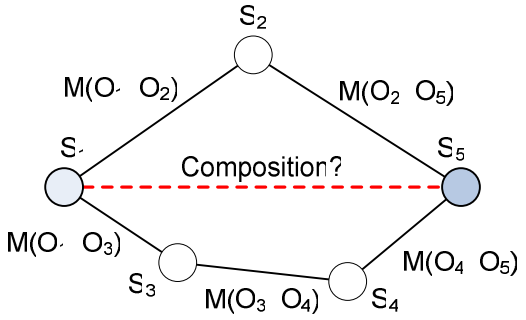

그림 3. 매핑 결합

[그림 2]에서 빨간 점선과 같이 S_1과 S_5간의 통신을 위해서는

- Compose(M(O_1, O_2), M(O_2, O_5)) 뿐만 아니라,
- Compose(M(O_1, O_3), M(O_3, O_4), M(O_4, O_5))

와 같은 매핑 결합이 가능하다. 따라서, 쿼리 변환에 있어 보다 적합한 매핑 결합을 선택할 수 있어야한다.

이를 위해서 본 논문에서는 주어진 쿼리의 분석을 통한 휴리스틱 접근법을 제안한다. 일반적으로 분산 정보 시스템들 간의 통신을 위한 쿼리는 다음과 같이 표현된다.

$$q ::= c \mid \neg q \mid q \wedge q' \mid q \vee q'$$

이와 같은 쿼리의 분석을 통해 Query-activated Class C_Q가 송신자의 온톨로지로부터 다음과 같이 얻을 수 있다. ([그림 4]에서의 녹색의 작은 원에 해당)

$$C_Q = \{c \mid c \in q, c \in O_{Sender}\}$$

예를 들어, 분산 정보시스템이 쿼리 언어로써, 다음과 같이 SPARQL[2]를 이용한다면, 정보시스템 S_1으로부터 쿼리 Q1

```
SELECT ?x, ?y
WHERE {
      ?x ?p "Full professor".
      ?y ?p "Researcher".
}
```

가 생성되었다면, 우리는 Query-activated Class C_{Q1}를 다음과 같이 구할 수 있다.

C_Q = {"Full professor", "Researcher"}

3.1 Semantic coverage ratio

주어진 C_Q로부터 본 논문에서는 매핑 결합을 위한 경로를 설정하게 되는데, 이 때 Semantic coverage ratio (τ_Q)를 다음과 같은 휴리스틱을 이용하여 계산하고 비교한다.

- 쿼리 Q에 의한 Query-activated Class C_Q가

[2] SPARQL, http://www.w3.org/TR/rdf-sparql-query/

보다 많은 Correspondence 클래스와 일치할수록, τ_Q가 증가한다.

$$\tau_Q^{H1}(S_{Src}, S_{Dest}) = \frac{\left|\{c \mid c = e_{Src}, M(O_{Src}, O_{Dest})\}\right|}{\left|C_Q\right|}$$

- C_Q와 일치하는 Correspondence의 CF값이 높을 수록, τ_Q가 증가한다.

$$\tau_Q^{H2}(S_{Src}, S_{Dest}) = \sum_{e_k \in C_Q} CF_k$$

예를 들어, [그림 4]에서 쿼리 Q에 의해 Activated된 클래스가 녹색 원들이라고 하자.

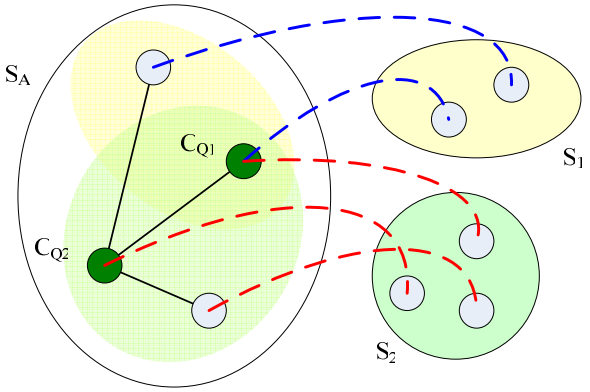

그림 4. 쿼리에 대한 매핑 결합의 선택

휴리스틱 H1에 의하면, $\tau_Q^{H1}(S_A, S_1) = \frac{1}{2} = 0.5$ 와 $\tau_Q^{H1}(S_A, S_2) = \frac{2}{3} = 0.67$ 임을 확인할 수 있으므로, 주어진 쿼리의 변환은 우선 S_2로 전달될 가능성이 크다.

3.2 Transformation Path Selection

최종적으로 우리가 찾아야 하는 정보는 쿼리의 소스(Source)로부터 Destination까지의 가능한 매핑 결합 경로들 중에서 최적의 경로이다. 그러므로, 앞에서 소개한 Semantic coverage ratio(τ_Q)의 Serial Aggregation을 수행한다.

$$\tau_Q(S_{Src}, S_{Dest}) = \max\left(\tau_Q(S_{Src}, S_i) \cdot \tau_Q(S_i, S_{i+1}) \cdot \ldots \cdot \tau_Q(S_j, S_{Dest})\right)$$

다음의 수식을 최대로 하는 매핑 경로를 통해 쿼리가 변환이 이루어져야 한다.

4. 토론

이미 Zimmermann and Euzenat[3]는 분산 정보시스템 환경에서 존재하는 온톨로지 매핑들의 결합(Composition)을 생성하는데 있어서 다음과 같은 세 가지 의미가 있음을 정리하였다.

- Simple semantic composistion
- Integrated semantic composition
- Contextualized semantic composistion

본 연구에서 제안하는 매핑 결합 방법론은 Simple semantic composition과 Integrated semantic composition에 한하여 적용하고자 한다.

5. 결론 및 향후 연구 방향

분산 환경에서 에이전트와 같은 자동화된 시스템들은 필요한 지식 및 정보의 검색을 위해 협업적 통신 기능(또는 서비스)를 필요로 한다. 시스템 간의 이질성 해결을 위해 메세지 변환이 효과적이다. 하지만, 변환을 거듭할수록, 초기에 생성된 의도(Context)가 퇴색되는 소위 Whispering problem이 있다. 본 연구에서는 온톨로지 매핑을 통해 쿼리 변환 중에 발생하는 Information loss를 최소화하고자 한다.

우선 가장 시급한 연구로는 제안된 휴리스틱들의 평가를 위한 시스템의 구현이다.

또한, 이와 같은 온톨로지 매핑 결합을 이용하여, Semantic P2P 환경[7]에서의 지식 공유시스템을 구현할 수 있다.

향후 연구로써, 우리는 블로그의 Trackback이나 RSS Feed와 같은 기능을 통해 정보 공유 서비스를 제공하고 있다. 하지만, 각 블로그나 사용자의 Context를 고려하지 않고 있으므로, 제공되는 정보가 대부분 무의미하다. 또한, 블로그 수의 증가 뿐만 아니라 Network Isolation 문제에 의해 효과적인 정보 공유가 이루어지기 어렵다. 이 문제점들을 해결하기 위해서, 본 논문에서 소개한 매핑 결합 기법을 이용하여 블로그 시스템의 BlogRoll과 같은 연결 정보를 활용하여 정보공유 서비스를 보다 향상시키고자 한다. 특히, 사회망 분석법을 적용하기 위해, Blog Overlay Network(BON) 플랫폼[9]을 설계하여, Context에 따른 Community identification을 수행하고자 한다. 이와 더불어, 본 논문에서 제안한 방법론이 블로거들의 Mental 모델에 의해 발생되는 Tag 정보들간의 Matching이 협업을 위한 해당 블로그에 저장되어 있는 정보들 간의 공유를 지원하고 있음을 보이고자 한다.

감사의 글

본 논문은 정통부 및 정보통신연구진흥원의 정보통신선도기반기술개발사업의 연구결과로 수행되었습니다.

참고 문헌

[1] Brandt, S.C., et al., "Ontology-based Information Management in Design Processes," In: Proc. 9th International Symposium on Process Systems

Engineering (2006).
[2] Chau, K. W., "An ontology-based knowledge management system for flow and water quality modeling", Advanced Engineering Software, Vol. 38, No. 3, 172-181 (2007).
[3] Zimmermann, A., Euzenat, J., "Three Semantics for Distributed Systems and their Relations with Alignment Composition," In: Proc. 5nd International Semantic Web Conference (ISWC'06). (2006) 16-29
[4] Euzenat, J., "An API for Ontology Alignment," In: Proc. Third International Semantic Web Conference (ISWC'04), (2004) 698-712
[5] Shvaiko, P., Euzenat, J., "A Survey of Schema-Based Matching Approached," Journal of Data Semantics IV, LNCS, Vol. 3730, 146-171 (2005)
[6] Jung, J.J., "Ontological Framework based on Contextual Mediation for Collaborative Information Retrieval," Information Retrieval, Vol. 10, No. 1. pp. 85-109 (2007)
[7] Jung J.J., Euzenat, J., "Towards Semantic Social Networks," In: Proc. 4th European Semantic Web Conference (ESWC'07). (2007) 267-280
[8] Lanzola, G., Gatti, L., Falasconi, S., Stefanelli, M., "A framework for building cooperative software agents in medical applications," Artificial Intelligence in Medicine, Vol. 16, No. 3, pp. 223-249 (1999)
[9] 정재은, 구철모, "블로그 환경에서의 정보 공유를 위한 상황 비교 기반의 블로그 오버레이 네트웍," Telecommunication Review, 제17권, 4호. (2007)