토픽맵 기반 온톨로지 시스템의 통합효과 측정을 위한 프로토타입 시스템 구축 및 평가에 관한 연구

Developing and Evaluating an prototype system for merging effects of ontology systems : Based on Topic Maps

도진국, 성균관대학교 문헌정보학과, jinkook@skku.edu
양선화, 성균관대학교 문헌정보학과, hanaby2002@hanmail.net

Jin-Guk Do, Dept. of Library and Information Science, Sungkyunkwan University
Seon-Hwa Yang, Dept. of Library and Information Science, Sungkyunkwan University

본 논문은 토픽맵 기반 온톨로지 시스템의 통합효과 측정을 위한 연구에 앞서 통합의 가능성과 통합 성능을 측정하기 위한 프로토타입 시스템 구축에 관한 연구이다. 프로토타입 시스템 구축을 통해 자동 통합 툴의 성능을 측정하고자 한다. 이를 위해 통합 전의 단일 토픽맵에서의 검색 결과와 통합 토픽맵에서의 검색 결과를 비교하여 정답율과 재현율을 평가함으로써 통합 토픽맵이 정보의 손실 없이 단일 토픽맵들을 완전히 통합한 것인지 확인할 수 있다.

1. 서론

최근 인터넷 또는 기업과 같은 조직 활동을 통하여 생산되고 유통되는 정보의 양이 기하급수적으로 늘어남에 따라 필요한 정보를 찾기 위하여 검색엔진을 이용하거나 조직 내의 지식관리시스템 등의 지식공유시스템을 이용한다. 그러나 기존에 구축되어 있는 검색시스템을 이용하여도 의도한 검색결과를 얻기 힘든 경우가 많다. 이는 단순히 키워드 매칭에 기반한 검색 패러다임이 더 이상 이용자들의 검색욕구를 충족시키지 못 함을 보여주는 것이다.

웹의 급속한 팽창으로 인한 검색 대상 범위의 확대는 보다 정교한 검색을 필요로 하게 되었으며, 지능화된 정보 검색 패러다임을 요구하고 있다. 이에 대한 해결책으로 최신 정보기술의 동향은 온톨로지, RDF/OWL, 토픽맵과 같은 시맨틱 웹 관련 기술을 제시하고 있다. 시맨틱 웹은 컴퓨터간의 정보 교환을 가능하게 하며 웹상에 존재하는 데이터의 의미를 컴퓨터가 이해하여 처리할 수 있도록 하는 새로운 정보기술로써 웹(WWW)의 창시자인 팀 버너스 리(Tim Berners-lee)가 고안하였다. 사람만이 웹에 산재한 정보의 의미를 파악하는 것이 아닌, 컴퓨터가 해석할 수 있는 일종의 지능적인 웹이 시맨틱 웹의 목적이다.

시맨틱 웹 구축을 위한 핵심 요소로 사용되는 온톨로지(Ontology)는 복잡한 개념들을 정의하고 의미적으로 연결할 수 있는 기술이다. 다르게 표현하면, 세상에 존재하는 모든 개념들을 정의하고 개념들 사이의 계층관계와 의미관계를 잘 정리해 놓은 사전 정도로 이해할 수 있다. 이렇게 정의되고 연결된 지식표현은 기계들이 이해할 수 있도록 표현되기 때문에 인간은 컴퓨터를 이용하여 기존의 웹에서 경험하지 못 한 질 높은 검색결과를 기대할 수 있다.

인터넷 상의 분산되어 있는 웹페이지들처럼 온톨로지 역시 기관별로 독립적으로 구축하여 사용되고 있어 의미적으로 대응되는 온톨로지들은 연결하고 통합하여 사용할 필요가 있다. 개방 및 분산 환경인 웹에서 모든 지식을 포함하는 하나의 전능한 단일 온톨로지를 생성하는 것은 비현실적이므로 대신 분야별로 세분화된 여러 온톨로지들을 단계적으로 생성하고 이들을 연결하고 통합하는 것이 현실적인 시맨틱 웹 구축의 방안이다(T. Berners-Lee 외, 2001).

이에 의미적으로 대응되는 다수의 온톨로지를 통합하여 하나의 새로운 온톨로지를 생성하는 것을 온톨로지 통합이라 한다. 온톨로지 통합은 온톨로지와 관련된 주요한 연구 중의 하나로 간주되고 있다. 온톨로지 통합과 관련된 연구는 전산학 분야의 온톨로지 매핑과 통합의 성능을 높이기 위한 방법론이 주를 이루고 있다. 반면 (문헌)정보학 분야에서 온톨로지 통합과 관련된 연구는 그 수가 적으며, 특히 단일 온톨로지 시스템을 통합하여 이용자를 대상으로 통합 시스템의 검색효율성이나 이용자의 만족도를 측정한 연구는 찾기 힘들었다.

이에 본 연구에서는 통합효과 측정을 위한 온톨로지 통합 시스템의 구축에 앞서 온톨로지 구축에 널리 사용되고 있는 국제표준 온톨로지 언어인 ISO 토픽맵(Topic Maps)를 이용하여 프로토타입 시스템을 구축해봄으로써 통합의 가능성과 자동 통합툴의 성능을 측정하고자 한다.

2. 이론적 배경

2.1 온톨로지 통합

온톨로지들이 분산된 환경에서는 온톨로지들 사이의 상호 연결, 재사용 및 지식 공유의 필요성이 대두되며 이러한 문제를 해결하기 위한 연구들이 수행되어 왔다. 온톨로지 통합에 관한 연구는 과거 관계형 데이터베이스나 객체지향 데이터베이스 및 XML 데이터베이스 등의 '스키마 통합'에 대한 연구에서 영향을 받았으며 많은 부분 이들 연구와 유사성을 가진다(김정민 외, 2006).

그러나 대부분 스키마 통합 및 온톨로지 통합에 관한 연구에서는 통합을 위한 다양한 기법들의 제시에 초점이 맞추어져 있으며 현실적으로 적용이 가능한가에 대해서는 고려하지 않고 있다. 대부분 범용성을 지원하기 위해 특정 데이터 모델을 그래프 모델로 변환한 다음 두 그래프의 노드와 에지 사이의 매핑을 계산하는 알고리즘을 제시하고 있다. 그러나 현재 온톨로지를 표현하는 데이터 모델이 RDF/OWL과 토픽맵(Topic Maps)로 표준화되고 있고 이들 모델은 온톨로지 표현을 위한 구문(syntax), 의미(semantic), 제약조건(constraints)을 가지고 있으며 이러한 특성을 매핑 계산 시에 고려함으로서 그래프 모델의 노드와 에지들 사이에 노드명이나 경로 등의 단순 비교에서 찾을 수 없는 매핑 규칙들을 필요로 하고 있다.

2.2 토픽맵(Topic Maps)

토픽맵은 지식구조를 표현하고 지식구조와 정보자원을 연결하기 위해 만들어진 새로운 ISO 표준이다. 2000년 ISO/IEC 13250으로 발표되었고, 2001년 비정규기관인 Topicmaps.org에 의하여 XML Topic Maps(XTM) 1.0 표준규격이 발표 되었다. 2002년 5월에 두 번째 ISO/IEC 13250 개정판이 발표된 상태이다. 현재는 XTM2.0을 주요 토픽맵 구문으로 사용하고 있다.

토픽맵과 토픽맵간의 유사한 토픽에 대한 통합(merge)하는 방법은 표준사양에서 제공하는 <mergeMap> 요소에 의하여 쉽게 적용할 수 있으며 상당히 유연하다. 동일한 의미를 가진 토픽은 자동적으로 단일화 된 토픽으로 통합가 가능하다. 토픽맵 간의 개념 또는 의미 분석 결과에 따라 매핑규칙을 적용함으로써 유사한 토픽들을 통합하는 것을 가능하게 한다.

통합에는 토픽맵 통합과 토픽의 통합이 있고 통합은 <mergeMap> 으로 이루어지거나 응용프로그램의 필요에 의해 이루어진다. 두 토픽맵이 통합될 때 주제 식별자가 같거나 룰을 따르면, 같은 기본 이름을 가진 토픽은 통합 된다. 두 토픽이 통합될 때 이들 토픽의 특징(이름 , 어커런스, 관계)들은 중복이 제거된 상태에서 합쳐져서 하나의 토픽이 생성 된다. 이름들은 비트(bit) 단위로 비교되고, 어커런스는 같은 클래스의 인스턴스이고 같은 자원을 참조하는지의 여부가 비교되며, 관계는 같은 클래스 같은 역할, 동일한 역할을 수행하는 토픽인지가 비교된다.

<mergeMap>은 특정 토픽맵을 <mergeMap>을 지닌 토픽맵과 통합한다. 그리고 <mergeMap>은 이 특정 토픽맵의 모든 특성에 대한 범위 집합(scope set)에 자원참조(resource reference), 주제 식별자, 그리고 토픽맵을 추가할 수 있다. 또한 통합 할 경우 기존 토픽맵 그래프를 하나의 Scope로 구성하여 추후 원래의 토픽맵만을 검색하고자 할 경우에는 Scope기반으로 쉽게 필터링이 가능하게 된다.(XTM 1.0 Specification, 2001)

3. 본론

3.1 실험의 설계

실제 토픽맵 기반 온톨로지 시스템의 구축은 Ontopia의 토픽맵 편집기인 온톨폴리(Ontopoly)를 사용하여 구축하였다. 그리고 토픽맵 브라우저인 옴니게이터(Omnigator)에 내장된 통합(Merge) 기능을 이용하여 구축된 단일 온톨로지 시스템들을 통합하였다. 성능측정을 위한 검색기는 옴니게이터(Omnigator)를 활용하였다.

가. 도메인의 선정

단일 온톨로지 시스템을 직접 구축하여 통합 성능을 평가하기 위해 일반적이고 연구자에게 익숙한 도메인을 선정하였다. 토픽과 연계가 풍부한 '여행'을 도메인으로 선정하였다.

나. PSI 정책

본 연구에서는 공인 식별자 대신에 자체 식별자 체계를 구축하여 사용하였다. PSI Identifier를 "http://psi.skku.edu/terms/"로 정의하고 terms 디렉토리 아래에 정의되는 토픽타입, 어소시에이션타입, 어커런스타입의 name을 등록하였다. 단어는 모두 소문자로 표기하되 합성어는 "_" 기호로 연결하였다.

다. 단일 온돌로지 시스템 구축

구축한 두 개의 단일 온톨로지 시스템의 토픽타입, 연관관계타입, 어커런스타입의 통계치는 아래의 표와 같다.

<표 1> 단일온톨로지 시스템의 구조적특징

온톨로지	A	B
토픽타입수	300	276
연관관계타입수	884	813
어커런스타입수	2713	2531

라. 온톨로지 시스템 통합

통합방향의 결정에 있어서는 온톨로지 크기가 작은 쪽에서 큰쪽으로 정하였다. 단일 온톨로지들의 통합 결과 통계치는 아래의 표와 같다.

<표 2> 단일온톨로지 시스템의 통합결과

통합 온톨로지(A,B)		
통합결과	토픽타입수	631
	연관관계타입수	1250
	어커런스타입수	3056

3.2 실험 결과 분석

병합된 시스템의 통합 성능 측정을 위해 단일 토픽맵에서의 검색결과와 통합 토픽맵에서의 검색 결과를 비교하여 정답율과 재현율을 평가해 봄으로써 통합 토픽맵이 정보의 손실 없이 단일 토픽맵을 완전히 통합한 것인지 확인해 보았다.

재현율과, 정확율을 측정하기 위한 수식은 아래와 같다(김정민 외, 2006).

$$\text{정확율} = \frac{I}{R} \qquad \text{재현율} = \frac{I}{P}$$

<그림 1> 통합성능 측정을 위한 수식

P는 단일 토픽맵들로부터 검색된 결과 집합이고 R은 통합 토픽맵으로부터 검색된 결과 집합이다. 그리고 I는 P와 R의 교집합이다. 정확율과 재현율이 1에 가까울수록 통합 토픽맵이 단일 토픽맵을 손실 없이 통합한 것으로 본다(김정민 외, 2006).

단일 토픽맵과 통합 토픽맵에 대하여 질의어를 이용한 통합 전과 후의 검색 결과를 비교하였다. 10개의 질의어를 작성하였다. 예들어, A온톨로지에서만 검색될 수 있는 질의어, B온톨로지에서만 검색될 수 있는 질의어, A,B 양쪽에서 검색될 수 있는 질의어를 포함하였다.

4. 결론 및 제언

실험의 결과 전체적으로 90% 이상의 정확율과 재현율을 보였다. 이는 온톨로지 표준 기술언어인 토픽맵에 기반한 Ontopia의 토픽맵 통합 툴이 통합의 성능이 우수하다는 것을 말해준다. 그리고 추후 통합시스템의 검색 효율성과 이용자 만족도 측정을 위한 연구에서 앞서 전제 단계인 자동 통합 툴이 단일 시스템을 제대로 통합할 수 있음을 확인할 수 있었다.

참고문헌

고영만. 2006. 시소러스 기반 온톨로지에 관한 연구. 『정보관리』, 성균관대학교 정보관리연구소 제5집 : 5-22.

오삼균, 김홍기 외. 2006. 국가지식정보 온톨로지 표준개발. 학술정보문화진흥원 연구과제.

김정민, 신효필, 김형주. 2006. T-MERGE 연산자에 기반한 분산 토픽맵의 자동통합. 『정보과학회지:소프트웨어 및 응용』, 33(9) : 787-801.

T. Berners-Lee, J. Hendler, O. Lassila. 2001. "The Semantic Web" Scientific America, 279.

S. Pepper. 2002. The TAO of Topic Maps. Ontopia. [Online].

S. Pepper, LM. Garshol. 2002. The XML Papers: Lessons on Applying Topic Maps. Ontopia. [Online].