

## 고객 상품 추천을 위한 온톨로지 자동 생성

구미숙<sup>0</sup> 황정희 류근호

충북대학교 데이터베이스 연구실

{gumisug<sup>0</sup>, jhwang, khryu}@dmlab.chungbuk.ac.kr

### Automatic Ontology Generation for Item Recommendation to Customer

MiSug Gu<sup>0</sup> JeongHee Hwang KeunHo Ryu

Dept. of Computer Science, Chungbuk National University

#### 요 약

최근 인터넷 기술의 발전으로 인해 정보량이 급증함으로써 이들 정보자원을 효과적으로 검색하기 위한 방법으로 메타데이터를 이용하여 필요한 정보 자원에 정확하게 접근하는 방법이 다양한 분야에서 제안되고 있다. 메타데이터는 정보자원을 효과적으로 검색하고 데이터를 재가공하여 다양한 각종 정보자원에 대한 정보 및 기록 관리를 할 수 있다. 이 논문에서는 정보를 효율적으로 검색하기 위하여 XML을 이용한 온톨로지 기반의 메타데이터를 이용한다. 홈쇼핑 사이트의 고객인 소비자를 대상으로 효율적인 정보 추천 및 검색을 위해, 상품 토픽맵 온톨로지를 구축하고 소비자에게 알맞은 쇼핑 정보를 전달하기 위한 정보 검색 시스템을 설계 구축 한다. 온톨로지의 자동적 구축은 데이터 마이닝 기법인 COBWEB의 개념 계층적 클러스터링 알고리즘을 이용하였다. 기존의 전문가에 의한 수동적인 온톨로지 구축을 자동화 시키므로써, 대량의 온톨로지를 구축하여 정보검색에 효율을 기할수 있다..

#### 1. 서 론

최근 인터넷 기술의 발전으로 정보의 양이 급증함으로써, 기존의 웹 검색에 비해 정확한 정보 검색을 위해 시맨틱 웹이 대두되고 있다[1]. 시맨틱 웹의 기반인 온톨로지는 특정 분야의 지식을 표현하기 위한 기본지식 체계를 제공하여 사용자가 원하는 정보에 대해 정확한 검색 결과를 제시해 준다[2]. 온톨로지란 어떤 특정 도메인의 정보들과 그 정보들 간의 관계를 정의해 놓은 것으로서, 기존의 택소노미(taxonomy)처럼 계층 구조를 가지고 있다. 이 논문에서는 온톨로지 구축을 위해서 온톨로지 도메인을, 홈쇼핑 사이트의 고객인 소비자를 대상으로 효율적인 정보 추천을 위해, 각 홈 쇼핑 사이트에서 판매하고 있는 상품을 대상으로 상품 토픽맵 온톨로지를 구축하여 고객에게 알맞은 쇼핑 정보를 전달하는 것을 목적으로 한다. 그리고 온톨로지 구축을 위한 데이터 마이닝 기법과 ISO 표준인 XTM을 이용한 온톨로지 자동 생성 기법을 제시한다.

온톨로지 구축을 위해서, COBWEB이라는 개념을 표현하는 점진적 개념 클러스터링 (incremental conceptual clustering) 알고리즘을 이용한다. 클러스터링을 통해 비슷한 항목들마다 별개의 클러스터를 형성하며, 기본 계층(basic level)에 대한 분류부터 시작해서 계층적 분류 구조를 만들어 간다. 이 과정에서 가장 중요한 것은 항목들 간의 유사도를 측정하는 것이다. 클러스터간 유사도 (inter-cluster similarity)와 클러스터내 유사도 (intra-cluster similarity)의 두 가지 유사도를 기반으

로 하는 카테고리 유틸리티(category utility, CU)를, 클러스터간의 비유사성(inter-cluster dissimilarity)과 클러스터내 유사성(intra-cluster similarity)을 최대화시키는 클러스터를 생성하기 위한 평가 방법으로 사용된다[4,6]. 이 논문에서는 온톨로지를 생성하기 위해서 상품 품목을 토픽 타입으로 설정하고, 사용자의 요청에 의해서 웹 에이전트가 각 홈쇼핑 사이트를 검색하는 학습 (learning)을 통해서 같거나 유사한 품목끼리 클러스터를 생성해가는 과정을 클러스터링 기법을 적용하여 온톨로지를 구축하였다. 토픽맵은 토픽, 어소시에이션, 어커런스등으로 구성된다[3]. 클러스터의 계층 구조를 이루고 있는 각 노드를 온톨로지 계층구조상의 상품의 품목으로 매치시켜 토픽맵의 토픽으로 정하고, 상하 계층 구조는 품목간의 수퍼클래스와 서브클래스라는 연관 관계가 된다. 이러한 구조의 온톨로지는 시맨틱웹의 기반이 되며, 고객이 각 홈쇼핑 사이트의 품목들에 대한 정보에 쉽게 접근이 가능하게 한다.

#### 2. 관련연구

##### 2.1 온톨로지

온톨로지는 특정 분야의 지식을 표현하기 위한 기본 지식 체계를 제공한다. 이런 면에서 택소노미와 유사하지만, 온톨로지의 가장 큰 특징은 온톨로지 내에 정의된 용어간의 상호 연관 관계를 체계적으로 기술한다. 온톨로지는 개념(concept)과 관계(relation)로 구성되고, 개념은 온톨로지 내에 정의된 키워드를 말하고, 관계는 온톨로지 내에 정의된 개념간의 관계를 의미하며, 이를 이원 관계 (binary relation)라 한다[2].

##### 2.2 XTM

XTM(XML Topic Map)은 주제 중심으로 개념을 명세 화

이 연구는 한국전자통신연구원의 정보통신 서비스 연구단의 연구비 지원으로 수행되었음

하고 개념들 간의 연관 관계를 정의한 모델로서 ISO의 표준안으로, 초기에는 전자 색인을 위한 데이터 모델로 고안되었으나 현재는 지식 관리 시스템의 지식 맵, 콘텐츠 관리 시스템의 콘텐츠 맵 그리고 시맨틱 웹 온톨로지 등의 데이터 모델로 사용되고 있다. 토픽 맵 모델의 세 가지 핵심 요소는 토픽(Topic), 연관관계(Occurrence), 어소시에이션(Association)등으로 구성되어 있다. 토픽은 서브젝트를 의미하고, 어소시에이션은 서브젝트와의 연관 관계, 어커런스는 서브젝트에 대한 리소스(resource)가 위치한 정보를 가리키는 역할을 한다[3].

2.3 데이터 마이닝을 이용한 온톨로지 생성 기법

데이터 마이닝이란 데이터로부터 알려지지 않고 유용한 정보를 추출하는 기법으로, 클러스터링은 오브젝트 집합의 유사성 구조에 의해 그 나뉘어 지는 범주의 수를 기준으로 분류하는, 무감독 분류법 (unsupervised classification)이다. COBWEB이라는 개념을 표현하는 점진적인 개념 클러스터링 알고리즘을 이용하여 텍소노미 구조를 생성하여, 온톨로지 구축 언어인 RDF와 RDF(S)의 계층구조와 매치시키는 과정을 통하여 온톨로지를 자동으로 생성하는 방법을 제시하고 있다[5].

3. 데이터 마이닝 알고리즘

이 논문에서는 데이터 마이닝 기법을 이용하여 자동적으로 온톨로지를 생성하는 기법을 제안한다. 개념 계층적 클러스터링 알고리즘인 COBWEB을 이용하여 토픽맵의 클래스 계층구조를 생성한다.

기본 계층에 대한 분류부터 시작해서 계층적인 분류 구조를 만들어 가는 기법으로, 언덕 오르기 탐색 기법(hill-climbing search)을 기반으로 수행된다. 현재의 검색(search) 상태를 평가한 후에 그것을 확장하고, 그 하부 계층을 평가하고, 더 확장을 하기 위해 가장 좋은 자식 노드(best child)를 선택한다. 그리고 그 부모계층(parent)보다 더 나은 선택할 자식 노드(child)가 없으면 중단한다.

COBWEB은 카테고리 유틸리티라는 경험적 평가(heuristic measure)에 의해서 클러스터를 평가한다. Fisher는 같은 클래스 내의 오브젝트의 유사도(similarity)와 다른 클래스내의 오브젝트의 비유사도(dissimilarity)를 최대화시키는 클러스터를 생성하는 클러스터링 방법을 주장하였다[6].

알고리즘은 각각의 인스턴스(instance)에 대해서 루트 카테고리(root category)로부터 시작된 트리를 모두 검색하여 계층 구조를 생성해 간다. 이런 특성을 기반으로 계층 구조인 텍소노미는 다음 과정을 거쳐서 생성된다.

1. 기존에 존재하는 클래스에 따라 오브젝트 분류
2. 새로운 클래스 생성
3. 두개의 클래스를 하나의 클래스로 결합하는 병합(merging) 단계 수행
4. 한개의 클래스를 몇개의 클래스로 나누는 분할(splitting)단계 수행

각 계층에서 텍소노미를 평가하기 위해 다음 식

CU(category utility)를 사용한다[5].

$$CU(C_1, C_2, \dots, C_n) = \frac{\sum_k [(P(C_k)[P(A_i = V_{ij} | C_k)]^2 - \sum_i \sum_j P(A_i = V_j)^2)]}{n}$$

식 1 카테고리 유틸리티(category utility) 식

여기서, n은 카테고리나 클래스의 개수, C<sub>k</sub>는 특정 클래스, A<sub>i</sub>는 i속성 값 중의 하나, V<sub>ij</sub>는 속성 A<sub>i</sub>에 대한 j값 중의 하나를 의미한다.

4. 데이터 마이닝 알고리즘을 이용한 온톨로지 생성

이 논문은 홈쇼핑 사이트의 고객을 대상으로 효율적인 정보 추천과 검색을 위해, 각 홈쇼핑 사이트에서 판매하고 있는 상품을 대상으로 상품 토픽맵 온톨로지를 구축하여 소비자에게 알맞는 쇼핑 정보를 전달하는 것을 목적으로 한다. 이 논문의 상품 토픽맵 온톨로지를 이용한 검색 시스템은 다음과 같은 과정으로 구축 된다.

첫째, 웹 에이전트가 각 홈쇼핑 사이트에서 취급하고 있는 상품품목에 대한 정보를 찾는 학습을 통해서, 같은 상품이나 유사한 상품을 분류하여 유사도를 계산하여 클러스터로 만든다.

예를 들어, 초기에 학습한 품목이 가전제품이라면, 가전제품이 가지고 있는 세부 항목에 대한 계층구조를 만든다. 이때 각 홈쇼핑 사이트마다 같은 품목에 대한 명칭을 유사하게 표기할 수 있으므로 워드넷을 통해서 유사어를 같은 품목으로 취급하는 전 처리 단계를 수행한다. 이 과정에서 두 가지의 유사도를 계산한다.

- 첫째, 클래스 내 유사도로, 유사도 값이 크면 클수록 더 많은 클래스를 가지며, 이것은 조건부 확률인 다음 식으로  $P(A_i = V_{ij} | C_k)$  구한다.
- 둘째, 클래스 간 유사도로, 유사도 값이 크면 클수록 그 값을 더 적은 오브젝트가 공유하게 된다. 이에 대한 함수식은  $P(C_k | A_i = V_{ij})$ 이다.

이를 기반으로 카테고리 유틸리티(CU)를 계산하여, 클러스터를 평가 하여 클러스터링 과정을 진행 시킨다. 이렇게 가져온 자료를 초기화 값으로 놓고 계층 구조를 생성한다.

둘째, 기존의 계층구조에 학습하여 추가된 컴퓨터 품목을 삽입하여 새로 업데이트된 계층구조를 생성하므로 점진적인(incremental) 클러스터링을 가능하게 한다.

셋째, 기존의 계층구조에 새로운 항목이 추가되었지만 기존의 품목과 겹쳐지는 부분이 있으면 같은 레벨의 두 노드를 취해서 병합을 하고 카운트 값을 합하여 새로운 노드를 생성한다.

넷째, 기존의 계층구조가 최상의 계층 구조가 아닐 때 최상의 노드를 생성하기 위해서 분리단계를 수행한다. 다음 그림은 COBWEB을 이용해서 생성된 계층구조와 클러스터에 의해서 생성된 온톨로지 계층구조이다.

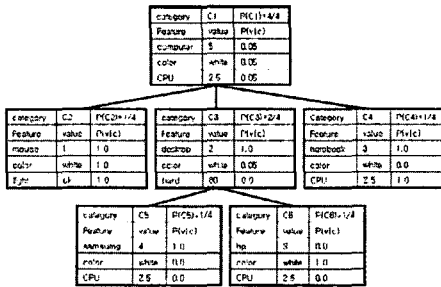


그림 1 COBWEB에 의해 생성된 컨셉트 맵소노미

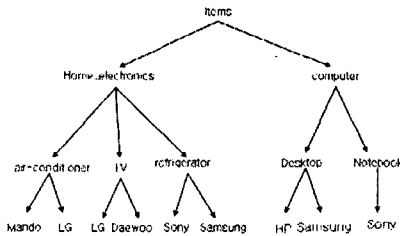


그림 2 클러스터에 의한 온톨로지 계층구조

이와 같이, COBWEB에 의해서 생성된 계층구조를 토픽 맵을 이용하여 텍소노미의 노드를 토픽 맵의 토픽으로, 계층구조의 상하계층을 슈퍼클래스 서브클래스로 매치시켜서 토픽맵을 이용한 온톨로지를 생성한다.

이 논문에서는 XTM을 이용하여 온톨로지를 설계 생성하였다. 토픽은 상품 품목을, 어소시에이션은 품목간의 슈퍼클래스와 서브클래스의 연관 관계를, 어커런스는 홈쇼핑 사이트의 주소를 가리키도록 정하였다. 이러한 과정을 거쳐 상품 토픽맵이 자동으로 구축된다.

다음은 토픽맵을 이용한 상품 토픽맵 온톨로지 구축을 위한 XTM 문서를 보여준다.

```

<!-- items_computer -->
<topic id="computer">
  <instanceOf>
    <topicRef xlink:href="#Items">
      </instanceOf>
    <baseName>
      <baseNameString>COMPUTER</baseNameString>
    </baseName>
  </topic>
  <!-- association -->
  <association id="items_computer">
    <instanceOf>
      <topicRef xlink:href="#super_sub_class">
        </instanceOf>
      <member>
        <roleSpec>
          <topicRef xlink:href="#superclass">
            </roleSpec>
          <topicRef xlink:href="#Items">
            </member>
          <member>
            <roleSpec>
              <topicRef xlink:href="#subclass">
                </roleSpec>
              <topicRef xlink:href="#computer">
                </member>
              </member>
            </association>
  </association>
  
```

그림 3 상품 품목 온톨로지 구축 XTM문서

구축된 온톨로지 문서는 XTM 문서이므로 온토피아(Ontopia)사의 토픽맵 톨인 옴니게이터(Omnigator)를 사용하여 유효성 검사를 한다. 온톨로지를 데이터베이스에 저장하기 위해서는 Saxparser와 Tmparser를 통해 파싱 과정을 거치고, 파싱된 온톨로지는 하이버네이드를 사용하여 객체관계형(object relation)의 클래스 상태로 데

이터베이스에 저장한다. 이렇게 각 클래스 별로 온톨로지 데이터베이스에 저장된 온톨로지는 각 상품이 토픽 클래스, 토픽들의 연관관계가 어소시에이션 클래스, 홈쇼핑 사이트 주소가 어커런스 클래스로 구성된다. 이와 같은 온톨로지를 이용한 정보 검색 시스템은 홈쇼핑 사이트의 고객에게 상품 품목에 대한 정보를 쉽게 접근하게 해준다. 다음 그림은 구축된 온톨로지를 토픽맵 브라우저인 TMBrowse를 통해서 웹 브라우저 한 결과이다.

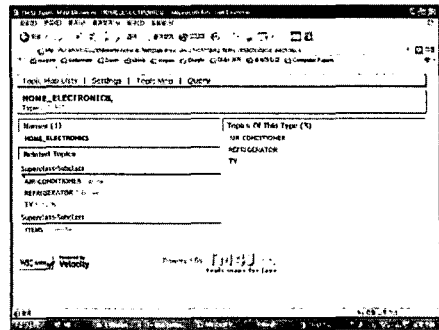


그림 4 TMBrowse를 이용한 온톨로지 브라우저

토픽맵 온톨로지를 웹으로 보여진 그림이며 가전제품에 대한 온톨로지 계층구조를 보여주고 있다.

### 5. 결론

이 논문에서는 대량의 다양한 정보를 사용자에게 좀더 효율적이고 정확한 추천과 검색 결과를 제시하기 위해서, 데이터 마이닝 기법인 개념 계층적 클러스터링 알고리즘과 토픽맵을 이용하여 상품 토픽맵 온톨로지를 구축하였다. 기존 웹의 키워드 검색에 의한 정보 검색은 사용자가 원하는 정보를 정확하게 전달하지 못하고, 웹 페이지를 디스플레이 하는 정도의 단점이 있기 때문에, 시맨틱 웹의 장점인 사용자가 원하는 정보를 정확하게 전달하기 위해 온톨로지를 구축하여 시맨틱 웹을 구현하였다. 현재 우리는 다양한 온톨로지 언어를 이용하여 온톨로지를 구축하여 토픽맵 온톨로지와의 통합을 시도하고, 다양한 데이터 마이닝 기법 적용을 시도하고 있다.

### 참고문헌

- [1] 황치근, 김창희, 정계동, 최영근, "부산 환경에서 온톨로지 기반의 메타데이터를 이용한 상품 검색 시스템 설계", 한국정보 과학회, 2004
- [2] 이정현, 방건동, 박세형, 백두권 "온톨로지 기반 설계 문서관리 시스템 설계 및 구현", 한국정보 과학회, 2001
- [3] S. Pepper, B. Moore, "XML Topic Maps(XTM) 1.0", TopicMaps.Org.
- [4] 김병희, "정보이론 기반의 병합식 클러스터링 기법 연구, 서울대학교 컴퓨터 공학부, 2003
- [5] P.Clerkin, P.Cunningham, C.Hayes, "Ontology discovery for the semantic web using Hierarchical Clustering", PKDD, 2002
- [6] Fisher, D.H. "Knowledge acquisition via incremental conceptual clustering", Machine Learning, 2, 139-172, 1987