온톨로지 병합을 위한 의사지원트리 기반 충돌 탐지 및 해결 기법

Conflict Detection and Resolution Method for Merging of Ontologies based on Decision Support Tree

정현숙[1], 김정민[2], 이성주[1]

[1] 조선대학교 컴퓨터공학과
E-mail: {hsch,sjlee}@chosun.ac.kr
[2] 서울대학교 컴퓨터공학부
E-mail: jmkim@idb.snu.ac.kr

요 약

본 논문에서는 토픽맵 기반의 온톨로지 병합 과정에서 발생할 수 있는 충돌의 유형을 트리 구조로 정의하고 충돌 탐지 및 해결을 통하여 두 온톨로지를 하나로 병합하는 기법을 제안한다. 병합충돌은 의미적 대응 요소들의 유사값에 기반하여 엘리먼트기반, 구조기반, 임시기반의 트리 구조로 분류되고 이 충돌 트리를 이용하여 두 매핑 요소사이의 병합충돌을 탐지하고 해결한다. 실험을 위해 토픽맵 질의언어 tolog를 사용하여 동서양 철학온톨로지 및 독일 문학온톨로지들의 병합 전과 후의 질의 결과를 비교하고 이를 정확율과 재현율로 병합 성능을 평가하였으며 그 결과 손실없는 병합이 가능함을 보였다.

Key Words : Ontology merging, Taxonomy of merge conflicts, Conflict detection, Topicmaps

1. 서 론

현재의 온톨로지는 시맨틱웹 뿐만 아니라 콘텐츠 관리, 전자도서관, 지식 관리 등 다양한 분야에서 지식 색인 및 구조화 도구로 사용되고 있다. 토픽맵은 ISO 표준으로서 RDF/OWL과 함께 온톨로지를 표현하는 데이터 모델로 사용된다[1]. 본 논문에서는 토픽맵 기반의 온톨로지들 사이의 병합에서 발생하는 충돌의 유형을 정의하고 이를 탐지 및 해결함으로써 손실없는 온톨로지 병합 알고리즘을 제안한다.

온톨로지 병합은 두 소스 온톨로지 요소들의 합집합을 구하는 것과 유사하다. 즉, 두 온톨로지의 공통 요소를 찾은 다음 이들 사이의 중복을 제거하면서 두 소스 온톨로지의 모든 요소를 하나로 합치는 것이다. 그러나 관계형 데이터의 조인이나 합집합을 구하는 것처럼 단순한 과정이 아니다. 의미적으로 대응되는 공통 요소를 찾는 것, 공통 요소 사이의 중복을 제거하는 것, 병합 과정에서 발생하는 데이터의 충돌을 파악하고 해결하는 것 및 서로 다른 유형의 요소들을 효과적으로 병합하는 것 등 여러 문제점들을 고려해야 한다.

토픽맵 표준 문서인 TMRM(Topic Masps Reference Model)[2]과 TMDM(Topic Maps Data Model)[3] 에서는 요소들 사이의 유사성 기반이 아닌 완전하게 일치하는 요소들 사이의 병합 과정을 정의하고 있다. 그러나 대부분의 경우 서로 다른 전문가들에 의해 생성된 온톨로지들에서 개념명, 속성유형 및 속성값 등이 일치하는 요소들을 찾기는 어렵다. 따라서 유사값에 근거하여 상호 매핑되는 요소들 사이의 병합이 다루어져야 한다.

또한 온톨로지 병합을 다루는 이전의 연구들은 대부분 두 소스 온톨로지들 사이에 의미적으로 대응되는 공통 요소를 효과적으로 찾기 위한 온톨로지 매칭(ontology matching)에 집중되어 있으며 매핑 요소들을 병합하는 과정에서 발생하는 문제를 정의하고 해결하는 방법에 대해서는 간과하고 있다[4][5].

두 온톨로지로부터 의미적으로 대응되는 요소들을 찾는 매칭 기법은 본 연구의 선행연구[6]에서 수행되었으며 본 논문에서는 온톨로지 병합 과정에서 발생하는 매핑 요소들 사이의 병합 충돌을 정의한다. 먼저 유사값에 기반하여 병합 충돌 트리를 정의하고 이 트리를 기반으로 병합 충돌을 탐지 및 해결한다.

병합 충돌의 유형은 크게 온톨로지 요소 수

준에서의 상이한 값에 의한 충돌인 엘리먼트기반 충돌(element-level conflict), 개념화 수준의 차이에서 오는 구조기반 충돌(structure-level conflict), 그리고 병합 과정에서 일시적으로 발생하는 데이터의 불일치에 의한 임시적 충돌(temporal conflict)로 나누어진다.

2. 관련연구

PROMPT[7]는 온톨로지 매칭 및 병합 도구로서 사용자 대화 형식의 점진적인 병합 방식을 제공하는 특징을 가진다. 먼저, 문자열 기반 비교 기법에 따라 노드명이 완전히 일치하는 요소들 사이에 매핑 관계를 설정한 다음 매핑 요소들과 이들에게 적용 가능한 연산들을 사용자에게 보여준다. 사용자가 연산의 실행을 요구하면 두 매핑 요소에 연산을 적용한 다음 이로 인해 발생하는 충돌이 있을 경우 충돌 리스트를 사용자에게 보여주는 등 대화식 방식으로 하나씩 처리해 나간다.

Chimerae[8]는 대용량의 온톨로지를 병합하고 테스트하기 위한 환경을 제공한다. 이 시스템에서 온톨로지 매칭은 독립된 프로세스가 아니라 병합 연산자의 하위 타스크로서 실행된다. 병합할 후보를 탐색하는 과정에서 용어들 사이의 유사성을 계산하고 문자열 기반 매칭 기법에 의해 산출된 유사값에 따라 병합할 대상을 결정한다. PROMPT와 유사하게 사용자에게 탐색한 병합 후보들을 보여준 다음 선택에 따라 대화식으로 병합하는 방식을 가지고 있다.

Pottinger[9]는 데이터베이스 스키마, UML 모델, 온톨로지 모델 등을 병합할 수 있는 범용적인 알고리즘을 제안하고 있다. 범용 병합 요구조건(generic merge requirements)을 정의하고 있으며 충돌 유형 및 해결 기법들을 설명하고 있다. 그러나 상이한 모델들 사이의 범용적인 병합을 위해 자체적으로 정의한 E-R 모델로 변환함에 따라 토픽맵의 병합에 있어서는 토픽맵 모델의 특성을 반영하지 못하는 문제점을 가진다.

3. 토픽맵 병합 정의

토픽맵 병합의 목적은 두 토픽맵의 개체 집합에 대하여 중복을 제거한 합집합을 구하는 것이다. 정의 1에서는 토픽맵 병합 연산의 입력 및 출력 값과 토픽맵 개체들의 합집합 연산을 정의하고 있다.

정의 1(토픽맵 병합). 토픽맵 집합 S가 주어졌을 때 집합 S에 대한 병합 연산은 다음과 같이 정의된다.

$$merge(S \times S) \rightarrow S$$

두 토픽맵 $TM_A, TM_B \in S$의 병합은 다음과 같이 각 토픽맵의 요소들의 합집합으로 정의된다.

$$\textbf{\textit{merge}}(TM_A, TM_B, M_{AB}) = \{ \text{all entities of } TM_A \} \cup \{ \text{all entities of } TM_B \}$$

병합 연산의 입력 값은 두 토픽맵 TM_A, TM_B와 그들 사이의 매핑 값인 M_{AB}이다. M_{AB}는 두 토픽맵의 요소들 중에서 의미적으로 1-대-1의 대응관계를 가지는 두 토픽의 쌍들의 집합이다. 병합 연산의 출력 값은 두 토픽맵의 요소들의 합집합으로써 두 토픽맵과 별개의 새로운 토픽맵이다.

매핑 값 M_{AB}는 (a, b, SIM_{name}, SIM_{occ}, SIM_H, SIM_{assoc}, SIM)의 7-튜플로 정의된다[7]. 여기서, a와 b는 각각 두 토픽맵의 토픽이고 SIM 값들은 두 토픽의 매핑 정도를 가리키는 복합 유사값으로 [0..1] 범위의 값이다. S_{name}은 문자열 비교에 따라 산출된 값으로 두 토픽명이 어느 정도 유사성을 가지는지 가리키는 토픽명 기반 유사값, S_{occ}는 두 토픽의 속성들 사이에 속성타입 및 속성값이 어느 정도 유사성을 가지는지 가리키는 토픽속성 기반 유사값, S_H는 두 토픽의 계층구조에서 자식 토픽들이 어느 정도 유사성을 가지는지 가리키는 계층구조 기반 유사값, S_a는 연관관계 타입 토픽들 사이에 두 토픽의 멤버들과 역할이 어느 정도 유사성을 가지는지 가리키는 연관관계 기반 유사값이다. 그리고 SIM은 이들 4가지 유형의 유사값들을 조합하여 평균값으로 계산한 단일 유사값이다. 따라서 a와 b 두 토픽이 의미적으로 대응되는지 여부는 단일 유사값 SIM이 특정 기준값(threshold)을 초과하는지에 의해 결정된다.

4. 병합 충돌 트리 정의

4.1 병합 충돌 트리

의미적으로 대응관계에 있지만 표현 및 구조상의 차이로 인해 발생하는 병합 충돌은 그림 1과 같이 크게 엘리먼트 기반 충돌, 구조 기반 충돌 및 임시적 충돌로 나누어진다.

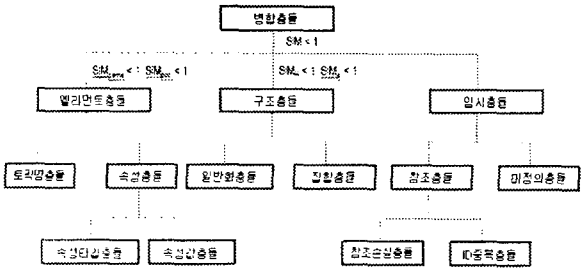

그림 1. 온톨로지 병합 충돌 트리

엘리먼트기반 충돌은 단일 토픽 자체에서 발생하는 충돌로서 토픽명의 차이로 인한 충돌 및

속성의 차이로 인한 충돌로 세분화된다. 속성 충돌은 속성타입 충돌과 속성 값 충돌로 나누어진다. 속성타입 충돌은 두 토픽의 속성 값은 같지만 속성타입이 다른 경우 발생하며 이와 반대로 속성 값 충돌은 속성타입은 같지만 속성 값이 다른 경우 발생한다.

구조기반 충돌은 두 매핑 토픽의 지식 표현 수준의 차이에서 비롯되는 충돌이다. 예를 들어, 두 매핑 토픽이 동일한 토픽명 Philosopher로 정의되어 있지만 토픽맵 A의 Philosopher는 그 하위에 동서양 및 시대적 구분 없이 Kant, Hegel, Mencius 등 철학자 인스턴스 토픽들로 연결되어 있는 반면 토픽맵 B의 Philosopher는 그 하위에 시대적으로 Ancient Philosopher, Medieval Philosopher, Modern Philosopher 등으로 분류한 다음 다시 각각의 시대 분류 토픽 하위에 철학자 인스턴스 토픽을 연결한 경우 두 Philosopher 토픽은 구조적 충돌을 가지는 것이다.

임시적 충돌은 토픽맵 병합 과정에서 발생하는 충돌로서 참조 충돌(reference conflict)과 미정의 요소 충돌(undefined element conflict)로 세분화된다. 참조 충돌은 토픽들 사이의 참조 값의 부정확에 의해 발생하는 충돌로서 참조 손실 충돌은 상위 토픽이 다른 토픽맵의 매핑 토픽과 병합됨으로 인하여 하위 토픽이 가지는 상위 토픽으로의 참조가 손실되는 경우이고 ID 중복 충돌은 병합으로 인해 동일한 ID를 가지는 토픽이 다수 발생하는 경우이다. 미정의 요소 충돌은 병합 전 토픽의 속성 타입, 연관관계 타입, 역할 타입이 병합으로 인해 존재하지 않는 타입이 되는 경우이다.

4.2 병합 충돌 탐지 및 해결

4.2.1 토픽명 충돌

토픽명 충돌은 두 매핑 토픽의 토픽명이 일치하지 않은 경우에 발생하므로 토픽맵 매핑 단계에서 산출된 토픽명 비교 연산의 결과인 SIM_{name} 유사값을 검사함으로써 토픽명 충돌 여부를 판단할 수 있다.

정의 2(토픽명 충돌 탐지). 두 매핑 토픽 t_a, t_b와 매핑 행렬 M 이 주어졌을 때,
1) $SIM_{name}(t_a, t_b) = 1$, 토픽명 일치
2) $SIM_{name}(t_a, t_b) < 1$, 토픽명 상이(충돌 발생)
2-1) 토픽명의 포함관계 존재. t_a.Name $\subset$ t_b.Name 또는 t_a.Name $\supset$ t_b.Name(부분 충돌)
2-2) 토픽명의 포함관계 없음(완전 충돌).

토픽명 충돌이 발생할 경우 해결 방법은 두 토픽명 사이에 한 토픽명이 다른 토픽명을 부분 문자열(substring)로 포함하고 있느냐에 따라 달라진다. 포함관계가 있을 경우 병합 토픽의 토픽명은 두 토픽명 중에서 다른 토픽명을 포함하는 토픽명으로 설정된다. 포함관계가 없는 경우는 완전 충돌이 발생한 경우로서 이 경우 시스템에서 자동적으로 우선되는 토픽명을 결정할 수 없으므로 병합 토픽의 토픽명으로 두 토픽명을 모두 지정한다.

4.2.2 속성 충돌

SIM_{occ}가 1인 경우 두 토픽은 같은 속성타입과 같은 속성 값을 가지는 것으로 다중 속성들이 완전히 일치함을 가리킨다.

정의 3(속성 충돌 탐지). 두 매핑 토픽 t_a, t_b와 매핑 행렬 M 이 주어졌을 때,
1) $SIM_{occ}(t_a, t_b) = 1$, 속성 일치
2) $SIM_{occ}(t_a, t_b) < 1$, 속성 상이(충돌 발생)
2-1) t_a.OccType <> t_b.OccType and t_a.OccVal = t_b.OccVal(속성타입 충돌)
2-2) t_a.OccType = t_b.OccType and t_a.OccVal <> t_b.OccVal(속성 값 충돌).

두 매핑 토픽의 $SIM_{occ} < 1$인 경우 각각 토픽의 다중 속성들 사이에 쌍을 지어 속성타입과 속성 값을 비교하여 속성타입 충돌 또는 속성 값 충돌이 발생했는지 탐지한다.

4.2.3 일반화 및 집합 충돌

일반화 충돌은 두 매핑 토픽이 서로 다른 수준의 개념화를 가질 때 발생한다. 따라서 이 충돌을 탐지하기 위해서는 두 매핑 토픽의 하위 토픽들이 동일한 트리 레벨에서 매핑되는지를 판단해야 한다.

집합 충돌은 두 매핑 토픽이 서로 다른 범위의 하위 토픽들을 포함하는 경우에 발생한다. 예를 들어 TM_1의 Philosopher 토픽은 하위에 Korean Philosopher, Chinese Philosopher만 가진 반면 TM_2의 Philosopher 토픽은 Korean Philosopher, Chinese Philosopher, Indian Philosopher, Western Ancient Philosopher 등 더 많은 종류의 철학자 분류를 가지는 경우이다. 일반화 충돌과 집합 충돌은 병합보다는 매핑 단계에서 해결된다. 즉, 구조적 매칭 기법에서 하위 토픽들 사이의 매칭 여부를 상위 토픽의 구조적 유사성에 반영하기 때문에 하위 토픽들 사이에 겹치는 부분이 많을수록 상위 토픽의 유사성이 높아진다.

4.2.4 임시 충돌

임시 충돌의 탐지는 병합 토픽의 속성, 상위 토픽, 연관관계, 역할 타입 등이 병합 토픽맵에 정의되어 있는지 여부를 검사함으로써 가능하다.

정의 4(임시 충돌 탐지). 병합 토픽맵 TM_C와 병합 토픽 t_C가 주어졌을 때,
1) t_c.OccType$_k$ $\notin$ T_o:{OccType$_i$|1≤ i ≤ n}(속성타입 미정의 충돌)
2) t_c.AssocType$_k$ $\notin$ R_a:{AssocType$_i$|1≤ i ≤ m}(연관관계타입 미정의 충돌)
3) t_c.RoleType$_k$ $\notin$ T_r:{RoleType$_i$|1≤ i ≤ l}(역할타입 미정의 충돌)

4) $t_c.TopicType_k \notin T_c.\{TopicType_i | 1 \leq i \leq p\}$(참조 충돌)

5. 병합 실험 및 평가

실험데이터는 전문가 집단에 의해 생성된 동서양 철학온톨로지와 야후 백과사전에서 실험을 위해 생성한 근대 및 현대 철학온톨로지, 독일 문학온톨로지 등을 사용한다.

병합 모듈의 성능을 평가하기 위해 실험 데이터의 각 토픽맵 쌍에 대하여 토픽맵 검색어인 tolog 질의어를 이용한 병합 전과 후의 검색 결과를 비교하였다. 예를 들어 (T_1,T_2) 토픽맵 쌍에 대해서 T_1 토픽맵에서만 검색될 수 있는 질의어, T_2 토픽맵에서만 결과를 검색할 수 있는 질의어, T_1과 T_2의 양쪽 토픽맵에서 결과를 검색할 수 있는 질의어를 실행하여 그 결과를 비교한다.

성능 평가 척도는 정보검색에서 사용하는 정답율(precision)과 재현율(recall)을 사용한다. 병합 전의 두 소스 토픽맵에서의 검색 결과와 병합 토픽맵에서의 검색 결과를 비교하여 정답율과 재현율을 평가해 봄으로써 병합 토픽맵이 정보의 손실없이 두 소스 토픽맵을 완전히 병합한 것인지 여부를 확인할 수 있다. 아래 정답율과 재현율을 구하는 수식에서 **P**는 두 소스 토픽맵으로 부터 검색된 결과 집합이고 **R**은 병합 토픽맵으로 부터 검색된 결과 집합이다. 그리고 **I**는 **P**와 **R**의 교집합이다. 그림 2는 각 토픽맵쌍의 정답율과 재현율을 보이는 그래프이다. 전체적으로 90% 이상의 정답율과 재현율을 보이고 있으며 병합 토픽맵이 소스 토픽맵을 손실없이 병합함을 보이고 있다.

$$precision = \frac{I}{R} \qquad recall = \frac{I}{P}$$

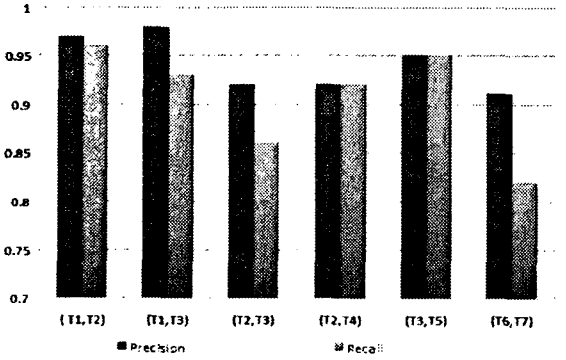

그림 2. 병합 성능 평가 그래프

두 소스 토픽맵 사이에 매핑되는 토픽이 많을 경우 재현율이 낮게 나오고 있다. 그 이유는 두 소스 토픽맵의 매핑 토픽들이 병합 토픽맵에서는 하나의 토픽으로 병합되기 때문에 동일한 질의어에 대해 병합 토픽맵의 질의 결과 수가 더 적게 나오기 때문이다.

6. 결론 및 향후연구

본 논문에서는 온톨로지 병합 과정에서 발생 가능한 매핑 요소 사이의 충돌 유형을 정의하고 매핑 유사값에 기반하여 충돌을 탐지 및 해결하는 기법을 제안한다. 온톨로지 병합 연산은 두 온톨로지 요소들의 합집합을 구하는 것으로 이 과정에서 충돌 탐지와 해결을 내부적으로 처리하며 중복을 배제한 병합 온톨로지를 생성한다.

토픽맵으로 구현된 여러 철학온톨로지를 실험데이터로 하여 병합 모듈의 성능을 평가한 결과 수작업에 비해 짧은 실행시간에 손실없이 병합할 수 있음을 보였다. 향후에는 일대일 매핑을 넘어서 다중매핑 요소들 사이의 병합처리도 가능하도록 연구하고자 한다.

참 고 문 헌

[1] M. Biezunski, M. Bryan, and S. Newcomb, ISO/IEC 13250 TopicMaps, 2002.

[2] P. Durusau, S. Newcomb, and R. Barta, Topic Maps - Reference Model, ISO/IEC JTC1/SC34, Version 6.0, http://www.isotopicmaps.org/tmrm, 2006.

[3] L.M. Garshol and G. Moore, Topic Maps - Data Model, ISO/IEC JTC1/SC34, http://www.isotopicmaps.org/sam/sam-model, 2006.

[4] P. Shvaiko and J. Euzenat. "A survey of schema-based matching approaches", University of Trento, Technical Report #DIT-04-087, 2004.

[5] F. Giunchiglia and P. Shvaiko. "Semantic matching", In The Knowledge Engineering Review Journal, 18(3), 2004.

[6] 김정민, 신효필, 김형주. "분산 토픽맵의 다중전략 매핑 기법", 정보과학회논문지(소프트웨어 및 응용), 게재예정.

[7] N. Noy and M. Musen. "PROMPT: Algorithm and Tool for Automated Ontology Merging and Alignment", In Proceedings of the National Conference on Artificial Intelligence(AAAI), 2000.

[8] D. L. McGuinness, R. Fikes, J. Rice, and S. Wilder. "An environment for merging and testing large ontologies", In Principles of Knowledge Representation and Reasoning: Proceedings of the Seventh International Conference(KR2000), 2000.

[9] R. A. Pottinger and P. A. Bernstein. "Merging Models Based on Given Correspondences", In Proceedings of the 29th VLDB Conference, Berlin, Germany 2003.