데이터마이닝 기법들을 이용한 (반)자동적인 온톨로지 생성

공유근° 양진혁 김지영 이윤수 정인정
고려대학교 전산학과
{kongjac°, grjinh, jykim2002, arzhna, chung}@korea.ac.kr

(Semi-)Automatic Creation of Ontology using Data Mining Technologies

Yu Gn Kong° Jin Hyuk Yang Ji Young Kim Yun Su Lee In Jeong Chung
Dept. of Computer Science, Korea University

요 약

시맨틱 웹의 구현을 가능하게 하는 핵심기술은 도메인 내의 개념들과 개념들 사이의 관계들을 정형적으로 기술한 온톨로지이다. 그러나 온톨로지 생성을 위한 기존 관련연구들의 상당한 부분들은 몇몇의 휴리스틱을 가지는 수작업 형태를 띠고 있다. 기존 연구들의 수작업을 통한 온톨로지 생성은 어려운 작업일 뿐 아니라 시간이 많이 소비되는 문제점을 갖는다. 따라서 본 논문에서는 도메인 온톨로지를 (반)자동으로 생성하는 방법론을 제안한다. 제안한 방법론에서는 데이터마이닝 기법들인 AOI(Attribute-Oriented Induction)와 ID3 알고리즘을 사용한다. 우리는 제안한 접근법이 온톨로지 자동 생성에 있어 실현 가능한 접근법임을 예제로써 증명한다.

1. 서 론

현재 웹은 사람을 위한 표현위주의 웹 데이터를 사용하고 있다. 이 데이터들에는 잠재적으로 알려지지 않은 유용한 정보들을 내포하고 있지만 정보를 이끌어 내기란 쉽지 않다[2]. 또한, 증가하고 있는 데이터를 효과적으로 처리하기 위해서는 기계와 사람에 의해 이해되고 해석되어질 수 있는 형태의 정보(개념)가 필요하다.

이러한 요구에 따라 시맨틱 웹이 1990년대 말 Tim Berners Lee[16]에 의해 만들어졌다. 시맨틱 웹[13]은 기계가 웹 데이터의 의미를 처리할 수 있고, 도메인내의 개념들과 개념들 간의 관계들을 정형적으로 기술하고 있는 온톨로지[11]를 기반 구조로 가진다. 온톨로지를 기반으로 우리는 기존 표현 위주의 웹이 가지는 기계에 의해 인식되어질 수 없고, 재사용이 어렵다는 등의 단점들을 보완할 수 있다.

그러나 아직까지 온톨로지 생성에 관하여 기존 관련 연구들의 상당한 부분들은 온톨로지 생성에 관하여 도메인 전문가들을 지원하기 위한 휴리스틱을 제공하는 수작업 형태를 띠고 있다. 이러한 연구들은 온토로지 생성이 힘들고 시간이 많이 소요되기 때문에 시맨틱 웹 기술들의 넓은 범위의 어플리케이션을 위한 확장적인 해결책이 아니다[9].

따라서 우리는 본 논문에서 도메인 온톨로지를 (반)자동적으로 생성하는 방법론을 제안한다. 본 논문에서 제안한 방법론에서는 데이터마이닝 기법들인 AOI[1, 4]와 결정트리의 한 종류인 ID3 알고리즘[1]을 사용하여 그 결과를 토대로 온톨로지를 생성한다. 우리는 제안한 시스템의 온톨로지 생성 단계에 대해 예제를 통하여 시스템이 온톨로지 자동 생성에 있어 실현 가능한 접근법임을 증명한다.

이 논문의 구성은 다음과 같다. 2장에서는 관련연구를 언급하고, 3장에서는 데이터마이닝과 시맨틱 웹의 연계성을 언급한다. 4장에서는 우리가 제안한 시스템 아키텍처와 온톨로지 생성 단계를 살펴보고, 5장에서는 (반)자동적인 온톨로지 생성 방법론을 예증한다. 마지막으로, 6장에서는 결론 및 향후과제를 논한다.

2. 관련연구

온톨로지에 관련된 연구는 시맨틱 웹의 하부구조를 이루기 위한 온톨로지 생성, 병합, 수정과 관련하여 다양한 생성 기법들이 소개되었다. 이 장에서는 온톨로지 생성 기법 중 인공지능 기법과 데이터마이닝 기법을 이용한 기존의 연구사례들을 보인다.

[5]는 특정 도메인에서 자유문맥 문서들로부터 온톨로지 지식을 자동적으로 추출하기 위한 방법에 기반을 둔 도메인 독립적인 온톨로지를 기계학습과 통계학적인 기술을 적용한 ONTOSTRUCT 어플리케이션을 이용하여 온톨로지를 자동적으로 생성한다.

[7]은 데이터마이닝 기법(결정트리와 규칙 집합)을 통해 지식 표현을 하기 위해 결과를 얻어내고, 그 결과를 OntologyBuilder 어플리케이션을 통해 온톨로지 언어로 변환한다.

[6]은 계층 클러스터링 알고리즘인 COBWEB을 통하여 개념 계층들(concept hierarchies)을 만들어 내고 이 개념 계층들을 통해 RDFS[13]를 만들어 냄으로써 자동적인 온톨로지 생성을 보인다.

[8]은 데이터마이닝 기법을 기존의 Grid 어플리케이션에 적용한 Knowledge Grid를 이용하여 지식을 추출함으로써 DAMON 온톨로지를 생성한다.

3. 데이터마이닝과 시맨틱 웹

이 장에서는 논문의 핵심 기술인 데이터마이닝 기법들 그리고 시맨틱 웹과 데이터마이닝의 관계에 대하여 논의한다.

3.1 데이터마이닝 기법들

데이터마이닝 기법들[1, 3]에는 속성 지향 귀납추리(AOI), 결정트리(decision trees), 연관 규칙(association rules), 분류(classification) 등이 있다. 우리가 본 논문에서 AOI와 ID3을 사용하는 이유는 이 기술들이 사람과 기계에 의해 직관적으로 이해될 수 있을 뿐만 아니라 기계적인 처리 절차를 가시화 할 수 있기 때문에 타당성 검증이 용이하기 때문이다.

AOI를 이용한 기법은 데이터베이스의 잡음이나 오류데이터를 없애거나 줄여서 대용량의 데이터로부터 정보 집약적인 형

태로 정보를 축약할 수 있다.

지식 표현의 한 방법인 ID3은 결정트리 생성하는 알고리즘으로 속성들에 대한 정보의 수준을 판별할 수 있도록 할 수 있다.

3.2 데이터마이닝과 시맨틱 웹의 관련성

데이터마이닝과 시맨틱 웹은 인공지능과 깊은 연관이 있는 응용으로써 지식의 재사용을 위해 기존 정보의 확장의 형태를 띠고 있고 기계에 의해 이해되어질 수 있으며 개념(정보)의 관계나 패턴을 도출함에 의해서 미래에 증가적인 정보를 수용할 수 있다는 공통점을 바탕으로 하고 있다.

따라서, [9]에서 언급한 바와 같이 시맨틱 웹과 데이터마이닝이 결합된 형태의 시맨틱 웹 마이닝의 발전은 시맨틱 웹의 전망을 밝게 하고 있다.

시맨틱 웹 마이닝[9]은 웹에서 새로운 의미 구조들을 발견하기 위해 데이터마이닝의 결과를 개선시키고자 사용한다. 바꾸어 말하자면, 시맨틱 웹 마이닝은 시맨틱 웹을 만드는데 있어서 의미 기반의 유용한 정보를 이끌어내기 위하여 데이터마이닝 기법을 사용한다는 것이다.

본 논문에서는 시맨틱 웹에서 데이터마이닝의 활용을 언급하고 있다. 제안한 방법론은 데이터마이닝 기법을 적용하여 시맨틱 웹의 기초가 되는 온톨로지를 (반)자동적으로 생성할 수 있는 방법론에 관하여 구체적인 사례로써 예증한다.

4. 온톨로지 자동 생성 시스템

이 장에서는 온톨로지 생성하기 위한 시스템 아키텍처 및 구현 단계를 보인다.

4.1 시스템 아키텍처

본 논문에서 제안한 시스템 아키텍처는 데이터마이닝 과정인 전 처리, AOI 그리고 ID3을 거쳐서 온톨로지를 생성시키는 단계로 이루어져 있다.

전 처리 단계를 거친 데이터들은 실세계에서 사용하는 데이터에서 나타날 수 있는 오류들을 줄이거나 제거한다. 이렇게 정재 되어진 정보를 통하여 AOI를 실행하기 때문에 정확한 정보를 집약할 수 있다. 또한, 이 정보들을 통하여 ID3 알고리즘을 수행하여 속성들에 대한 정보 집약 수준을 판별할 수 있다. 이와 같은 단계를 거쳐서 우리는 기존 데이터로부터 온톨로지 정보 추출하는데 있어 정확도를 높였다. 또한, 우리는 각각의 단계에 따른 Visualization을 통하여 얻어진 온톨로지가 타당함을 명확하게 보일 수 있다.

그림 1은 제안한 시스템의 전체적인 아키텍처를 나타낸 것이다. 이 아키텍처를 기반으로 다음 장에서는 온톨로지 생성에 관한 구체적인 예를 보일 것이다.

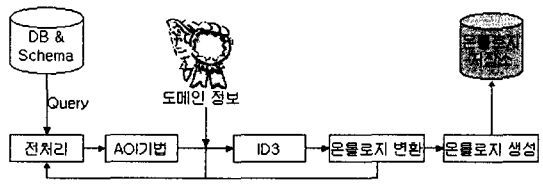

그림 1 시스템 아키텍처

4.2 온톨로지 (반)자동 생성 단계

시스템 아키텍처에 적용되어지는 구현 단계는 다음과 같다.

① 데이터베이스로부터 쿼리를 통하여 도메인 온톨로지에 적절한 후보 데이터들과 속성들을 검색한다.

② 전 처리(Insertion, Smoothing) 과정을 통하여 잘못된 속성 값들을 조정한다.

③ 중복성이 없는 속성 값을 갖는 속성들을 제거한다.

④ 데이터베이스의 개념 계층들을 토대로 데이터마이닝의 drill-down, roll-up을 수행하여 속성 값을 일반화 시킨다.

⑤ 도메인에 필요한 정보들이 제대로 처리되었는지 도메인 정보를 토대로 판단한다. 만약 도메인에 필요한 정보들이 잘 처리되었다면 다음단계로, 아니면 처음단계로 돌아간다.

⑥ 정보들이 집약된 속성들에 대해 ID3 기법을 적용하여 속성들에 대한 각각 정보의 값을 얻어내고, 그 결과 값을 다음 단계에 넘겨준다.

⑦ 결과 값을 토대로 알맞은 온톨로지를 생성한다. 만약 필요한 정보들이 누락된 것이 있다면 처음단계로 돌아간다.

위의 단계들을 거쳐 생성된 온톨로지는 저장소에 저장된다. 저장될 온톨로지의 형태는 위한 RDF[13], DAML+OIL[10]과 DAML-S[14] 같은 형식이 될 수 있다.

5. 구현사례

이 장에서는 DARPA Agent Markup Language 사이트에서 제공하는 예제[12]를 통하여 실제로 우리가 제안한 시스템 아키텍처와 온톨로지 생성단계를 가지고 (반)자동적인 온톨로지 생성에 관하여 예증한다.

먼저 우리는 데이터베이스에 있는 Person 테이블에서 쿼리를 통하여 적절한 후보 집합을 검색한 후, 그 데이터에 대하여 잡음이나 오류데이터를 없애기 위하여 전 처리(insertion, smoothing) 과정을 거쳤다. 다음 순서로 중복을 가지지 않는 속성에 대하여 information gain 값이 작을 것으로 기대되는 불필요한 속성들을 제거한다.

```
person table concept hierarchy(Set-grouping hierarchies)
---------------------------------

{young_age, middle_age, old_age} ⊂ all(age)
{11...30} ⊂ young_age
{31...50} ⊂ mid_age
{51...60} ⊂ old_age

...

{poor, medium, rich} ⊂ all(income_per_month)
{0...9999} ⊂ poor
{10000...14999} ⊂ medium
{15000...} ⊂ rich

...

{small, medium, large} ⊂ all(shoe_size)
{0...230} ⊂ small
{235...265} ⊂ medium
{270...} ⊂ large

...
```

그림 2 person 테이블의 Set-grouping hierarchies

그림 2는 데이터베이스 개념 계층의 한 종류인 Set-grouping hierarchies를 각각의 속성에 대하여 나타낸 것으로, 이 개념 계층을 참고하여 단계 ④에 해당하는 drill-down과 roll-up과 같은 기능을 수행함으로써 각각의 속성들의 값에 대하여 일반화된 값을 얻는다.

다음 단계로 처리되어진 값들이 전 처리와 AOI를 거치며 도메인에 적합한 속성들과 값이 생성되었는지 도메인 정보를 가지고 판단하거나 도메인 전문가에 의해 판독되어진다.

다음 단계로 ID3 알고리즘을 전단계의 결과 값에 적용하여 각각의 속성들에 대한 information gain값과 entropy값을 얻어내어 속성들의 가중치를 판별한다.

마지막으로 ID3을 통해 얻어진 결과인 속성들의 가중치를 통하여 도메인 온톨로지에서 다루어져야 할 적합한 속성들을 찾아 낼 수 있다. 이렇게 선별된 속성들은 도메인 온톨로지에서 개념을 정의할 때 유용한 정보들을 포함할 수 있도록 사용될 수 있다. 따라서 우리는 이러한 속성들을 토대로 온톨로지 정보로 변환하고 저장소에 저장시킨다.

```
Weka를 이용한 Id3 classifier
==============

shoe_size = 205~230
| age = 11~20
| | shirt_size = small: Woman (80%)
| | shirt_size = medium: Woman (100%)
| | shirt_size = large: Woman (100%)
| age = 21~30
| | shirt_size = small: Woman (100%)
| | shirt_size = medium: Woman (60%)
| | shirt_size = large: null (0%)
| age = 31~40: Woman (100%)
| age = 41~50: Woman (100%)
shoe_size = 235~260
| age = 11~20
| | shirt_size = small: Woman (50%)
| | shirt_size = medium: Woman (100%)
| | shirt_size = large: null (0%)
| age = 21~30
| | shirt_size = small: Woman (80%)
| | shirt_size = medium: Woman (100%)
| | shirt_size = large: Woman (100%)
| age = 31~40
| | shirt_size = small: Woman (100%)
| | shirt_size = medium: Woman (80%)
| | shirt_size = large: null (0%)
----
```

그림 3 Weka 어플리케이션을 이용한 ID3 classifier

그림 3은 Weka 어플리케이션[15]을 이용한 ID3 알고리즘을 적용하여 나온 결과의 일부를 보인 것이고, 그림 4는 ID3 단계를 거쳐서 나온 결과를 Jena 툴킷[17]을 사용하여 DAML+OIL 온톨로지로 자동 변환[7]한 예제 소스를 보인 것이다.

```
....

<daml:Class rdf:about="#Person">
  <rdfs:comment>every person is a man or a woman</rdfs:comment>
  <daml:disjointUnionOf rdf:parseType="daml:collection">
    <daml:Class rdf:about="#Man"/>
    <daml:Class rdf:about="#Woman"/>
  </daml:disjointUnionOf>
</daml:Class>

<daml:Class rdf:ID="Adult">
  <daml:intersectionOf rdf:parseType="daml:collection">
    <daml:Class rdf:about="#Person"/>
    <daml:Restriction>
      <daml:onProperty rdf:resource="#age"/>
      <daml:hasClass rdf:resource="http://www.daml.org/2001/03/daml+oil-ex-dt#over17"/>
    </daml:Restriction>
  </daml:intersectionOf>
</daml:Class>

<Person rdf:ID="Adam">
  <rdfs:label>Adam</rdfs:label>
  <rdfs:comment>Adam is a person.</rdfs:comment>
  <age><xsd:integer rdf:value="43"/></age>
  <shoesize><xsd:decimal rdf:value="9.5"/></shoesize>
</Person>

<daml:ObjectProperty rdf:ID="hasHeight">
  <rdfs:range rdf:resource="#Height"/>
</daml:ObjectProperty>

<daml:Class rdf:ID="Height">
  <daml:oneOf rdf:parseType="daml:collection">
    <Height rdf:ID="short"/>
    <Height rdf:ID="medium"/>
    <Height rdf:ID="tall"/>
  </daml:oneOf>
</daml:Class>

....
```

그림 4 DAML+OIL 온톨로지 예제

6. 결론 및 향후과제

우리는 이 논문에서 데이터의 의미를 처리할 수 있는 형태의 시맨틱 웹 기반구조를 구성하기 위하여 데이터마이닝 기법을 적용한 (반)자동적인 온톨로지 생성에 관한 시스템 아키텍처와 생성 단계를 살펴보았다. 또한 기존의 수작업 형태의 온톨로지 생성 부분을 (반)자동적으로 발전시킴에 따라 온톨로지의 생성에 있어 기존의 취약점들을 해결하였다.

우리는 현재 온톨로지 자동생성 하는데 있어 복잡한 개념들의 관계를 이끌어내고 더 나아가 개념들의 제약사항들을 데이터마이닝 기법을 통해 기존 자료들로부터 자동적으로 이끌어내고 추가적인 정보들이 포함된 온톨로지를 생성할 수 있는 연구를 진행하고 있다.

7. 참고문헌

[1] Jiawei Han, Mecheline Kamber, Data mining, Morgan Kaufmann, 2001.

[2] Jiawei Han, Kevin, Chen-Chuan, Chang, Data Mining for Web Intelligence, November 2002.

[3] 나민영, 2000년대의 DB응용기술, '데이터마이닝' 특집호, 1997년 9월호

[4] William J. Frawley, Gregory Piatetsky-Shapiro, Christopher J. Matheus, Knowledge Discovery in Databases : An Attribute-Oriented Approach, AAAI AI MAGAZINE, 1992

[5] Melania Degeratu, Vasileios Hatzivssiloglou, Building Automatically a Business Registration Ontology, dg.o2002 Proceedings, 2002

[6] Patrick Clerkin, Padraig Cunningham, Conor Hayes, Ontology Discovery for the Semantic Web Using Hierarchical Clustering, Trinity College Dublin Computer Science Department, Technical Reports, April 2001

[7] Armin Wrobel, Oliver Wurmli,, Data Mining for Ontology Building, Diploma Thesis - Dep. of Computer Science WS 2002/2003

[8] Mario Cannataro, Carmela Comito, A Data Mining Ontology for Grid Programming, 1st Workshop on Semantic in Peer-to-Peer and Grid Computing at the Twelfth International World Wide Web Conference, 2003

[9] Bettina Berendt, Andreas Hotho, Gerd Stumme, Towards Semantic Web Mining, ISWC 2002, LNCS 2342, pp.264-278, 2002

[10] Deborah L. McGuinness, Richard Fikes, James Hendler, Lynn Andrea Stein, DAML+OIL : an ontology language for the Semantic Web, IEEE Intelligent Systems, Vol. 17, No. 5, pages 72-80, September/October 2002

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

[12] http://www.daml.org/2001/03/daml+oil-ex

[13] Stefan Decker, Sergey Melnik, Frank van Harmelen, Dieter Fensel, Michel Klein, Jeen Broekstra, Michael Erdmann, Ian Horrocks, The Semantic Web : the roles of XML and RDF, IEEE, vol 4, pages 63-70, Sept-Oct, 2000

[14] The DAML Services Coalition, DAML-S: Semantic Markup For Web Services, In Proceedings of the International Semantic Web Workshop, 2001

[15] http://www.cs.waikato.ac.nz/ml/weka/

[16] Tim Berners-Lee, Mark Fischetti, Weaving the Web: the original design and ultimate destiny of the World Wide Web by its inventor, HarperCollins, 1999

[17] http://www.hpl.hp.com/semweb/jena.htm