메타데이터 기반 정보시스템간 의미 유사도 측정 방법

임정은[1] 최오훈[1] 나홍석[2] 백두권[1]
[1]고려대학교 정보통신대학 컴퓨터학과, [2]한국디지털대학교 디지털정보학과
{jelim[O], pens}@korea.ac.kr, hsna99@kdu.edu, dkbaik@korea.ac.kr

A Methodology for Semantic Similarity Measurement among Metadata based Information System

Jung-Eun Lim[1], O-Hoon Choi[1], Hong-Seok Na[2], Doo-Kwon Baik[1]
[1]Department of Computer Science and Engineering,
College of Information and Communications, Korea University
[2]Dept. of Information and Computer Science, Korea Digital University

요 약

특정 도메인의 정보시스템간에 정보를 공유하기 위해서, 정보 시스템들은 도메인별로 사용되는 메타데이터를 각기 정의하여 사용하기 때문에 각각의 정보 시스템간의 정보 공유시 메타데이터의 이질성 문제가 발생되지 않는다. 그러나, 메타데이터의 불일치 문제는 이기종 도메인간에 정보를 공유할때 발생된다. 본 논문에서는 메타데이터를 이용하여 구축된 정보시스템 간의 상호운용성을 증진하기 위하여 메타데이터의 의미적 유사성 측정 방법을 제안한다. 이를 위하여 메타데이터 레지스트리(MDR)에 정의되어 있는 메타데이터에 대한 개념 모델을 정의하고, 개념모델의 인스턴스간에 의미유사성을 측정하는 방법을 제안한다. 제안한 방법을 사용한 결과 도메인이 다른 정보시스템간에 정보공유를 위한 의미적으로 유사한 최적의 메타데이터를 선택할 수 있다.

1. 서 론

이기종 정보시스템간의 정보공유뿐만 아니라 이질적인 도메인에 대한 정보공유에서 각각의 도메인에 따른 다양한 정보시스템간의 이질성이 발생되고 있다. 종류로는 시스템 이질성, 구조적 이질성, 도메인 불일치, 의미적 이질성(동음이의어, 동의어)등이다. 이러한 이질성을 해결하기 위한 기존의 연구는 구조적 이질성을 해결하는 것이었다. 그러나, 구조적 이질성을 해결하였음에도 여전히 이질성은 존재한다. 이러한 이질성의 근본적인 원인은 정보시스템들에 존재하는 메타데이터가 서로 다르게 정의되어 의미적 불일치(동음이의어, 동의어)를 야기한다는 것이다.

현재의 정보시스템에서 메타데이터의 의미(semantics)는 단일 정보시스템 내에서만 식별 가능하므로, 의미적 상호운용성 증진을 위해서는 정보시스템들의 메타데이터간에 의미를 식별할 수 있는 방법이 필요하다. 대표적인 정보시스템인 미국 환경청 MDR인 EDR(Environmental Data Registry)[1]에서 메타데이터는 10,000건 이상이며, 한국전자거래진흥원에서 만든 REMCO시스템[2]의 경우 2,000건 이상의 전자상거래관련 메타데이터가 존재하므로, 사람이 직접 도메인간에 의미적으로 비슷한 메타데이터를 파악하기가 어렵다. 따라서, 서로다른 도메인의 정보시스템들간 수많은 메타데이터에 대한 정보 공유를 위해서 사람이 판단하기 전에 의미적으로 유사한 메타데이터를 선정하는 작업이 필요하다.

본 논문에서는 정보시스템간에 의미적으로 유사한 요소끼리 정보공유를 하도록 메타데이터간의 유사성을 측정할 수 있는 방법을 제안하였다. 이 방법은 정보시스템이 가지고 있는 메타데이터에 대한 개념모델을 정의하고, 메타데이터가 가진 속성값과 워드넷을 이용해서 개념모델의 인스턴스를 생성하여, 정보공유가 필요한 메타데이터의 인스턴스간에 유사성을 측정한다. 논문의 구성은 다음과 같다. 2장은 관련연구, 3장은 메타데이터의 개념모델(Concept Model), 4장에서는 의미검색을 위한 유사성 측정 방법, 5장을 끝으로 결론을 맺는다.

2. 관련 연구

2.1 MDR(Metadata Registry)시스템

메타데이터 레지스트리(이하 MDR)[3]은 ISO/IEC 11179 표준을 적용함으로써 표준화된 메타데이터를 저장하고, 관리한다. ISO/IEC11179 표준은 데이터베이스들의 상호운용성을 높이기 위해 ISO/IEC JTC1/SC32 WG2에 의해서 개발되었다. ISO/IEC 11179에서 메타데이터는 *객체 클래스(Object class), 속성(Property), 표현(Representation)*의 3 부분으로 구성되며, 이 세 요소가 모여서 하나의 *데이터 요소*를 만들게 된다. 이 논문에서 *데이터 요소 개념*은 의미적 유사성을 측정하기 위한 중요한 요소로 사용된다.

데이터 요소 개념 = 객체 클래스 + 속성
데이터 요소 = 데이터 요소 개념 + [표현]

*객체 클래스(Object Class)*는 메타데이터를 갖게 되는

대상을 의미한다. 명확한 범위와 의미 그리고 동일한 규칙을 따르는 특성과 행위를 가짐으로써 식별되는 실세계의 생각, 추상 또는 사물을 뜻한다. 객체 클래스의 이름은 개념 영역(conceptual domain), 데이터 요소 개념(data element concept), 데이터 요소(data element)를 구성하는 한 요소로 사용된다. 객체 클래스 용어는 그 자체로 개념 영역의 이름으로도 사용 가능하다.

속성(*Property*)은 하나의 객체 클래스가 공통으로 갖는 특성을 말한다. 특성은 각각의 이름을 갖는다. 데이터 요소 항목명은 객체 클래스 용어와 특성 용어를 합쳐서 구성될 수 있다. 특성은 독립적으로 또는 복합적으로 데이터 요소의 이름으로 사용된다.

2.2 메타데이터를 이용한 정보공유방법

정보시스템에서 기존의 정보검색은 주로 메타데이터의 속성인 이름과 정의만을 검색키워드와 매칭시키는 방법을 이용하였다. 정보시스템 A에 '회사_이름', '학교_명', 정보시스템 B에 '법인_명', '학교_이름', '사람_이름', '프로젝트_명'등 메타데이터가 존재한다고 가정했을 때,

1) 기존의 방법으로 키워드 '회사_이름'으로 정보시스템 B의 메타데이터를 검색하면 "no result"라는 결과를 얻을 것이다. 사실 '회사_이름'과 '법인_명'은 유사한 의미를 가지는 메타데이터이지만, 매칭되는 키워드가 없으므로 검색되지 않는다.

2) 키워드 '이름'으로 정보시스템 B의 메타데이터를 검색하면, '이름'이 포함된 모든 메타데이터가 결과로 반환된다. 그러나, 메타데이터 '학교_이름', '사람_이름'은 '이름'이라는 속성은 같지만, 메타데이터가 개발된 도메인에 따라서 객체 클래스가 달라지기 때문에, 단순히 메타데이터 속성인 이름과 정의에 대한 검색 키워드 매칭으로는 의미까지 고려하기에 상당히 제한적이다.

따라서, 정보공유 할 메타데이터를 구분짓는 주요 특징으로 객체 클래스, 속성등을 활용한다면 의미를 고려한 메타데이터의 검색이 될 것이다.

3. 의미 상호운용성을 위한 개념 모델(Concept Model)

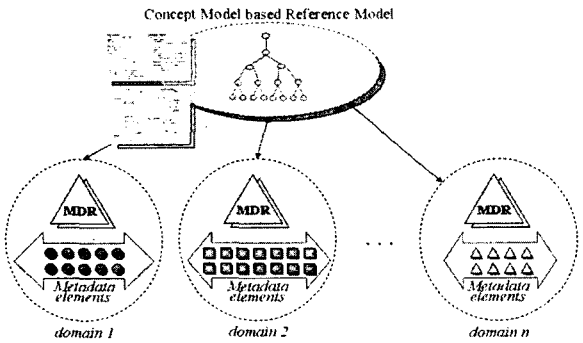

그림 1 메타데이터 기반 정보시스템간의 개념 모델

그림 1은 의미상호운용성을 위한 일반적인 아키텍쳐이다. 본 논문에서는 그림 1과 같이 정보시스템에 정의된 메타데이터에 대한 개념모델을 정의하고, 메타데이터가 가진 속성값과 워드넷을 이용해서 개념모델의 인스턴스를 생성하여 참조모델로 두고, 정보공유가 필요한 메타데이터의 인스턴스간에 유사성을 측정한다.

3.1 개념 모델

우리는 MDR과 WordNet을 이용하여 메타데이터에 대한 개념 모델을 정의하였다. 이것을 의미 유사성 측정에 활용하도록 한다. 개념 모델(*Concept Model*)은 아래와 같이 구성된다.

Concept Model := (DEC, Property, N_set, Context)

- *DEC* : 메타데이터 이름의 동의어집합.
- *Property* :메타데이터에서 속성의 동의어집합.
- *N_set* : 비교하고자하는 메타데이터의 객체클래스(Object class)의 동의어집합과 해당 객체클래스와 의미적 이웃관계에 속하는 객체클래스들의 집합. 의미적 이웃이란 온톨로지의 계층 구조 상에서 특정한 메타데이터의 *객체클래스*와 의미적으로 인접한 개념들의 집합
- *Context* : 메타데이터가 사용되는 환경

개념모델에서 메타데이터의 *DEC*와 *Context*는 정보시스템의 MDR에서 추출하고, *Property*와 *N_set*은 WordNet에서 추출한다.

위에서 정의한 개념 모델을 BNF로 표현하면 다음과 같다.

표 1 개념 모델의 BNF표현

개념 모델의 BNF표현

<Concept Model>::= {
DEC:{<dec_set>}
Property :{<prop_set>}
N_set :{<object_set>,<neighbor_set>}
Context : <context>
}
 	
<dec_set>::={<syn_sets>}
<prop_set>::={<syn_sets>}
<object_set> ::={<syn_sets>}
<neighbor_set>::={}|{<syn_sets>}
<context>::=<word>
<syn_sets>::={<syn_set>}|<syn_sets>,{syn_set}
<syn_set>::=<word>|<syn_set>,<word>

3.1.1 객체 클래스의 의미적 이웃의 범위

개념모델에서 *N_set*은 메타데이터의 '객체 클래스'와 의미적으로 이웃관계에 속하는 개념들의 집합이다. 의미적 이웃이란 '객체 클래스'에 대한 워드넷(WordNet) 계층 구조상에서 의미적으로 인접한 개념들의 집합을 말한다. 이때 의미적 이웃집합은 워드넷 계층 구조상에서 메타데이터의 '객체 클래스'를 중심으로 방향성 없는 호(undirected arc)의 개수가 같은 범위에 있는 것으로 선출한다.

의미적 이웃집합 선출을 위해 거리에 사용되는 미터법은 자기 자신(객체 클래스)을 0이라고 가정하기 때문에, 개념들의 의미적 이웃에 자기 자신도 포함된다. 아래그림은 메타데이터의 '객체 클래스'에 대한 워드넷 계층 구조를 보여준다.

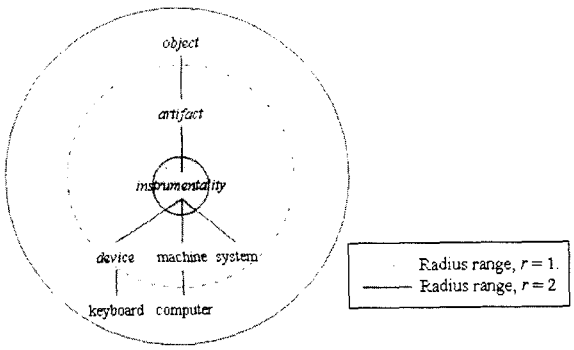

그림 2 의미적 이웃집합의 예

그림 2에서 의미적 이웃집합을 아래와 같이 표현할 수 있다.

- 반지름의 범위가 $r=1$일 때,
  N_set = {*instrumentality, artifact, device, machine, system*}
- 반지름의 범위가 $r=2$일 때,
  N_set = {*instrumentality, artifact, object, device, machine, system, keyboard, computer*}

4. 메타데이터 의미 유사성 측정 방법

정보시스템간의 상호운용성 증진을 위해 메타데이터 의미 유사성을 측정하는 순서는 다음과 같다.

1) 개념 모델의 N_set을 사용하여 유사성을 검사할 후보 메타데이터들을 선정한다.
2) 선택된 후보 메타데이터들에 대한 개념 모델 인스턴스를 생성한다.
3) 인스턴스에 매칭 프로세스 기반 유사성 측정 함수를 적용하여 각각의 유사성을 계산한다.
4) 가장 유사성이 높은 메타데이터를 최종 후보로 선정하여 정보시스템간의 정보공유에 사용한다.

4.1 유사성 측정 함수

앞에서 정의한 개념 모델을 이용하여 서로 다른 정보시스템에서 메타데이터간의 유사성을 측정할 수 있는 함수를 제시한다. 아래 (1)은 이질적 도메인의 정보시스템 a와 b가 존재할 때, 각 정보시스템의 메타데이터 p, q간의 의미유사성을 측정하기 위한 함수이다.

$$S(a^p,b^q)=w_{dec}\times S_{dec}(a^p,b^q)+w_{pr}\times S_{pr}(a^p,b^q)+w_n\times S_n(a^p,b^q)+w_{ct}\times S_{ct}(a^p,b^q) \quad (1)$$

- 유사도(S): S_{dec}, S_{pr}, S_n, S_{ct} 는 각각 메타데이터의 객체 클래스의 동의어 집합, 속성의 동의어 집합, 객체 클래스의 의미적 이웃 집합, 문맥에서 계산된 유사도
- 가중치(W) : W_{dec}, W_{pr}, W_n, W_{ct} 는 메타데이터의 객체 클래스의 동의어 집합, 속성의 동의어 집합, 객체 클래스의 의미적 이웃 집합, 문맥의 가중치

아래 (2)는 개념모델의 각 항목별로 의미유사성 측정하기 위한 함수이다.

$$S_i(a,b)=\frac{A\cap B}{|A\cap B|+\alpha(a,b)|A/B|+(1-\alpha(a,b))|B/A|} \quad (2)$$
(단, $0\le\alpha\le1$)

- a, b는 비교하려고 하는 메타데이터
- A, B는 a, b의 메타데이터 요소 개념의 동의어 집합(*DEC*), 속성 동의어 집합(*Property*), 의미적 이웃들의 집합(N_set) 또는 문맥(*Context*)
- i은 메타데이터 요소 개념의 *dec, pr, n, ct*

위의 함수(2)를 이용하여 개념모델의 각 항목별로 유사성을 측정하고, 그 결과를 함수(1)에 적용하여 궁극적으로 메타데이터간에 유사성을 측정할 수 있다.

5. 결론

최근 정보시스템간의 의미적인 정보공유를 하기위한 연구가 진행되고 있다. 본 논문에서는 도메인이 다른 정보시스템간의 상호운용성 증진을 위한 메타데이터 의미 유사성 측정방법을 제안하였다. 정보시스템내에 존재하는 수많은 메타데이터 가운데 후보를 선정하고, 선정된 메타데이터의 유사성을 측정하여, 궁극적으로 가장 의미적으로 유사성이 높은 메타데이터들간에 정보공유를 할 수 있다. 제안한 방법을 사용해서 나온 결과로 도메인이 서로 다른 정보시스템간의 정보공유를 위한 최적의 메타데이터 후보를 제공할 수 있다.

향후 의미적 연관관계의 개념을 도입하여 보다 의미적인 공유가 될 수 있도록 지속적인 연구가 필요할 것이다.

참고문헌

[1] Environmental Protection Agency (EPA), Environmental Data Registry (EDR, USA): http://www.epa.gov/edr/
[2] 한국전자거래진흥원, 전자거래 중앙등록저장소(REMKO, Registry&Repository of ebXML in KOREA), http://www.remko.or.kr:8000/jsp/index.jsp
[3] ISO/IEC 11179 Information technology-Metadata registries
[4] ISO 15836:2003 Information and documentation - The Dublin Core metadata element set