온톨로지의 점증적 갱신에 기반한 스키마 매칭

이준승° 이경호
연세대학교 컴퓨터과학과
jslee°@icl.yonsei.ac.kr khlee@cs.yonsei.ac.kr

Schema Matching Based on An Incremental Ontology Update

Jun-Seung Lee Kyong-Ho Lee
Department of Computer Science, Yonsei University

요 약

본 논문은 점증적으로 갱신되는 온톨로지에 기반한 스키마 매칭 알고리즘을 제안한다. 스키마 매칭에 사용되는 온톨로지는 전문가에 의하여 작성된 정적인 것으로 모든 어휘관계를 포괄하기는 힘들다. 제안된 방법은 이전의 매칭 결과와 사용자 피드백에 따라 점증적으로 온톨로지를 갱신하여 매칭의 성능을 향상시킨다. 특히, 제안된 온톨로지는 분할, 병합 관계를 기술하고 있어 단순한 매칭관계뿐만 아니라 복합매칭관계 추출을 가능케 한다. 성능평가를 위한 실험결과, 점증적 온톨로지의 적용이 매칭 성능을 매우 향상시킴을 알 수 있었다.

1. 서 론

스키마 매칭은 입력으로 받은 두 스키마가 포함하고 있는 요소들 사이의 의미적 관계를 찾는 것으로, 정보의 상호운용에 중요한 역할을 한다. 특히, 다른 스키마로 작성된 XML 문서를 각각의 스키마에 적합하게 변환하기 위해선 더욱 정확한 스키마 매칭이 선행되어야 한다. 현재 사용되고 있는 문서변환 시스템은 전문가의 수동 스키마 매칭에 의존하고 있다. 하지만 스키마의 양이 증가함데 따라 자동화된 스키마 매칭의 필요성이 커지고 있다.

현재까지 진행되어온 스키마 매칭에 관한 연구의 대부분은 어휘 사이의 관계를 포함하고 있는 부가정보에 기반한다. 일반적으로 동의어 사전이 사용되지만 특정분야는 사용되는 어휘의 의미가 서로 다를 수 있다. 따라서 많은 연구에서 동의어 사전과 함께 특정분야에서 사용되는 어휘간의 의미적 관계가 정의된 도메인 온톨로지를 이용한다.

대부분의 도메인 온톨로지는 그 분야에 대해 충분한 지식을 지닌 전문가에 의해 정의된다. 하지만 전문가라 하더라도 다양한 스키마에서 사용된 모든 어휘를 포괄하는 것은 어려운 일이다. 또한 기존의 정적인 온톨로지는 어휘간의 의미 변화에 따라 온톨로지를 재구축 해야하는 어려움이 있다. 따라서 본 연구에서는 이전의 스키마 매칭결과에 따라 점증적으로 갱신되는 온톨로지에 기반한 스키마 매칭 방법을 제안한다. 또한 단순한 매칭관계 뿐만 아니라 변환과정에 적용될 연산에 따라 단순매칭과 복합매칭으로 분류하여 계산할 수 있는 방법을 제안한다.

소스와 타겟요소 사이에 일대일(one-to-one) 매칭 관계를 형성하는 단순매칭은 특별한 연산 과정 없이 문서 변환에 적용이 가능하다. 하지만 모든 매칭결과가 단순매칭으로 나타나지는 않는다. 여러 요소가 하나의 요소와 매칭되는 복합매칭이 발생할 수 있다.

복합매칭은 소스와 타겟요소 사이에 일대다 혹은 다대일의 관계의 매칭으로 분할(spilt) 또는 병합(merge)과 같은 연산을 포함한다. 즉, 분할연산이 적용되는 복합매칭에서는 하나의 소스요소에 포함되어 있는 정보가 적절히 나뉘어 여러 타겟요소로 매칭된다. 병합연산이 적용되는 경우, 여러 소스요소에 포함되어 있는 정보가 합쳐져 하나의 타겟요소로 매칭된다.

제안된 방법의 성능평가를 위하여 부동산 목록에 사용되는 실제 스키마 5개를 이용하여 실험하였다. 실험결과, 온톨로지를 적용하지 않은 경우 정확률은 97%으로 높았지만 58%의 재현률로 많은 매칭관계를 추출하지 못했다. 하지만 온톨로지 적용결과 재현률은 평균 73%로 찾지 못했던 많은 매칭을 찾아냄을 확인하였다. 특히, 복합매칭을 효과적으로 추출하였다.

이 논문은 2003년도 한국학술진흥재단의 지원에 의하여 연구되었음 (KRF-2003-003-D00429).

2. 관련연구

기존의 스키마 매칭에 관한 연구가 활발히 진행되었다. 하지만 Rahm 등 [1]의 스키마 매칭에 관한 서베이에 따르면 대부분의 연구가 특정 도메인에 사용되는 어휘사전을 활용하고 있다. 예를 들어 Madhavan 등 [2]이 제안한 Cupid는 XML 스키마의 계층구조를 활용하여 매칭관계를 계산하고 있지만 기본적으로 실험에 사용된 데이터인 invoice에 대한 어휘사전을 사전에 정의하고 있다.

특히 Xu [3] 가 제안한 방법은 계층적 구조의 온톨로지를 활용하여 복합매칭을 찾을 수 있는 방법을 제안한다. Xu가 제안한 온톨로지는 특정 도메인에 포함될 모든 요소들의 관계를 방향 그래프로 표현되며 각 요소에 포함될 수 있는 어휘리스트를 활용하여 온톨로지와 입력된 스키마 사이의 매칭을 찾는다. 즉, 입력된 두 스키마와 온톨로지의 분석을 통해 스키마 사이의 복합매칭을 계산한다. 하지만 제안된 온톨로지는 전문가에 의해 수동으로 설계되어 입력받는 모든 스키마를 포괄하기 어렵고 사용되는 어휘목록 역시 구축에 많은 비용이 소요되며, 갱신이 필요한 경우 수동으로 일일이 입력해야하는 어려움이 있다.

Doan 등 [4]이 제안한 COMAP는 학습기법을 활용하여 복합매칭을 찾는 방법으로 미리 학습된 분류기를 이용하여 스키마에 따라 작성된 문서를 입력받아 스키마 사이의 매칭을 찾는다. 특히 COMAP는 소스와 타겟사이에 재사용되는 데이터에 대한 정보를 이용하여 수식이 포함된 복잡한 연산이 적용된 매칭을 계산할 수 있는 방법을 제안한다. 하지만 COMAP는 관계형 데이터베이스를 대상으로 제안되어 계층 구조의 XML 스키마의 특성을 반영하지 못한다. 또한 학습 데이터가 다양한 입력 스키마를 포괄할 수 있어야 정확한 매칭결과를 얻을 수 있다.

본 논문에서는 복합매칭의 관계를 정의할 수 있으며 동적 갱신이 가능한 온톨로지 구조에 기반한 스키마 매칭 방법을 제안한다. 특히, 구조적 정보를 활용하여 매칭을 선택하기 때문에 XML 스키마에 적합하며, 기존의 매칭 결과를 활용하여 갱신된 온톨로지를 이용하기 때문에 더욱 향상된 매칭결과를 기대할 수 있다.

3. 문서모델 및 온톨로지

본 절에서는 스키마 매칭에서 발생할 수 있는 매칭관계를 정리하고, 어휘 사이의 매칭관계를 기술하고 있는 온톨로지의 구조

를 설명한다. 특히 제안된 방법은 매칭관계를 크게 변환과정에서 특별한 연산이 필요없이 변환이 가능한 단순매칭과 분할이나 병합과 같은 연산이 필요한 복합매칭으로 분류한다.

3.1. 스키마 트리 모델

본 절에서는 스키마를 효과적으로 표현할 수 있는 문서 모델을 간략히 정의한다. XML 스키마를 표현하기 위해 뿌리 노드(root node)를 포함하며 형제 노드간에 순서가 존재하는 순서 트리(ordered tree)에 기반한 문서 모델을 제안한다. 문서 모델은 XML 스키마에 정의된 요소와 속성을 노드로 갖는다. 각각의 노드는 레이블(label)과 값(value)를 갖는다. 레이블은 XML 스키마 문서에서 정의된 요소와 속성의 이름이고 값은 XML 스키마에 정의된 요소나 속성의 데이터 타입으로 단말노드만 값을 갖게 된다. 본 논문에서는 제안된 문서 모델로 표현된 XML 스키마를 스키마 트리라고 정의한다.

3.2. 온톨로지 구조

온톨로지는 특정 분야에서 사용되는 어휘 사이의 의미적 매칭 관계와 관계도를 정의한다. 매칭관계는 단순매칭과 복합매칭으로 나뉘어 정의되어 있으며, 특히 복합매칭의 경우 적용되는 연산에 따라 병합관계와 분할관계로 구분하여 정의하고 있다.

관계도는 두 어휘가 이루고 있는 관계의 정도를 나타내는 값으로 0부터 1사이의 실수로 표현된다. 즉, 큰 관계도 값은 두 어휘 사이에 매칭이 선택될 가능성이 높다는 것을 나타낸다. 관계도는 스키마 매칭의 결과에 따라 증가하거나 감소하여 온톨로지를 점증적으로 갱신하는 역할을 한다.

소스 \ 타겟	Name	FirstName	LastName	SirName
Name	=	①분할:1.0	②분할:1.0	null
FirstName		=	null	null
LastName			=	③단순:0.7
SirName				=

그림 1. 도메인 온톨로지의 예

그림 1은 도메인 온톨로지의 구조를 개념적으로 표현한 예이다. 그림에서 ①과 ②는 Name과 (FirstName, LastName)이 1.0의 관계도를 갖는 분할연산이 적용되는 복합매칭임을 나타내고, ③은 SirName과 LastName은 0.7의 관계도를 갖는 단순매칭 관계임을 표시한다.

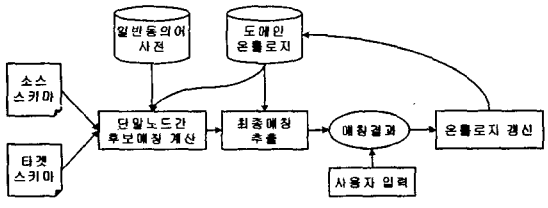

그림 2. 제안된 스키마 매칭 과정

4. 매칭 알고리즘

본 절에서는 매칭을 계산하는 과정을 구체적으로 설명한다. 그림 2와 같이 제안된 알고리즘은 크게 세 단계로 구성된다. 먼저, 요소가 포함하고 있는 이름과 데이터 타입에 기반하여 단말노드사이에 후보매칭을 계산하고, 계산된 후보매칭 사이의 경로유사도를 비교하여 최종 매칭을 추출한다. 끝으로 매칭결과를 이용하여 온톨로지를 갱신한다.

4.1. 단말노드간 후보매칭 계산

입력된 소스, 타겟 스키마가 스키마 트리로 표현이 되면, 단말노드를 비교하여 후보매칭을 계산한다. 단말노드 사이의 유사도 계산을 통하여 임계값 이상인 매칭을 후보매칭에 포함시킨다. 단말노드의 유사도는 식(1)과 같이 노드 레이블 사이의 언어적 유사도와 데이터 타입 유사도의 합으로 정의한다. 여기서 가중치 w_l과 w_t는 각각 언어적 유사도와 데이터 타입 유사도의 가중치를 의미한다. 언어적 유사도와 데이터 타입 유사도에 대한 자세한 설명은 다음과 같다.

$$\text{단말노드유사도}(N_s,N_t) = w_l * \text{언어적유사도}(N_s,N_t) + w_t * \text{데이터타입유사도}(N_s,N_t) \quad (1)$$

두 노드의 레이블 사이의 유사도를 나타내는 값으로 먼저 도메인 온톨로지에 두 어휘관계가 정의되어 있는지 확인한다. 도메인 온톨로지에 포함되어 있는 어휘라면 관계도를 두 노드의 언어적 유사도로 갖는다. 도메인 온톨로지에 포함되어 있지 않다면 대문자나 특수기호를 기준으로 토큰화 과정을 거쳐 각 토큰사이의 유사도를 계산한다. 최종적으로 두 노드의 언어적 유사도는 전체 토큰사이의 유사도의 평균으로 계산한다.

이때, 정확률의 향상을 위해 도메인 온톨로지 뿐만 아니라 일반 동의어 사전[5], 축약어 사전을 함께 활용한다. 두 어휘의 관계가 일반 동의어 사전을 통해 검색이 된다면 기본 유사도를 부여하고, 각 토큰이 축약어 사전에 정의되어 있다면 해당 토큰은 전체 이름으로 대치된다.

스키마 트리를 구성하는 단말노드는 다양한 종류의 데이터 타입을 값으로 갖는다. 특히 데이터 타입이 서로 다른 노드간의 변환은 노드가 포함하는 정보에 손실을 가져올 수 있다. 데이터타입 유사도는 이와 같이 서로 다른 데이터 타입을 갖는 노드간의 변환에 의하여 발생 가능한 정보 손실을 표현하기 위해서 제안되었다.

본 논문에서는 노드의 데이터 타입간의 유사한 정도를 해당 노드가 포함하는 정보의 손실 정도에 따라 '동일', '비손실 변환가능', '손실 변환가능', 그리고 '변환불가'의 4단계로 구분한다. 동일한 데이터 타입의 변환의 경우, 가장 큰 유사도를 부여하고, 정보 비손실 변환가능, 정보 손실 변환가능, 그리고 변환불가의 단계의 따라 점점 낮은 유사도를 부여한다.

4.2. 최종매칭 추출

본 절에서는 단말노드 매칭 과정을 통해 계산된 후보 매칭에서 경로유사도와 온톨로지를 이용하여 최종 매칭을 추출하는 과정을 설명한다. 경로유사도는 후보 매칭에 포함된 단말노드가 형성하고 있는 경로사이의 유사도로 단말노드의 구조적 정보를 반영한다.

경로유사도는 두 경로사이에 매칭되는 중간노드의 유사도의 평균으로 계산한다. 그러기 위해선 먼저 중간노드 사이의 매칭관계를 찾아야 한다. 두 중간노드의 유사도가 임계값 이상의 값을 보이는 경우 매칭되며 중간노드 N_s와 N_t간의 유사도는 다음의 식으로 계산한다.

$$\text{중간노드유사도}(N_s,N_t) = w_l * \text{언어적유사도}(N_s,N_t) + w_s * \text{구조적유사도}(N_s,N_t) \quad (2)$$

언어적 유사도는 단말노드 매칭과 동일하게 두 중간노드의 레이블 사이의 유사도를 나타내고, 구조적 유사도는 해당 중간노드의 서브트리의 매칭되는 단말노드의 비율로 나타낸다.

경로유사도는 매칭된 중간노드 사이의 유사도 평균으로 다음의 식으로 계산한다.

$$\text{경로유사도}(Ps,Pt) = \frac{\sum Ps\text{와 } Pt\text{사이에 대응관계를 갖는 중간노드의 유사도}}{|Ps|+|Pt|} \quad (3)$$

Ps:소스 경로, Pt:타겟 경로

후보매칭 집합에서 계산된 경로유사도를 비교하여 가장 높은

값을 보이는 매칭을 최종 매칭으로 선택한다. 가장 높은 경로 유사도를 보이는 매칭이 유일하면 그 매칭은 단순매칭으로 분류한다. 하지만 가장 높은 경로유사도를 보이는 매칭관계가 두 개 이상인 경우 도메인 온톨로지를 검색하여 복합 매칭관계를 확인한다. 매칭관계가 정의되어 있지 않거나 단순매칭으로만 정의되어 있다면, 단말노드 유사도를 비교하여 가장 높은 유사도를 나타내는 매칭을 선택한다.

4.3. 온톨로지 갱신

시스템에 의해 계산된 매칭은 완벽할 수 없기 때문에, 사용자에 의한 보정을 필요로 한다. 온톨로지의 갱신은 시스템의 매칭결과와 사용자의 보정결과를 분석하여 수행된다. 온톨로지 갱신의 경우는 다음과 같이 구분된다.

- 온톨로지에 어휘 추가: 시스템이 정확한 매칭을 찾거나, 시스템이 찾지 못한 매칭을 사용자가 후처리에 의해 추가한 경우, 두 어휘의 관계가 온톨로지에 정의되어 있지 않다면 두 어휘의 관계는 온톨로지에 추가된다. 특히 사용자가 복합매칭으로 연결을 시킨다면 온톨로지에는 적용된 복합매칭으로 기술된다.
- 온톨리지의 관계도 증가: 시스템이 정확한 매칭을 찾거나, 시스템이 찾지 못한 매칭을 사용자가 후처리에 의해 추가한 경우, 두 어휘의 관계가 온톨로지에 정의되어 있다면 해당 온톨로지의 관계도는 증가한다.
- 온톨로지의 관계도 감소: 시스템이 찾은 결과를 사용자가 후처리에서 제거하는 경우, 누적된 매칭 횟수를 감소시킴으로 그에 따른 관계도를 감소시킨다.

5. 실험 결과

제안한 알고리즘의 성능을 평가하기 부동산 거래를 위해 실제 사용되는 스키마를 이용하여 실험하였다. 실험데이터는 Doan의 실험에서 사용된 것으로 5개의 소스스키마와 1개의 타겟스키마로 복합매칭관계를 포함하고 있다.

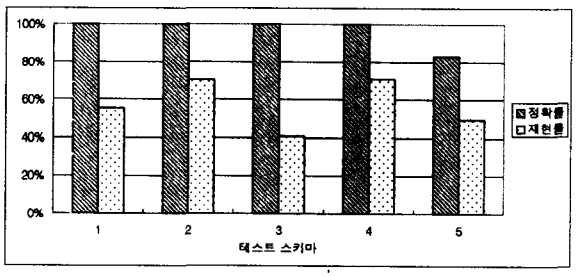

그림 3. 도메인 온톨로지를 적용하지 않은 실험결과

먼저 도메인 온톨로지 없이 실험을 실시하였다. 그림 3은 각 소스스키마에 대한 매칭결과를 정확률(precision)과 재현률(recall)로 나타낸 것이다. 평균 97%의 정확률과 57%의 재현률을 보인다. 재현률이 낮은 이유는 온톨로지가 정의되지 않아 복합매칭을 찾을 수 없기 때문이다. 또한 단순한 어휘사전만을 사용함에 따라, 다양한 어형의 변화를 고려하지 못하기 때문에 단말노드 사이의 후보매칭 계산과정에서 실패하는 경우가 많다. 예를 들면, bed와 bedrooms 같은 경우 일반 동의어 사전에서는 찾을 수 없는 관계로 언어적 유사도가 0이 된다.

도메인 온톨로지의 점증적 갱신을 이용한 실험결과는 그림 4와 같다. 소스스키마 1을 이용하여 매칭을 계산하고 사용자에 의해 잘못된 매칭과 누락된 매칭을 입력하게 하였다. 이때 사용자는 누락된 매칭이 단순매칭인지 복합매칭인지 구분하여 정의할 수 있다. 매칭결과 사용자의 후처리로 생성된 온톨로지는 스키마 2의 실험에 적용된다. 소스스키마 3, 4, 5에 대해서도 같은 방법으로 실험한다.

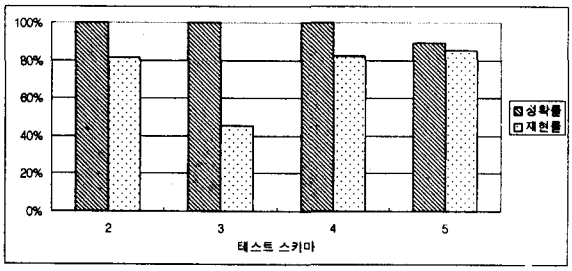

그림 4. 온톨로지의 점증적 갱신에 기반한 매칭결과

실험결과 재현률은 평균 73%로 온톨로지 없이 수행된 실험결과(57%)에 비해 향상됨을 확인할 수 있었다. 하지만 스키마 3에서 재현률이 여전히 낮은 이유는 사용된 어휘 구조가 이전에 수행된 스키마 1, 2와 다른 구조를 보여 온톨로지의 갱신에도 불구하고 많이 향상되지 못했다. 즉, 동일한 어휘로 정의되어 있지 않는 기존의 스키마 매칭 결과는 다음 매칭결과에 큰 도움을 주지 못한다.

제안된 방법은 모든 실험에서 평균 97%이상의 높은 정확률을 보였고, 온톨로지를 적용한 후 새롭게 찾아진 대부분의 매칭은 분할연산이 적용되는 복합매칭관계로 단순매칭과 복합매칭으로 구분하여 매칭을 확인할 수 있었다.

6. 결론 및 향후연구

본 논문에서는 온톨로지의 점증적 갱신을 이용한 스키마 매칭 방법을 제안하였다. 특히 제안된 방법은 온톨로지를 이용하여 단순매칭 및 복합매칭을 구분하여 계산할 수 있다. 이것은 문서변환시 적용해야할 연산을 판단할 수 있게 함으로 더욱 정확한 문서변환을 가능하게 한다. 제안된 방법의 평가를 위한 실험결과, 온톨로지를 점증적으로 갱신함에 따라 성능이 현저히 향상됨을 확인할 수 있었다. 또한 복합매칭 관계도 정확히 계산되었다.

하지만 일부 실험에서는 온톨로지를 적용하여도 성능에 많은 향상을 보이지 않는 경우도 있었다. 이것은 기존에 사용된 스키마와 입력된 스키마의 어휘가 동일한 형태가 아니면 온톨로지의 갱신에도 불구하고 성능이 향상되기 힘들기 때문이다. 따라서 향후 연구로 더욱 정교한 온톨로지의 구축과 갱신 방법을 개발할 것이다.

참고문헌

[1] Erhard Rahm and Philip A. Bernstein, "A survey of approaches to automatic schema matching," VLDB, vol. 10, issue 4, pp. 334-350, 2001.

[2] Jayant Madhavan, Philip A. Bernstein, Erhard Rahm, "Generic Schema Matching with Cupid," Proceedings of the 27th VLDB Conference, Roma, Italy, 2001.

[3] Li Xu, David W. Embley, "Discovering direct and indirect matches for schema elements," Proceedings. 8th Conference on DASFAA, pp. 39-46, 2003.

[4] Anhai Doan, "Learning to Map between Structured Representations of Data," Ph.D. Dissertation, 2002.

[5] George A. Miller, "WordNet: A Lexical Database for English," Communications of the ACM, Vol. 38, No. 11, pp. 39-41, 1995.