의미적 시각미디어 웹 서비스를 위한 온톨로지 반자동 생성

김하영°, 이충우, 황재일, 서보원, 나연묵
단국대학교 전자컴퓨터공학과
{hykim°, cwlee, jihwang, bwsuh}@dblab.dankook.ac.kr, ymnah@dku.ac.kr

Semiautomatic Ontology Construction for Semantic Visual Media Web Service

Hayoung Kim°, Chungwoo Lee, Jaeil Hwang, Bowon Suh, Yunmook Nah
Department of Electronics and Computer Engineering, Dankook University

요 약

웹 서비스는 사용자의 요청에 적합한 서비스 제공자의 정보를 제공하여 주는 시스템으로 사용자는 원하는 서비스를 웹 서비스에서 검색, 통합하는 등으로 새로운 서비스를 조합할 수 있다. 이러한 웹 서비스는 다양한 형태의 검색자원을 가질 수 있는데 HERMES는 웹 서비스 시각미디어 검색 시스템의 일종이다. 오늘날의 웹 서비스는 시맨틱 개념을 접목시켜 검색 성능을 향상시키고 정확성을 증대시키기 위해 온톨로지를 주로 활용한다. 시맨틱 개념의 핵심기술인 온톨로지는 단어와 관계들로 구성된 사전으로서 어느 특정 분야에 관련된 단어들을 계층적 구조로 표현한 것이다. 본 논문은 온톨로지의 반자동 생성을 위해 Mining Extractor를 구축하여 HERMES를 개선하는 방법을 제안한다. Mining Extractor는 대상 도메인을 필터링하고 도메인간의 계층구조를 파악하여 온톨로지를 구축하는 것을 목적으로 한다. 이를 위해 워드넷(WordNet)과 데이터 마이닝 기법의 연관규칙을 적용하였다.

1. 서 론

웹 서비스(Web Service)는 네트워크상에서 서로 다른 종류의 컴퓨터들 간에 상호작용을 하기 위한 소프트웨어 시스템으로 서비스 지향적 분산 컴퓨팅 기술의 일종이다[1][2]. 프로세스의 재사용적인 측면을 고려할 때 사용자로부터 검색 키워드를 받아 적절한 웹 서비스를 찾게 하는 것이 매우 중요한 데 웹 서비스 상에서의 검색 방식은 아직 미흡한 실정이다. 이러한 한계점을 극복하기 위해 시맨틱 개념을 접목시켜 검색성능을 향상시키고 정확성을 증대시키기 위한 온톨로지 구축에 관한 연구가 진행되고 있다[3]. 본 연구는 서비스제공자의 서비스에 관한 온톨로지 구축을 반자동으로 하는 데에 목적을 두고 있다. 이를 위하여 웹 서비스 기반의 분산 시각미디어 검색 프레임 워크인 HERMES(tHE Retrieval framework for visual MEdia Service)를 개선하였다. HERMES는 웹 상에 분산되어 있는 멀티미디어 자원을 제공하는 웹 서비스들을 찾아내고 웹 서비스가 가진 자원들을 검색하여 이용자에게 제공하는 역할을 한다[4]. 본 논문이 제안하는 온톨로지 반자동 생성도구인 Mining Extractor는 대상 도메인을 필터링하고 도메인간의 계층구조를 파악하여 온톨로지를 구축하는 것을 목적으로 한다. 이를 위해 워드넷(WordNet)과 데이터 마이닝 기법의 연관규칙(association rule)을 적용하였다[5][6].

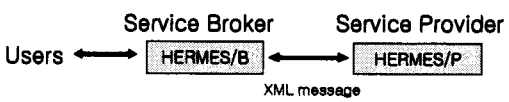
* 본 연구는 정보통신부 및 정보통신연구진흥원의 대학 IT연구센터 육성·지원 사업(IITA-2006-C1090-0603-0031)의 연구결과로 수행되었음

2. 관련 연구

본 장에서는 기존 HERMES의 구조와 온톨로지를 살펴보고 온톨로지 구축시 Mining Extractor에서 사용 될 워드넷과 데이터마이닝 기법의 일종인 연관규칙에 대하여 살펴본다.

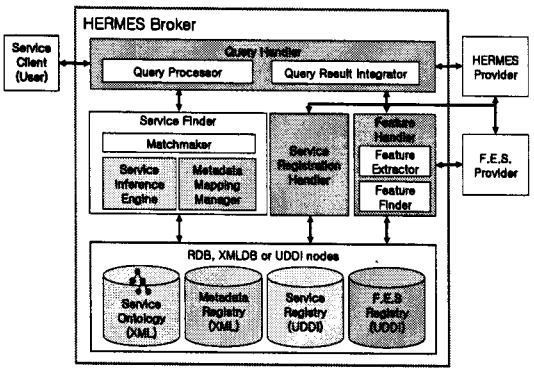
2.1 HERMES

지식기반 시각미디어 검색 프레임워크인 HERMES는 (그림 1)과 같이 웹 서비스를 기반으로 설계되었으며



(그림 1) HERMES 아키텍처[4]

HERMES 중개자(broker)와 HERMES 서비스 제공자(provider)로 구성되어 있다[4]. HERMES 중개자는 사용자의 질의에 맞는 가장 적절한 서비스 제공자(Service Provider)를 찾아 해당 사용자의 질의를 선택된 서비스 제공자가 사용하는 질의 형식으로 변환하여 서비스 제공자에 서비스 요청을 수행하며 사용자와 시각 미디어 서비스 제공자 사이의 중개자(Broker) 역할을 한다. (그림 2)와 같이 사용자가 질의어를 통해 서비스를 요청하면 Query processor는 이를 받아 Service Finder의 Matchmaker에게 서비스 제공자 목록을 요청한다. Matchmaker는 Service Ontology에 접근 가능한 Service Inference Engine과 Service Registry(UDDI)를 이용하여 사용자의 요청에 적합한 서비스 목록을 제공한다[7]. 이후 Metadata Mapping Manager를 이용하여 각 서비스의 제공자마다 정의한 시각 미디어 메타데이타 스키마에 적합한 형태로 사용자 질의를 변환한 뒤 해당 서비스 제공자에게 질의를 전송한다.



(그림 2) HERMES 중개자 아키텍처[4]

이 때 서비스 제공자는 HERMES 중개자로부터 전달 받은 질의를 Provider Specific Ontology를 이용하여 다시 분석하고 이 결과를 가지고 가장 적합한 이미지를 찾아서 HERMES 중개자에게 전달하면 HERMES 중개자는 다시 클라이언트에게 이 결과를 보내주게 된다.

2.2 온톨로지

온톨로지(Ontology)는 단어와 관계들로 구성된 사전으로서 어느 특정 도메인에 관련된 단어들을 계층적 구조로 표현하고 추가적으로 이를 확장할 수 있는 추론 규칙을 포함한다. 이러한 온톨로지는 시맨틱 웹에서 지식을 나타내는 데 주로 사용된다[3][8]. 온톨로지를 표현하기 위해 스키마와 구문구조 등을 정의한 언어가 온톨로지 언어이며 현재 DAML+OIL(DARPA Agent Markup Language + Ontology Interface Layer), OWL(Web Ontology Language)등과 같은 온톨로지 언어가 정의되어 있다. 이 중에서 W3C(World Wide Web Consortium)의 표준안으로 제시한 DAML+OIL은 웹 리소스에 대한 시맨틱 마크업 언어이며 W3C의 RDF(Resource Description Framework)와 RDF 스키마 표준에 기반을 두고 이들을 확장한 프레임 기반의 온톨로지 표현이다[9][10]. 이러한 온톨로지는 Jena, Protege 등과 같은 온톨로지 생성도구를 이용하여 특정 도메인에 대해 구축할 수 있다[11][12].

2.3 워드넷

워드넷(WordNet)은 1985년 Princeton 대학의 George A. Miller, Christiane Fellbaum 등 심리학자, 언어학자, 전산학자등을 중심으로 구축되었으며 2006. 11월 Unix/Linux/Solaris를 위한 3.0버전과 2005. 5월 Windows를 위한 2.1버전까지 발표되어 있다[5]. 기존의 사전이 음절순으로 제작된 것과는 달리 워드넷은 개념을 바탕으로 네트워크를 구축한 대용량 지식베이스이다. 워드넷에서는 동의관계, 반의관계, 상의관계, 하의관계, 분의관계, 양식관계, 함의관계를 표현하고 있다. 현재도 각지에서 세계 각국의 언어로 확장하는 연구가 진행되고 있으며 정보검색, 자동번역, 문장분석 등과 같은 여러 분야에서 활용되고 있다. 또한 온톨로지를 구축하기 위한 기반으로 사용되어 범용성이나 통합문제를 어느 정도 해결할 수도 있다[13].

2.4 연관규칙

연관규칙(association rule)이란 데이터 항목들 간의 조건-결과 식으로 표현되는 유용한 패턴을 말한다. 데이터베이스가 총 n개의 트랜잭션 데이터로 구성되며 전체 m개의 항목으로 구성된다고 하고 이를 I 라 하자. 연관규칙 R은 조건부와 결과부로 구성되며 항목집합인 X와 Y에 대하여 'X가 일어나면 Y도 일어난다'는 의미로 다음과 같이 표현할 수 있다.

$$R : X \Rightarrow Y$$

여기서 $X,Y \subseteq I$ 이고, $X \cap Y = \Phi$이어야 한다. 따라서 연관규칙을 탐사함은 적절한 항목집합 X와 Y를 선택하는 문제로 볼 수 있으며 이를 위해 몇 가지 척도를 고려하고 있다. 우선, 항목집합 X 및 규칙 R에 대한 지지도(support)는 각각 다음과 같이 정의된다.

- support(X) = 집합 X의 항목을 동시에 포함하는 트랜잭션수의 전체 수(n)에 대한 비율
- support(R) = support(X∪Y)

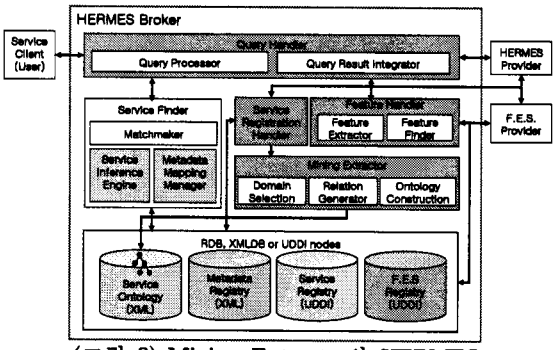
즉, 규칙 R에 대한 지지도는 집합 X 또는 집합 Y에 있는 항목을 동시에 포함하는 트랜잭션수의 비율을 나타낸다. 연관규칙 R의 가치를 평가할 때 통상 다음과 같이 정의되는 신뢰도(confidence)를 사용한다.

- confidence(R) = support(X∪Y) / support(X)

이 신뢰도는 조건부 확률의 개념으로 집합 X(조건)가 발생한다고 할 때 집합 Y(결과)도 동시에 발생할 확률을 의미한다. 즉, 트랜잭션에 X의 항목들을 포함하는 경우 Y의 항목들도 동시에 포함할 확률을 나타내며, 신뢰도가 큰 규칙일수록 의미가 크다고 하겠다 [6][14][15].

3. Mining Extractor

기존 HERMES의 온톨로지는 수동적, 고정적이며 QoS(Quality of Service)요소 추출에 주로 사용되었다. (그림 3)의 Mining Extractor를 이용하여 생성된 온톨로지는 사용자의 연관검색 활용에 그 목적을 두고 있으며 도메인 선정, 관계 생성, 온톨로지 생성의 순으로 수행된다. 온톨로지 구축의 예를 보이기 위해 이미지 등록 사이트인 Flickr[16]의 인기 있는 태그 집합 'Travel', 'Europe', 'France', 'Party'등을 도메인 선정의 자원으로 활용하고 구현에 필요한 도구들을 살펴본다.



(그림 3) Mining Extractor와 HERMES

3.1 도메인 선정

(그림 4)는 Domain Selection 의 수행 알고리즘이다. 서비스 제공자가 등록한 태그 집합 T의 관계를 유추하기 위한 방법으로 태그 집합을 구성하고 있는 Ti, Tj의 아이템 교집합을 점유율 Sij로 표현하여 비교하였다. 점유율이 높을수록 소집합 Ti는 대집합 Tj에 대해 귀속력이 높다.

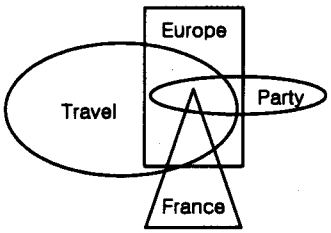
```
Algorithm Select (T)

 // T = {T1, T2, T3, T4, ...}
 // n : number of tag set T
 // Ti : i th tag name
 // Tj : j th tag name
 // Sij = number of (Ti and Tj) / minimum (Ti or Tj)
 //     : share between i th tag and j th tag
 // Smin : minimum share of Service Provider

begin
   for (i=1,j=1; i != j, i<=n, j<=n; i++, j++)
     if (Sij >= Smin)  Tag_Pair_Positioning (Ti, Tj);
end
```

(그림 4) Domain Selection 알고리즘

(그림 5)는 분류 된 태그집합 간의 관계를 도식화한 것으로 집합의 아이템이 공유되어 있음을 볼 수 있다.



(그림 5) 태그 집합 아이템 공유 관계

Select 알고리즘을 수행한 예측 결과는 (표 1)과 같고 서비스 제공자가 정한 최소 점유율 Smin이 45%라고 할 때 최소 점유율 이상의 분류집합을 온톨로지 구축 도메인의 대상으로 삼는다.

(표 1) Ti에 대한 Tj의 점유율 Sij

Ti	Tj	Sij
Travel	Europe	65%
Travel	France	40%
Travel	Party	48%
Europe	France	45%
Europe	Party	50%
France	Party	10%

3.2 관계 생성

(그림 6)은 Relation Generator의 첫 번째 단계인 태그명칭의 용어관계를 설정하는 수행 알고리즘이다. 워드넷을 이용하여 Ti와 Tj의 관계 Rij를 규정하는데 Ti가 Tj의 상위어이면 Rij=1, Ti가 Tj와 동의어이면 Rij=0, Ti

가 Tj의 하위어이면 Rij=-1 로 관계 값을 부여한다. 관계가 설정되지 못한 태그들은 다음 단계를 위해 집합 P로 귀속된다.

```
Algorithm Tag_Pair_Positioning (Ti, Tj)

 // Rij : relation value between Ti and Tj
 // P = { } : set of tag sets not related

 begin
    if (Ti is synonym of Tj) then Rij = 0;
    else if (Ti is superodinate of Tj) then Rij = 1;
    else if (Ti is hyponym of Tj) then Rij = -1;
    else Ti and Tj are inserted to P;
    if (P ≠ null) Tag_Relation_Finding (P);
 end
```

(그림 6) 태그명칭 용어관계 수행 알고리즘

(표 2)는 (표 1)의 결과를 자원으로 하여 Ti와 Tj의 태그명칭 용어관계 알고리즘을 수행한 예측결과 Rij이다.

(표 2) 워드넷을 활용한 관계 생성

Ti	Tj	Rij
Travel	Europe	
Travel	Party	
Europe	France	Europe > France
Europe	Party	
Travel	Trip	Travel = Trip

Rij가 규정되지 않은 태그들은 연관규칙을 이용하는 다음 단계를 수행한다. (그림 7)은 태그를 구성하는 아이템의 빈발도를 반영하여 관계를 규정짓는 연관규칙의 Apriori 알고리즘을 응용한 수행 알고리즘이다.

```
Algorithm Tag_Relation_Finding (P)

 // P = {P1, P2, P3, ...}
 // P1= {a, b, c, ...}, P2={a, c, d, ...}, P3={b, c, e, ...}, ...
 // j : number of frequent itemsets from Apriori algorithm
 // Lj : frequent itemsets of P
 // Cj : confidence of frequent itemset Lj
 // Cmin : minimum confidence of Service Provider

 begin
  run Apriori algorithm on itemsets of each tag
                                  to find related tags;
  eliminate related tags with Cj<Cmin;
  determine parent tag as relatively frequent tags;
 end
```

(그림 7) 연관규칙을 응용한 관계생성 수행 알고리즘

Apriori 알고리즘은 각 집합의 아이템을 바탕으로 아이템 집합 Lj를 유도하고 Lj의 신뢰도 Cj를 계산한다. 이 때 서비스 사용자가 지정한 최소 신뢰도 Cmin 이상인 아이템집합을 포함하는 태그집합들은 관계를 가지고 있다고 할 수 있다.

- 각 태그집합의 아이템

  Travel = {a, b, c, ···}

  Europe = {a, c, d, ···}

  Party = {b, c, e, ···}

- 신뢰도 계산

  support = support({a & c})=66.6%

  confidence = support({a & c}) / support(a)=100%

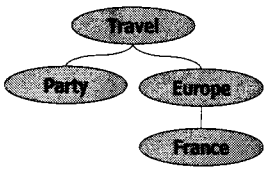
(표 3)은 연관규칙을 응용한 관계생성 알고리즘의 수행 예측 결과이며 신뢰도가 높다고 판단되어 지는 태그집합쌍을 나타내었다. 가장 빈발하게 출현 된 태그를 기준으로 다른 태그와의 관계를 결정한다.

(표 3) 아이템 집합에 대한 신뢰도

Itemset Lj	Confidence Cj	Related Tags
{a, c}	100%	Travel, Europe
{b, c}	100%	Travel, Party

3.3 온톨로지 생성

(그림 7)은 도메인 선정 및 관계를 규명한 결과를 바탕으로 예상되는 온톨로지를 표현한 그림이다.



(그림 7)태그집합의 온톨로지 구현 예

3.4 실험환경 구현

향후 구축 할 실험환경으로 온톨로지 구현이 용이한 자바플랫폼의 Eclipse 도구를 사용한다. 워드넷을 외부 라이브러리로 활용하기 위해 라이브러리를 추가하고 온톨로지 구축을 위해 jena 라이브러리를 추가한다[5][11]. 생성된 온톨로지는 Protege의 Owlviz탭을 이용하여 graph 형태로 확인할 수 있으며 이를 위해 Protege와 Graphviz를 설치한다[12][17].

4. 결론

오늘날의 웹 서비스 사용자는 단순한 분류 기준이 아

닌 의미적으로 연관 있는 대상의 검색을 목적으로 한다. 그러나 몇 몇 포털사이트를 제외한 서비스 검색 방식은 텍스트의 계층적인 분류나 클러스터를 대부분으로 하고 있다. 본 논문에서는 웹 서비스 기반의 시각미디어 검색 엔진인 HERMSE를 개선하여 새로운 구조를 제안하였다. HERMES에 등록되는 서비스를 대상으로 하여 시맨틱 웹을 실현시키는 온톨로지의 구축을 반자동으로 생성하기 위해 Mining Extractor를 추가하였다. Mining Extractor에 의해 반자동으로 생성되는 온톨로지는 서비스 제공자의 등록 비용절감과 날로 증가하는 서비스 통합발전에 기여할 것이다.

[참고문헌]

[1] Header Kreger, IBM Software Group, "Web Services Conceptual Architecture(WSCA 1.0)," http://www-4.ibm.com/software/solutions/webservices/pdf/WSCA.pdf, May 2001.

[2] W3C Web Service WG, "Web Services Architecture," http://www.w3.org/TR/2004/NOTE-ws-arch-20040211/, W3C Working Group Note 11 February 2004.

[3] Berners-Lee, T., Hendler, J. and Lassila, O., "The Semantic Web," Scientific American, 2001.

[4] Nah, Y., Lee, B. and Kim, J., "Visual Media Retrieval Framework using Web Service," LNCS, 3597. pp.104-113, July 2005.

[5] George A. Miller, Christiane Fellbaum, "WordNet", (http://wordnet.princeton.edu)

[6] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules Between sets of items in large databases," in *Proc. 1993 ACM-SIGMOD Int. Conf. Management of Data(SIGMOD'93)*, pp.207-216, Washington, DC, May 1993.

[7] David Ehnebuske, Christopher Kurt, "UDDI Core tModels: Taxonomy and Identifier Systems," UDDI.org(http://www.uddi.org/taxonomies/Core_Taxonomy_OverviewDoc.htm), August 23, 2001.

[8] Bill Andersen, "What is an ontology?," Ontology Works (http://www.ontologyworks.com), January 2001.

[9] Frank van Harmelen, Peter F. Patel-Schneider and Ian Horrocks, "Reference description of the DAML+OIL ontology markup language," March 2001.

[10] Smith, M. K., C. Welty and D. L. McGuinness, "Web Ontology Language (OWL) Guide Version 1.0," 2003.

[11] http://jena.sourceforge.net/, Jena Semantic Web Framework

[12] http://protege.stanford.edu/, The Protege Ontology Editor and Knowledge Acquisition System.

[13] 권혁철, "시맨틱웹의 가능성과 한계", *한국과학기술정보연구원: 지식정보인프라지 통권 15호, 2004*

[14] A. Maedche, S. Staab, "Semi-Automatic Engineering of Ontologies from Text," *Proc. of the 12th Int. Conf. on Software Engineering and Knowledge Engineering, 2000.*

[15] D. Braga, A. Campi, S. Ceri, M. Klemettinen, P. Lanzi, "Discovering Interesting Information in XML Data with Association Rules," *SAC, Proceedings of the 2003 ACM symposium on Applied computing table of contents,* pp.450-454, 2003

[16] http://www.flickr.com

[17] http://www.graphviz.org