데이터베이스를 이용한 도메인 온톨로지의 효율적인 생성

김태석 양진혁 정인정
고려대학교 전산학과
e-mail: {cstkts, grjinh, chung}@korea.ac.kr

Efficient Creation of Domain Ontology Using DataBase

Tae-Suk Kim Jin-Hyuk Yang In-Jeong Chung
Dept. of Computer Science, Korea University

요 약

월드와이드웹(WWW) 기술은 폭발적으로 증가하고 있는 웹 데이터들의 의미적인 정보를 효과적으로 처리하기에는 많은 문제점이 있다. 이러한 문제점을 해결하기 위하여 1999년 말에 제안된 시맨틱 웹은 온톨로지를 기반으로 하고 있다. 그러나 온톨로지 생성에 관한 많은 연구들은 많은 시간과 비용이 소비된다. 이와 같은 문제를 해결하기 위하여 우리는 데이터베이스에서 온톨로지를 생성할 수 있는 방법을 제안한다. 데이터베이스는 도메인을 잘 나타내고 있는 정보의 저장소이므로 데이터베이스로부터의 온톨로지 생성은 분석, 설계 등의 사전 작업이 필요하지 않다. 우리는 데이터베이스에서 스키마를 추출, 뼈대그래프를 생성하고 개념그래프로 확장하여 도메인을 잘 나타낼 수 있는 온톨로지를 생성한다. 끝으로 알고리즘을 통한 생성을 함으로서 제안된 생성방법을 검증한다.

1. 서론

최근 웹 데이터는 폭발적으로 증가하고 있다. 그러나 사용자가 원하는 의미적인 정보를 처리하기에는 적합하지 않다. 이러한 문제점을 해결하기 위해 Tim Berners-Lee는 새로운 개념의 웹인 시맨틱 웹[1]을 제안하였다. 시맨틱 웹은 사람뿐만이 아닌 기계가 데이터의 의미를 이해하고 처리한다. 따라서 기존 표현 위주의 웹에서는 기계가 인식하기 어렵고, 공유 및 재사용이 어렵다는 등의 단점들을 보완 할 수 있다. 시맨틱 웹은 데이터의 의미를 이해하고 처리하기 위하여 도메인 내의 개념들과 개념들 간의 관계를 정형적으로 기술하고 있는 온톨로지[2], 지식 저장소를 기반으로 한다. 온톨로지를 생성하기 위한 방법은 인지, 생성, 평가, 기록의 과정을 거친다[3]. 이 방법은 많은 부분을 휴리스틱에 의존하기 때문에 풍부한 어휘와 정형적 의미에 충실한 온톨로지가 생성 되지만 시간과 비용이 많이 소비된다. 이러한 문제를 해결하기 위해 많은 연구들[4, 5, 6, 7, 8]이 있었다. 그러나 이러한 연구들도 분석, 설계, 키워드 색출, 데이터 선별 등과 같은 과정으로 많은 부분에서 시간과 비용이 소비된다. 따라서 우리는 시간과 비용의 소비를 더 줄일 수 있는 새로운 온톨로지 생성방법으로서 데이터베이스에서 온톨로지를 생성하는 방법을 제안한다.

데이터베이스는 데이터모델을 통한 분석, 설계 단계를 거쳐 생성된 도메인을 잘 표현하고 있는 저장소이다. 또한 데이터베이스의 데이터들은 도메인을 나타내는 데이터들로만 이루어져 있기 때문에 도메인을 나타내는 키워드(keyword)들을 선별하는 사전 작업이 필요하지 않다. 따라서 데이터베이스에서의 온톨로지 생성은 많은 부분을 휴리스틱에 의존하고 있는 기존 생성 방법의 의존성을 상당히 많이 줄이고, 사전 작업의 노력을 줄임으로서 기존 생성방법보다 시간과 비용을 획기적으로 줄일 수 있다. 따라서 우리는 데이터베이스를 이용하여 온톨로지를 생성 하는 방법을 제안한다. 데이터베이스 스키마를 이용하여 뼈대그래프를 생성하고 이를 확장하여 개념그래프를 생성한다. 생성된 개념그래프에 데이터베이스 튜플(tuple)들을 매핑 함으로서 온톨로지를 생성한다.

본 논문의 구성은 다음과 같다. 2장에서는 우리가 제안하는 생성방법에서 사용하는 기존 기술들에 대해 간단히 언급한다. 3장에서는 기존의 생성방법들을 언급하고 문제점을 토론한다. 4장에서는 알고리즘과 함께 데이터베이스와 온톨로지의 매핑 관계를 정립하여 온톨로지의 생성방법을 논의 한다. 5장에서는 실험을 통하여 온톨로지가 생성하는 구체적인 예제를 보인다. 마지막으로 6장에서는 결론 및 향후과제에 대해 언급한다.

2. 기반 개념

2.1 데이터베이스 / Entity-Relationship(E-R) Model

데이터베이스는 상호 연관이 있는 데이터의 모임이다[9].

데이터베이스는 현재 정보를 다루는 대부분의 시스템에서 정보를 저장하는 저장소로서 상호 연관을 키 값 매칭에 의한 관계로서 잘 표현하고 있다. E-R 데이터모델은 실세계의 인식에 기본을 둔 실세계의 기본적인 객체를 나타내는 엔티티(entities)와 객체들 간의 관계로 구성된다[9]. E-R 모델은 데이터베이스 설계의 근간으로서 우리는 데이터베이스에 잘 표현되어 있는 테이블과 객체들 간의 상호 연관을 이용한다. 그러나 데이터베이스의 상호 연관은 키 값 매칭으로 관계를 표현 하고 있기 때문에 온톨로지의 관계를 표현하기에는 부적합하다. 키 값 매칭은 테이블과 테이블의 연결 고리의 역할 만 있을뿐 연결 관계의 의미를 담고 있지 않다. 따라서 온토로지를 생성할 때 관계를 정의하기 위해선 도메인 전문가 또는 데이터베이스 설계자의 최소한의 개입이 필요하다.

2.2 OWL (Ontology Web Language)

OWL은 온톨로지 기술을 위한 표준 마크업 언어로써 사람에게 정보를 제공해 줄 뿐만 아니라 기계가 이해 할 수 있는 계층적 구조로서 기계에게 정보를 제공해 준다. 그 구조는 기계가 지식을 직접 분석 처리 할 수 있어 그 활용 분야의 폭은 상당히 넓다. 풍부한 어휘(vocabulary)와 정형적 의미(formal semantics)를 포함하고 있기 때문에 기계의 해석을 요구하는 웹 컨텐트를 저작하는데 있어 XML, RDF 및 RDF-S와 같은 기타 온톨로지 기술 언어 보다 표현력이 뛰어나다[14].

3. 관련 연구

그동안 시맨틱 웹의 기반구조인 온톨로지의 생성에 관한 연구들의 상당부분은 도메인 전문가들의 휴리스틱에 의존하는 수작업 형태가 주를 이루고 있다. 이러한 생성 방법은 많은 시간과 비용이 소요 된다. 이러한 문제는 온톨로지 생성 분야의 연구에서 무시 할 수 없는 부분으로서 좀 더 효율적이고 효과적으로 온톨로지를 생성하기 위한 여러 가지 방법들이 제안되어 왔다. 온톨로지를 생성하기 위한 여러 방법들은 UML을 이용한 도메인 온톨로지 생성방법[7, 8], 형식적 개념 분석(Formal Concept Analysis)을 이용한 도메인 온톨로지 생성방법[4], 데이터 마이닝 기법(ID3, AOI, association rules mining, clustering)을 이용한 온톨로지 생성 방법[5, 6], 도구들(Ptotege 2000, etc)을 이용한 온톨로지 생성 방법들이 있다.

UML을 이용한 도메인 온톨로지 생성방법은 UML과 OWL(Web Ontology Language)의 매핑관계를 통해 UML로 작성된 도메인 분석 결과를 온톨로지로 생성한다. 이 방법은 휴리스틱에 의존하는 것 보다 시간과 비용이 절약 된다는 장점을 가지고 있으나 설계, 분석 과정이 반드시 필요하다는 단점이 있다. 형식적 개념 분석 방법은 문서들(documents)에서 중요 키워드(keyword)를 통하여 상위 온톨로지인 개념 트리 또는 그래프를 생성하여 온톨로지로 확장한다. 개념 트리 또는 그래프 생성이 자동적으로 가능하기 때문에 시간과 비용이 절약되는 장점이 있으나 키워드를 색출하는 사전작업이 필요하다. 또한 데이터타미이닝 기법을 이용한 생성 방법은 온톨로지 생성의 모체가 되는 저장소에서 필요로 하는 데이터를 선별, 선택하여야 하는 사전 작업이 반드시 필요하다.

분석, 설계, 키워드 색출, 데이터선별은 전체 생성과정의 많은 부분을 차지하고 있으므로 많은 시간과 비용이 필요하다. 이러한 시간과 비용의 소비는 효율적이고 효과적인 온톨로지 생성의 저해 요소가 된다.

4. 온톨로지의 생성 방안

온톨로지는 도메인 영역의 명백한 개념화이고 공유되고 공통된 이해를 제공한다[10]. 개념화를 포함하기 위한 여러 연구들은 온톨로지 생성을 위해 "skeletal" 방법론[3]을 따른다[11, 12]. "skeletal" 방법론은 애드혹(ad-hoc) 처리과정을 취하고 있으며 그 과정은 다음과 같다.

- ▶ 목적 인지(Identify Purpose)
- ▶ 온톨로지 생성(Build Ontology)
  - ▹ 온톨로지 캡춰(Ontology Capture)
  - ▹ 온톨로지 코딩(Ontology Coding)
  - ▹ 온톨로지 통합(Ontology Integration)
- ▶ 온톨로지 평가(Evaluate Ontology)
- ▶ 온톨로지 기록(Document Ontology)

우리는 "skeletal" 방법론을 따라서 개념화 부분인 온톨로지 캡춰, 온톨리지 코딩 부분에 초점을 둔다.

4.1 온톨로지 생성 알고리즘

step1) 도메인을 잘 나타내는 데이터베이스를 선택
step2) 데이터베이스에서 스키마를 읽어 들여 개념그래프(conceptual graph) 생성
step2.1) 스키마에서 다이어그램(diagram)을 추출
step2.2) 다이어그램을 기반으로 뼈대그래프를 추출
step2.3) 뼈대그래프에 관계를 추가하여 개념그래프를 생성
step2.4) 뼈대그래프의 각 노드에 속성(attribute)과 제한(restriction)을 추가
step3) 데이터베이스 튜플(tuple)들을 개념그래프에 매핑
step4) 온톨로지 기술언어(OWL)로 기술

4.2 매핑 관계의 정립

데이터베이스는 분석, 설계과정을 거쳐 생성된 저장소이기 때문에 데이터베이스에서 온톨로지의 생성은 분석, 설계과정을 거치지 않는 장점을 지니고 있다. 데이터베이스는 분석, 설계과정을 거쳐 구성된 저장소이므로 도메인을 잘 내포하고 있다. 따라서 데이터베이스에서의 온톨로지 생성은 도메인을 잘 나타내는 키워드 또는 문서의 선별 등의 휴리스틱에 의존하는 사전 작업이 필요하지 않다.

데이터베이스는 스키마를 지닌다. 스키마는 각 테이블과 테이블 사이의 관계를 정의 한다. 각 테이블은 분석, 설계과정에서 데이터모델에 근간을 두고 있다. 데이터모델에서 테이블 명은 튜플(tuple)들의 대표 명사로 이름 지어지며 각 테이블의 관계는 테이블명과 테이블 이름의 관계를 지어 주는 동사로 설계가 된다. 온톨로지의 개념화 단계에서 각 노드는 노드하위의 값들의 대표 행위나 명사로 이름 지어 질 수 있다[13].

<표 1> 데이터베이스 스키마와 온톨로지 매핑관계

데이터베이스 스키마	온톨로지
테이블 이름	개념그래프의 노드 이름
테이블 관계	개념그래프의 노드 관계
테이블 속성	각 노드의 속성
테이블 각 속성의 제한	각 노드의 제한
튜플	각 노드 하위의 객체

데이터모델의 상호연관은 데이터베이스 구축시 키 값 매칭으로 표현된다. 구축이 완료된 데이터베이스에서의 상호연관은

온톨로지를 위한 관계로 표현하기에 부적합하다. 키 값에 의한 매칭으로 표현되는 상호 연관은 테이블과 테이블의 연결 고리의 역할 만 있을 뿐 연결 고리 자체의 의미를 담고 있지 않다. 또한 데이터모델(E-R Model)에는 각 테이블을 설명하는 메타데이터가 존재 할 수는 있으나 구축 된 데이터베이스에는 메타데이터가 존재 하지 않을 수 있기 때문에 데이터베이스내의 관계를 온톨로지에 그대로 적용 할 수 없다. 따라서 관계 정의 시 최소한의 도메인 전문가의 개입이 필요하다.

5. 실험

본 절에서는 제안한 온톨로지 설계 및 생성 방안의 타당성을 실험을 통하여 검증한다.

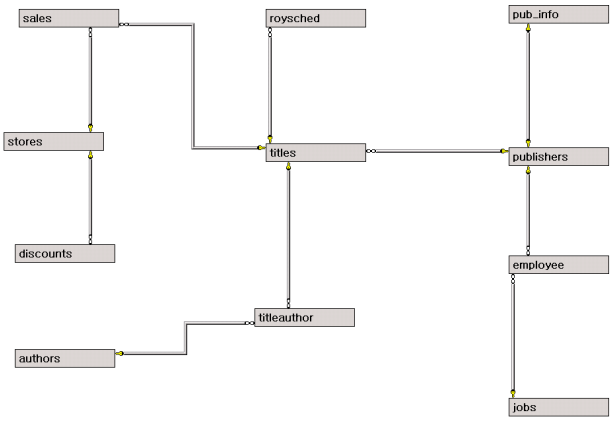
step1) 도메인을 잘 나타내는 데이터베이스를 선택

데이터베이스는 MS_SQL 2000 내에 서적관련 예제인 "pubs"를 선택하였다. "pubs"는 MS_SQL의 서적을 기준으로 서점, 저자 등의 데이터를 기술한 서적 도메인의 좋은 예제이다.

step2) 데이터베이스에서 스키마를 읽어 들여 개념그래프(Conceptual Graph) 생성

step2.1) 스키마에서 다이어그램(Diagram)을 추출

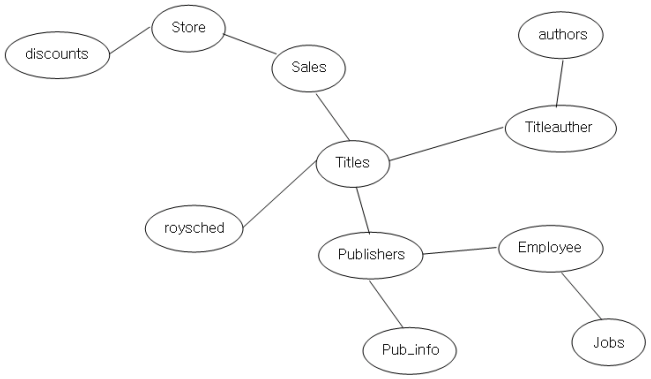
데이터베이스 관리 시스템을 통해 (그림 1)과 같은 스키마를 추출하여 도식화 하였다.



(그림 1) 다이어그램 추출

step2.2) 다이어그램을 기반으로 뼈대그래프를 추출

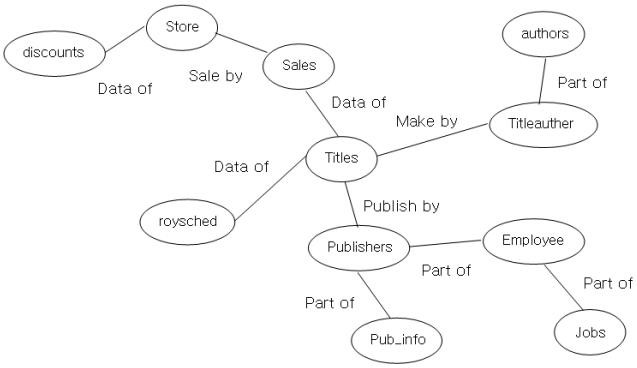
각 노드의 이름은 테이블 이름으로 매핑하고 각 관계는 테이블 키값 매칭 관계를 따라 노드간의 관계를 (그림 2)와 같이 선으로서 표기 한다.



(그림 2) 다이어그램에서 뼈대 그래프 생성

step2.3) 뼈대그래프에 관계를 추가하여 개념그래프를 생성

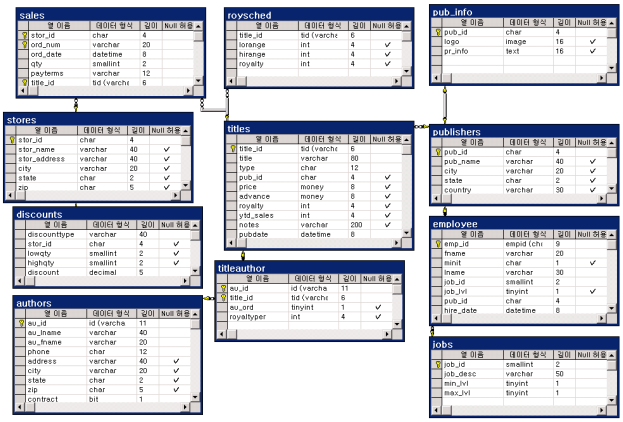
위에서 생성된 뼈대그래프를 기반으로 도메인 전문가에 의존하여 (그림 3)과 같이 각 노드간의 관계를 설정한다.



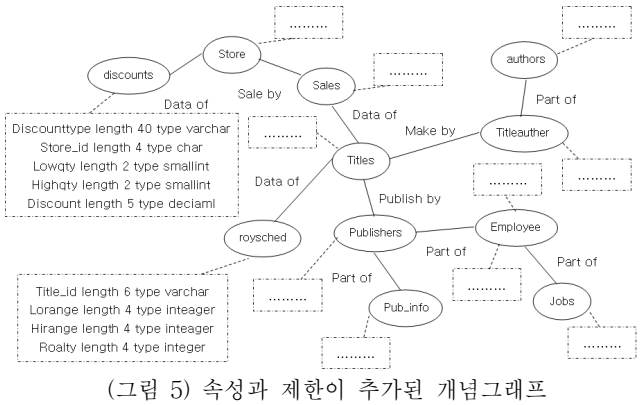
(그림 3) 개념그래프 생성

step2.4) 뼈대그래프의 각 노드에 속성(Attribute)과 제한(Restriction)을 추가

테이블의 속성과 제한(그림 4)을 뼈대그래프에 추가한다. 각 제한은 테이블의 속성의 제한을 중시하여 (그림 5)와 같이 매핑한다.



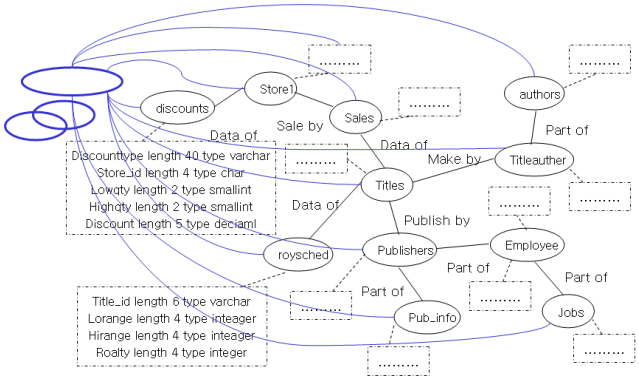
(그림 4) 데이터베이스 스키마 상의 각 속성과 제한



(그림 5) 속성과 제한이 추가된 개념그래프

step3) 데이터베이스 튜플(tuple)들을 개념그래프에 매핑

데이터베이스 튜플들은 개념에 해당되는 객체이다. 각 객체는 개념그래프의 구현(implement)에 해당된다.



(그림 6) 개념그래프의 구현

step4) 온톨로지 기술언어(OWL)로 기술

(그림 7)는 생성된 개념그래프를 이용하여 온톨로지 기술언어인 OWL로 생성된 온톨로지의 일부분이다.

```
                    ……
<owl:Class rdf:ID="titleauthor">
        <rdfs:subClassOf rdf:resource="#titles"/>
</owl:Class>
<owl:ObjectProperty rdf:ID="Part of">
        <rdfs:domain rdf:resource="#Publishers"/>
        <rdfs:range rdf:resource="#Employees"/>
</owl:ObjectProperty>
                    ……
<owl:Class rdf:ID="discounts">
        <rdfs:subClassOf rdf:resource="#store"/>
        <owl:Restriction>
                <owl:onProperty rdf:resource
          ="#DataOf">
        </owl:Restriction>
</owl:Class>
                    ……
<owl:ObjectProperty rdf:resource="Discounttype">
        <owl:Restriction>
                    ……
        </owl:Restriction>
</owl:ObjectProperty>
                    ……
```

(그림 7) 생성된 온톨로지

6. 결론 및 향후 연구과제

우리는 기존 온톨로지 생성방안보다 시간과 비용을 적게 소비하는 효율적이고, 효과적인 온톨로지 생성 방법을 제안하였다. 데이터베이스 스키마를 기초로 하여 뼈대그래프를 생성 후 관계, 속성, 제한을 추가하여 개념그래프로 확장하여 온톨로지를 생성했다.

온톨로지를 생성함에 있어 휴리스틱에 의존하여 온톨로지를 생성하는 것이 제안된 방법보다 풍부한 어휘와 정형적 의미에 충실한 온톨로지가 생성 되지만 시간과 비용이 많이 소비된다. 시간과 비용을 줄이기 위해 여러 생성방법들이 연구되어 왔다. 그러나 이러한 연구들은 분석, 설계, 사전 작업 등의 시간과 비용이 소비된다. 새로 제안한 데이터베이스에서의 생성도 약간의 휴리스틱에 의존하여 시간과 비용을 소비하지만 분석, 설계, 사전 작업을 배제함으로서 그 시간과 비용은 기존에 연구된 온톨로지 생성방안들 보다 적게 소비된다.

본 논문의 알고리즘은 관계를 정의함에 있어 약간의 휴리스틱에 의존하여 자동적이지 못하다. 그러나 관계를 복케블러리를 통하여 정의함으로서 휴리스틱 의존성을 더욱 낮추어 시간과 비용의 소비를 더 줄일 수 있는 가능성이 있다. 그러므로 향후 연구과제로 본 알고리즘의 보완을 통하여 (반) 자동적인 방법으로 향상 시키고자 한다.

참고문헌

[1] Stefan Decker, Sergey Melnik, Frank van Harmelen, Dieter Fensel, Michael Kelin, Jeen Broekstra, Michael Erdman, Ian Horroks, The Semantic Web : the roles of XML and RDF, IEEE, Vol 4, pp.63-70, Sept-Oct, 2000

[2] Asuncion Gomez-Perez, Oscar Corcho. Ontology languages for the Semantic Web, IEEE, Vol. 17, pp 54-60, Jan-Feb, 2002.

[3] Uschold, M. et.al. The Enterprise Ontology The Knowledge Engineering Review , Vol. 13, Special Issue on Putting Ontologies to Use (eds. Mike Uschold and Austin Tate), 1998. Also available from AIAI as AIAI-TR-195

[4] Thanh Tho Quan, Siu Cheung Hui, Alvis Cheuk M. Fong, Tru Hoang Cao. Automatic Generation of Ontology for Scholarly Semantic Web. ISWC 2004: Third International Semantic Web Conference,Hiroshima, Japan, November 7-11, pp 726-740, 2004.

[5] Armin Wrobel, Oliver Wurmli, Josef M. Joller, Siu Cheung Hui. DataMining for Ontology Building, IEEE Intelligent Systems, 2003

[6] 공유근, 양진혁, 김지영, 이윤수, 정인정, 데이터마이닝 기법들을 이용한 (반)자동적인 온톨로지생성. 2003 가을 학술발표 논문집, pp. 130-132 한국정보과학회, 2003.

[7] Stefan Wendler, Mapping XMI / UML to DAML+OIL, http://www.jdev.de/html/projets/ uml2daml/mapping/uml2daml_mapping.html, 2002

[8] 이윤수, 김태석, 양진혁, 정인정. 소프트웨어 공학적 방법을 이용한 온톨로지의 효율적인 설계 및 생성에 관한 연구. 제22회 한국정보처리학회 추계발표대회 논문집 제11권 제2호 (2004.11)

[9] Korth / Silberschatz "Database System Concepts" McGraw Hill.

[10] Reimer, U. Tutorial on Organizational Memories for Capturing, Sharing and Utilizing Knowledge. Intemational Conference on Enterprise Information Systems, ICEIS 2001, Setubal, Portugal, July 7-10, 2001.

[11] Nianbin Wang, XiaofeiXu. A Method to Build Ontology. The Fourth International Conference/Exhibition on High Performance Computing in the Asia-Pacific Region, 2000. 14-17 May 2000 Page(s):672 - 673 vol.2

[12] Rubén Prieto Díaz. A Faceted Approach to Building Ontologies. the 2003 IEEE International Conference on Information Reuse and Integration, IRI - 2003, October 27-29, 2003, Las Vegas, NV, USA

[13] Vijayan, Sugumaran, Veda C. Storey. Ontologies for conceptual modeling: their creation, use, and management. Data & Knowledge Engineering Volume 42 , Issue 3, September 2002, Pages: 251 - 271

[14] OWL : http://w3g.org/2004/OWL/