건강데이터 온톨로지를 위한 반자동 학습 모델

김광성, 황두성

단국대학교 전자계산학과

Semi-Automatic Learning Model for Health Data Ontology

Kim, Kwnagseong, Hwang, Doosung

Dankook University

E-mail : kwangseongkim@gmail.com, dshwang@dankook.ac.kr

요 약

웹 관련 기술의 발전과 더불어 정보시스템의 개발에서 기계가 자동 처리할 수 있는 데이터의 기술 방법으로 온톨로지의 사용이 보편화되고 있다. 온톨로지는 특정 영역의 개념과 그들간의 관계를 단순 명료하게 기술한다. 지식 발견을 위한 도메인 온톨로지 구축은 도메인의 이해, 데이터의 이해, 테스크의 이해, 온톨로지 학습, 온톨로지 평가, 정제 등 다단계를 통해 완성되나 전문성이 요구된다. 본 논문에서는 학습 기반 도메인 온톨로지 구축 방법을 제안하고 건강데이터를 위한 온톨로지 구축에서 응용하였다. 제안된 학습 기반 온톨로지 구축 방법은 건강데이터의 세부 영역별 개념과 관계를 밝히는데 유용하였다.

1. 서론

시멘틱 웹 기반 정보 서비스에 온톨로지는 정보의 구조화와 지식화 개념을 변화시키는 중요한 기술이 되고있다. 정보시스템의 데이터 처리 단위가 텍스트가 아닌 개념 중심으로 변화하면서 지식의 표현과 공유 기반 재사용을 가능하게 하는 온톨로지의 구축에 관한 방법이 연구되었으며, 정보 검색, 지식 관리 분야에서 온톨로지를 이용한 다양한 응용이 나타났다.

지식 관리 시스템에서 지식 발견은 유용한 정보를 발견하는 분야이다. 웹 정보시스템에서 지식 발견의 온톨로지의 도입으로 가능하다. 반 자동 또는 자동으로 프로세스에 의해 만들어 지는 온톨로지는 컴퓨터가 이해 가능하도록 정보를 형식화 및 구조화하여 의미 기반의 서비스를 사용자에게 제공한다[1]. 일반적으로 온톨로지 구축 방법은 도메인의 추상적 개념 요소와 개념 구조를 추출하고, 그들 간의 연관 관계를 파악하는 세부단계로 구성되나 응용 대상의 특징과 서비스에 반영하는 유연한 구축 방법을 선택한다.

본 연구에서는 반 자동 온톨로지 구축을 위한 5단계 프로세스[2]를 구체화하여 개선된 온톨로지 학습 프로세스를 제안하였다. 논문의 구성은 관련 연구에서 온톨로지 구축 방법을 고찰하고, 온톨로지 구축 프로세스를 제안한다. 제안된 온톨로지 학습 프로세스를 적용하여 건강 데이터 온톨로지가 구축된 사례를 보이고, 앞으로 연구 방향을 기술한다.

2. 관련연구

온톨로지란 관심 영역 내 공유된 개념화에 대한 형식적이고 명시적인 명세화다[2]. 의미를 재해석해 보면 특정 목적을 가지고 해당 분야에 있어서 발생하는 개념을 파악하기 위해 합의된 용어와 정의를 사용하여 기계가 처리할 수 있도록 형식적인 표현 방법과 명시적 개념들의 유형과 사용에 대한 규칙을 정의하는 것이다.

온톨로지는 개념(classes), 관계(relations), 속성(properties), 인스턴스(instances)그리고 공리(axioms)로 구성된다. 간단히 온톨로지에 대해 정의 내리자면 4-튜플(C, R, I, A)이다. C는 개념들의 집합, R은 관계들의 집합, I는 인스턴스들의 집합, A는 공리들의 집합이다[3].

CRISP-DM(CRoss Industry Standard Process for Data Mining)은 데이터 마이닝 수행에 필요한 정립된 6단계 프로세스 모델을 제시하고 있으며 그림 1과 같다[4].

그림 1 CRISP-DM 프로세스

비즈니스 이해는 비즈니스 관점에서 프로젝트 목적과 요구사항을 이해 한 후 데이터 마이닝 관점을 통해 문제점과 요구사항을 해결하기 위한 예비 계획을 수립한다. 데이터 이해는 데이터 수집과 다뤄질 데이터에 대한 이해 활동 그리고 문제를 해결에 도움이 되는 특정 데이터 집합에 대한 수집 활동을 수행하게 된다. 데이터 준비 단계에서는 모델링 도구에서 사용될 데이터를 선택 및 변환하는 작업을 수행한다. 평가 단계는 완성된 데이터 모델이 배치되기 이전 마지막으로 적합성에 여부 판단한다. 최종적으로 배치 단계를 통해 어플리케이션과 연동하여 데이터 마이닝으로 획득한 모델을 적용하게 된다.

지식 관리 시스템 개발 관점에서 CRISP-DM을 기초로하는 반 자동 온톨로지 구축 프로세스는 도메인 이해, 데이터 이해, 업무 이해, 온톨로지 학습, 온톨로지 평가 그리고 정제 프로세스로 제 정의 될 수 있다[2].

그림 2 METHONTOLOGY 개발 활동

METHONTOLOGY[5]는 온톨로지 구축 활동을 관리 활동, 개발 활동 그리고 지원 활동으로 구분 한다. 관리 활동은 프로젝트에 필요한 시간, 인적, 물질적 자원을 적절하게 배치한다. 개발 활동은 도메인에 대한 사전 조사부터 시작하여 온톨로지 구축 그리고 유지 보수 활동 전반적인 개발 활동을 포함한다. 지원 활동에서는 온톨로지 개발 단계에서 요구되는 부가적인 활동을 수행한다. 예를 들어 평가, 문서화 그리고 버전 관리 등이 있다. 그림 2는 METHONTOLOGY에서 제시한 활동을 나타내고 있다.

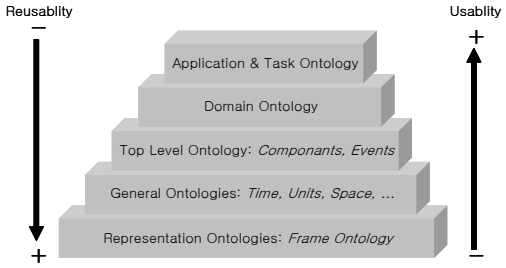

그림 3 온톨로지 재사용성과 사용성[6]

온톨로지는 개념화의 목적에 근거하여 다양하게 나눠진다. 상위 레벨 온톨로지로 일수록 재사용성이 강조되며, 하위 단계 온톨로지로 갈수록 사용성을 지향하게 된다. 온톨로지 구축 방법론 측면에서 온톨로지 개념화 목적 및 범위에 따라 상이한 개발 방법론이 사용된다. 건강 데이터 온톨로지는 특정 범위의 데이터를 통해 비만 상태를 분류하는 것으로 좁은 범위의 특정 문제를 해결하기 위한 온톨로지 구축 방법론이 요구된다. METHONTOLOGY의 경우 도메인 이상의 범위의 일반적인 개념과 용어에 대해 온톨로지를 설계하기 위해 적합한 방법론이며[6], 본 CRISP-DM[2]의 경우 초기 개발 단계에서 제시하는 비즈니스 이해, 데이터 이해, 업무 이해 단계는 한정된 범위의 문제를 정확한 이해를 통해 목적에 적합한 온톨로지 구축 기회를 제공한다.

3. 반 자동 온톨로지 구축 프로세스

반 자동 온톨로지 학습 프로세스는 도메인 및 데이터 이해 단계를 마친 후 수행된다. 앞에서 제시한 선행 연구에서는 온톨로지 학습 단계[2]

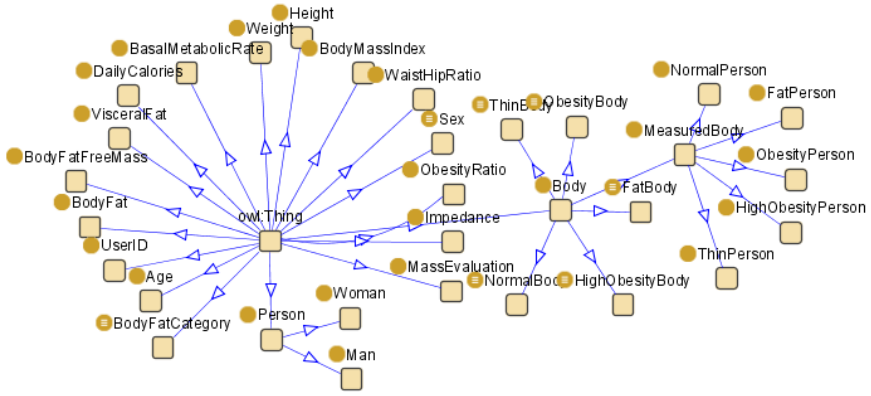

그림 3 건강 데이터 온톨로지 클래스

그리고 개념 및 형식화 그리고 구현[5] 단계에서 온톨로지 학습 프로세스가 이루어진다.

온톨로지 구축에서 가장 중요한 단계는 온톨로지 학습 프로세스에서 수행되는 개념화 단계이다. 이 단계에서 클래스, 관계, 공리가 도출되며 형식화 단계를 통해 온톨로지 언어로 표현된다. 일반적인 온톨로지 학습 프로세스 단계는 개념 유도, 관계 유도, 온톨로지 완성, 온톨로지 생성 그리고 온톨로지 수정 및 확장을 통해 온톨로지가 구현된다.

I. **용어 정의**: 도메인에서 사용되는 용어들을 정리한다. 용어집에는 이름, 동의어, 약어, 설명, 타입등 에 대한 정보를 나열한다.

II. **개념 유도**: 용어집을 기초로 도메인에서 사용되는 개념을 유도 한 후 추상화 와 일반화 단계를 반복적으로 수행하여 계층적인 개념 구조를 구성한다.

III. **관계 유도**: 개념 및 데이터들 간 관계를 파악 후 나타낸다. 관계는 이진 관계 형태로 나타나면 정의역과 공역이 정의되며, 관계는 클래스의 속성이 가질 수 있는 값의 범위를 결정 짓는 동시에 클래스를 제약 사항을 추가시켜 구체화 한다. ∀hasBody(domain:Person, range:Body)과 같이 표현된다.

IV. **공리 정의**: 온톨로지에서 클래스는 공리에 의해 정의 되며, 이후 추론을 할 수 있는 조건을 형성한다. 예를 들어 사람은 나이와 이름 그리고 신체를 가지고 있다라는 사실은 사림이기 위한 필요조건이 된다. 반대로 어떤 인스턴스가 사람인지 아닌지 판단하기 위해 사람이기 위한 충분 조건을 추가하여 인스턴스가 특정 클래스에 분류될 수 있도록 한다.

V. **형식화**: 개념, 관계, 개념간 계층구조 그리고 공리를 형식화된 언어로 표현 한다. 예를 들어 ∀(x,y)hasName ∧ ∀(x,y)hasAge ∧ ∀(x,y)hasBody ⇒ Persno(x)라고 표현 할 수 있다.

VI. **온톨로지 생성**: 형식화된 지식을 온톨로지 지원 도구를 사용하여 온톨로지 언어를 구현한다.

VII. **예외 및 오류사항 평가**: 온톨로지 개발 지원 도구들은 온톨로지에 대한 검증 할 수 있는 추론 엔진을 탑재 하고 있다. 이를 이용하여 선언된 표현과 추론된 표현간의 차이를 파악하고 잘못된 추론을 부분의 공리를 수정한다.

VIII. **평가**: 온톨로지 평가 방법에서 제시한 사항을 평가 한다.

4. 건강 데이터 온톨로지 구축

건강 데이터는 비만을 관리하기 위한 목적으로 체지방률에 따라 5가지로 비만 상태를 분류하였다. 가장 먼저 건강 데이터를 이해하기 비만 관련 용어 표 1이 정의되었다. 다음에 도메인에 대한 지식과 용어을 참고하여 개념 유도 과정을 통해 계층적인 개념구조도를 그림 4와 같이 설계한다. 이후 개념 및 자원들간 관계를 조사하여 관계에 대한 정의역과 공역을 결정 짓는다. 용어

를 설명할 수 있는 명시적인 공리를 추가함으로서 용어 사전이 완성된다. 비만 상태를 분류하기 위해 사용된 규칙은 Person(x) and BodyFatCategory(y) : ∀(x,y)hasBodyFatCategory(x, HighObesityFat) ⇒ HighObesityBody(x)를 사용하였다.

한글명	영문명	동의어	약어	타입
성별	sex			클래스
연령	age			클래스
체중	weight			클래스
표준체중	standard weight			클래스
체지방량	body fat mass			클래스
제지방량	fat free mass			클래스
내장지방면적	visceral fat area			클래스
비만도	obesity ratio			클래스
기초대사량	basal metabolic rate		BMR	클래스
1일필요대사량	daily calories			클래스
임피던스	impedance			클래스
체중평가	weight evaluation			클래스
체지방량평가	body fat mass evaluation			클래스
근육량평가	mass evaluation			클래스
허리힙비율	waist hip ratio		WHR	클래스
체질량지수	body mass index		BMI	클래스

표 1 건강 데이터 용어집 예

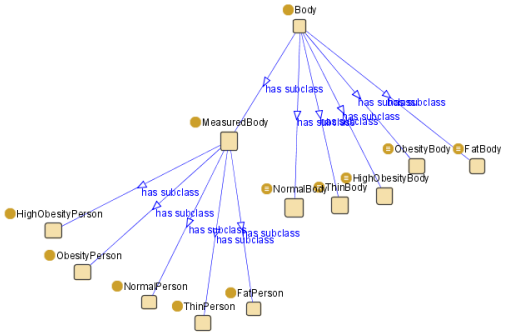

그림 4 Body 클래스

형식화 단계에서 용어 사전 내용을 기계가 이해 할 수 있는 형태로 변환한 후 온톨로지 개발 도구를 이용하여 형식화 된 내용을 온톨로지 언어로 표현 한다. 그림 5은 프로티지(Protege,[7]) 도구로 건강 데이터 온톨로지의 개념을 표현한 것이다.

5. 결론

본 연구에서는 건강 데이터에 존재하는 다양한 개념을 이용하여 비만 관리를 지원할 수 있는 온톨로지를 구축하였다. 개발 초기 단계에서는 도메인과 데이터 이해 단계에서 습득된 지식이 사용되었으며, 개념화 단계부터는 온톨로지를 구현하기 위한 온톨로지 엔지니어링 관련 기술이 사용된다. 개선된 온톨로지 학습 프로세스는 개념 및 자원들간 관계를 명확히 이해하고 정의하는데 유용하였으며, 이 결과 정확한 추론이 될 수 있는 형식화된 공리를 정의 할 수 있었다.

도메인 영역에서 특정 문제를 해결하기 위해 설계된 건강 데이터 온톨로지는 재사용성 보다는 사용성에 편재되어 개발된다. 이에 반해 좀 더 일반화된 온톨로지, 상위 단계 온톨로지는 특정 목적의 어플리케이션에 지향하지 않고 도메인에 대한 순수 지식 수준만을 정의하여 온톨로지에 대한 재사용성을 강조한다. 온톨로지 설계에 원래 목적은 재사용성과 공통으로 사용된 용어에 대해 합의된 의견을 제공하는 것이다[6]. 이러한 점을 고려할 때 보다 지식 단계 수준의 온톨로지 구축을 통해 다양한 생체 및 비만 관리 데이터가 통합된 환경 속에서 발전해 나아가야 있는 상위 레벨 온톨로지 구축 또한 의미를 가진다.

본 연구는 정의된 건강 데이터 온톨로지 구축하여 체지방률에 따른 비만 정도를 결정할 수 있다. 구축된 건강온톨로지에 다른 생체 정보를 포함하고 식단 및 운동 온톨로지와 연계하면 종합적인 비만을 관리하는 비만 온톨로지로 확장이 가능하다.

[참고문헌]

[1] Semi-automatic construction of topic ontology: Fortuna B and Mladenic D and Grobelnik M and Proceeding of the ECML/PKDD Workshop on Knowledge Discovery for Ontologies, 2005a

[2] Semantic Web Technologies: Knowledge Discovery for Ontology Constructionp: 9-27: Marko Grobelnik and Dunja Mladenic, WILEY, 2006

[3] Similarity for Ontologies-A Comperhensive Framework: Ehrig M and Haase P and Hefke M and Stojanovic N: Proceedings of 13th European Conference on Information Systems, May 2005

[4] CRISP-DM 1.0: Step-by-step data mining guide: Chapman P and Clinton J and Kerber R and Khabaza T and Reinartz T and Shearer C and Wirth R, 2000
[5] IEEE Inteligent Systems & their applications: Building a Chemical Ontology Using Methontology and the Ontology Design Enviroments 4(1)p37-46: Fermandez-Lopez M and Gomez-Perez A and Pazos A and Pazos j, IEEE, 1999
[6] Ontological Engineering: Gomez-Perez A and Fernandez-Lopez M and Corcho O, Springer, 2004
[7] http://protege.stanford.edu/
[8] Mining the Web: Analysis of Hypertext and Semi Structured Data. Morgan Kaufmann: Chakrabarti S, 2002