한국지능시스템학회 논문지 2008, Vol. 18, No. 2, pp. 157-162

대규모 태깅 데이터를 이용한 태깅 온톨로지 학습

Learning Tagging Ontology from Large Tagging Data

강신재*
Sin-Jae Kang
* 대구대학교 컴퓨터·IT공학부

요 약

본 논문은 대중에 의해 자유롭게 생성된 분류 체계인 폭소노미, 즉 대규모의 태깅 데이터로부터 태깅 온톨로지를 학습하는 방법을 제시하고 있다. 기존 소셜웹 시스템 간에는 태깅의 의미에 대해 공통의 합의가 이루어지지 않았기 때문에, 시스템마다 태깅 정보를 표현하기 위해 내부적으로 다른 방법을 쓰고 있으며, 따라서 소프트웨어 에이전트를 이용하여 시스템간의 정보처리를 자동으로 할 수가 없다. 이를 해결하는 방법으로 폭소노미를 위한 태깅 온톨로지가 필요하다. 태깅의 본질적인 속성을 분석하여 태깅 온톨로지를 정의하고, 태깅 데이터의 기계 학습을 통하여 유사 태그와 사용자 그룹 정보를 획득한 후, 태깅 온톨로지를 학습한다. 이의 활용 방안으로 학습된 태깅 온톨로지를 이용하여 모델링한 추천 시스템도 제안한다.

키워드 : 태깅 온톨로지, 온톨로지 학습, 폭소노미, 소셜웹, 클러스터링

Abstract

This paper presents a learning method of tagging ontology using large tagging data such as a folksonomy, which stands for classification structure informally created by the people. There is no common agreement about the semantics of a tagging, and most social web sites internally use different methods to represent tagging information, obstructing interoperability between sites and the automated processing by software agents. To solve this problem, we need a tagging ontology, defined by analyzing intrinsic attributes of a tagging. Through several machine learning for tagging data, tag groups and similar user groups are extracted, and then used to learn the tagging ontology. A recommender system adopting the tagging ontology is also suggested as an applying field.

Key Words : Tagging Ontology, Ontology Learning, Folksonomy, Social Web, Clustering

1. 서 론

인터넷이 보편화되고 성장함에 따라 다양하면서도 방대한 양의 웹 자원(resource)이 급격히 늘어나게 되었다. 이러한 자원을 관리하기 위해서는 분류, 색인, 검색, 추론 등의 작업들이 필연적인데, 웹 자원을 개념적으로 기술(description)하고 공유할 수 있는 방법이 확보된다면 이들 작업들을 보다 정확하고 효율적으로 수행할 수 있게 된다. 그나마 텍스트로부터 이러한 개념 기술을 자동으로 추출하기는 용이한 편이어서 여러 연구들이 이루어져 왔으나, 이미지, 동영상, 오디오 등 멀티미디어 형태의 자원으로부터 개념 기술을 추출하기에는 아직까지 어려움이 많다.

소셜웹(social web, web 2.0)과 폭소노미(folksonomy)는 최근 인터넷 분야에서 핫이슈가 되고 있는 키워드로, 소셜웹 이전의 웹이 일방적인 정보 제공의 형태였다면 소셜웹은 사용자들의 자발적인 참여와 개방성을 통해 블로그 등을 활용하여 정보 및 네트워크를 창조하고 공유하는 특성을 지닌다. 폭소노미는 대중(folk)에 의한 분류(taxonomy)를 의미하며, 각 사용자가 관심있는 웹 자원에 대해 자유 형태의 키워드인 태그(tag)를 자발적으로 부여하고 이를 대중이 함께 공유하는 형태를 가진다. 여기서 사용한 태그는 종래의 카테고리와는 다른 자유로운 분류법으로, 사용자가 웹 자원의 내용에 대해 이해한 후 수작업으로 부여하는 것이니 만큼 온톨로지 자동 구축, 추천 시스템(recommender system) 등의 분야에서 아주 중요한 정보로 활용될 수 있다.

del.icio.us(딜리셔스)[1]는 소셜 북마킹으로 유명한 웹사이트로 북마크를 저장하고, 다른 사용자들과 함께 공유하고, 다른 사용자들의 북마크를 볼 수 있는 웹사이트이다. 카테고리를 만들어 관리하는 대신 자유롭게 태그를 사용해 북마크를 관리할 수 있다. 이 사이트의 전체적인 소스 코드는 공개되어 있지 않지만, 사용자의 데이터를 API를 통해 XML이나 JSON 포맷으로 받을 수 있다[1].

Flickr(플리커)[2]는 온라인 사진 공유 커뮤니티 사이트로

접수일자 : 2007년 11월 20일
완료일자 : 2007년 12월 31일
이 논문은 2007학년도 대구대학교 학술연구비 일부지원에 의한 논문임
I thank Dr. Ying Ding at DERI Innsbruck for helpful discussions and providing tagging data.

1) http://del.icio.us
2) http://www.flickr.com

소셜웹의 대표적인 사이트이다. 이 서비스는 개인 사진을 교환하는 목적 이외에도 사진을 올려 저장하는 용도로 쓰이기도 한다. 사용자는 태그를 이용해서 사진들을 분류하는 것이 가능한데, 이것은 나중에 검색자가 장소 이름이나 주제 같은 것을 가지고 검색하는 일을 용이하게 해준다.

폭소노미를 이용하는 이러한 태깅 시스템의 주요한 문제점으로는 다의어(polysemy), 동형이의어(synonymy), 개념 범위의 불일치(discrepancies in granularity)가 있다[2]. 예를 들어 'apple'이라는 단어는 과일을 의미할 수도 있지만 컴퓨터를 의미할 수도 있기 때문에 사용자 검색 시 원치 않는 결과를 초래할 수도 있다. 단수/복수, 대문자/소문자와 같은 형태에 따라 동일한 의미이지만 다르게 인식되는 문제도 있으며, 너무 일반적이거나 혹은 너무 구체적인 단어를 태깅에 사용하여 발생하는 문제도 있을 수 있다. 이러한 문제를 해결하기 위한 한 방법으로 스태머(stemmer)과 워드넷(WordNet)을 이용하여 태깅에 사용된 태그를 전처리한 후, 태그의 공기 정보(co-occurrence information)를 분석하여 태그를 클러스터링하고, 그들 간의 의미관계를 추출하여 태깅 온톨로지의 학습에 사용하는 방법을 본 연구에서 사용하고자 한다. 이러한 결과물은 검색 시 질의어(태그) 확장과 태깅 시 연관 태그의 추천, 추천 시스템의 모델링 등에 활용될 수 있다. 태깅 온톨로지는 범용 온톨로지라기 보다는 태그를 사용하는 웹 사이트와 웹 애플리케이션에서 활용되는 도메인 온톨로지라고 할 수 있으며, 태그를 사용하는 서로 다른 웹 사이트 간 원활한 정보의 교류와 처리를 위해 사용되는 지식베이스이다.

딜리셔스는 하나의 웹 자원(북마크)에 대해 여러 사용자가 태깅을 할 수 있기 때문에 넓은 폭소노미(broad folksonomy)로 분류되고, 플리커는 주로 웹 자원(사진)의 생성자만이 태깅을 하기 때문에 좁은 폭소노미(narrow folksonomy)로 분류될 수 있다[3]. 따라서 플리커보다 딜리셔스의 태깅 정보가 웹 자원에 대한 여러 사용자의 다양한 견해를 표현하고 있다고 볼 수 때문에, 온톨로지의 구축 및 학습에 활용 가능한 양질의 정보로 간주할 수 있다. 본 연구에서는 태깅의 본질적인 속성을 분석하여 태깅 온톨로지를 정의하고, 딜리셔스 사이트로부터 폭소노미 정보를 자동으로 추출한 후, 기계 학습을 통하여 유사 태그와 사용자 그룹 정보를 획득하여, 태깅 온톨로지의 학습에 사용한다. 이의 활용 방안으로 학습된 태깅 온톨로지를 이용하여 모델링한 추천 시스템도 제안한다.

2장에서는 태그 온톨로지를 정의하게 된 기본 아이디어를 설명하고, 3장에서는 방대한 태깅 데이터로부터 태그 온톨로지를 학습하는 방법에 대해 설명한다. 4장에서는 태깅 온톨로지의 활용 방안으로 추천 시스템을 모델링하고, 5장에서는 관련된 기존 연구를, 6장에서는 결론과 향후 연구계획을 제시한다.

2. 태깅 온톨로지

기존 소셜웹 시스템간에는 태깅의 의미에 대해 공통의 합의가 이루어지지 않았기 때문에, 시스템마다 태깅 정보를 저장하기 위해 내부적으로 다른 표현법을 쓰고 있으며, 소프트웨어 에이전트를 이용하여 시스템간의 정보처리를 자동으로 할 수가 없다. 따라서 이를 해결하는 방법으로 폭소노미를 위한 태깅 온톨로지가 필요하다.

현재 전산처리를 위한 온톨로지의 정의 가운데 가장 공감대를 얻고 있는 것은 Gruber의 정의이다. Gruber는 온톨로지를 "특정 분야(domain)에서 공유된(shared) 개념(conceptualization)들을 전산처리가 가능한 형태로 형식화(formal)시키고 명시한(explicit) 지식베이스"로 정의하고 있다[4]. 본 논문에서 정의한 태깅 온톨로지는 범용 온톨로지라기 보다는 폭소노미를 적용한 소셜웹 사이트와 웹 애플리케이션에서 태깅 정보를 공유하기 위해 사용하는 도메인 온톨로지라고 할 수 있으며, 서로 다른 웹 사이트 간 원활한 정보의 교류와 처리를 위해 필수적인 지식베이스이다.

소셜웹 사이트의 기본 동작은 사용자(태거)가 관심있는 웹 자원(객체)[3]을 사이트에 추가하고 임의의 단어(태그)를 그 자원에 할당(태깅)하는 형태로 이루어진다. 즉 태거, 객체, 태그가 동시에 하나의 태깅에 관여하게 되는 것이다. 이를 개념적으로 표현하기 위해 태거, 객체, 태그를 태깅 클래스를 중심으로 묶고, 소셜 시스템의 태깅 정보로부터 획득할 수 있는 태깅한 날짜 등의 정보를 추가하여 그림 1과 같은 태깅 온톨로지를 정의하였다.

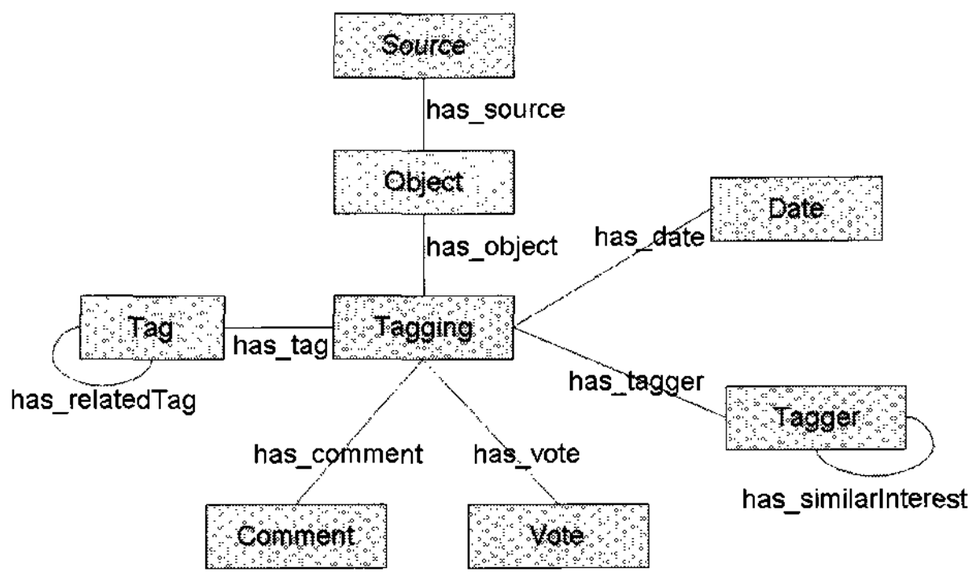

그림 1. 태깅 온톨로지

Fig. 1. Tagging ontology

하나의 실제 태깅은 태깅 온톨로지의 인스턴스에 해당하며, 한 사용자가 생성한 모든 태깅 정보를 모은 것이 개인에 의한 분류체계라면, 이러한 개인 분류를 모두 모은 것이 폭소노미라고 정의할 수 있다.

태깅 온톨로지의 대부분의 정보는 소셜웹 사이트로부터 추출된 태깅 정보로부터 확보할 수 있으나, 추천(recommendation) 시스템의 구현에 중요한 역할을 할 수 있는 "has_relatedTag", "has_similarInterest"와 같은 의미관계는 직접 얻을 수가 없다. 따라서 추출된 태깅 정보를 가공하여 기계 학습을 거치면서 해당 의미 관계를 추출하고자 한다. 태깅 정보(즉 태깅 온톨로지의 인스턴스)는 del.icio.us 자바 API[4]를 이용하여 구현된 인터넷 에이전트 크롤러(crawler)를 통해 자동으로 추출되어 RDF(resource description framework)[5]의 형태로 저장된다. 이는 W3C에 의해 표준화되고 있는 시맨틱웹 기반 기술을 적극 채택하여 OWL로의 확장 및 추론 등의 작업을 수행할 수 있는 토대를 마련하기 위함이다. RDF로 표현된 태깅 온톨로지 인스턴스의 예는 그림 2와 같다.

실제로 크롤러를 통해 del.icio.us 사이트로부터 462,733

3) 어떠한 웹 자원을 추가하고 공유하느냐에 따라 서비스의 종류를 구분해 볼 수 있는데, del.icio.us는 북마크를, Flickr는 사진을 공유하는 대표적인 소셜웹 사이트이다.

4) del.icio.us Java API (http://sourceforge.net/projects/delicious-java)

명의 사용자(태거), 404,388개의 태그, 483,564개의 북마크(객체)가 포함된 총 9,400,029개[5]의 태깅 인스턴스를 추출하였다. 정의된 태깅 온톨로지는 추후 FOAF[6]나 SIOC[6]에서 사용하고 있는 개념 스키마의 클래스들과의 연관도에 따라, 온톨로지 병합(merge)이나 온톨로지 매핑(mapping) 과정을 거쳐 확장될 수 있으며, 이를 통해 손쉽게 다른 사이트들과 정보를 공유할 수 있게 된다.

```
<rdf:RDF
   xmlns:j.0="http://uto.deri.at/"
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" >
  <rdf:Description rdf:about="http://uto.deri.at/2fdc4bc3-c33e-4b37-aa6f-b0e9e44179d4">
    <j.0:has_date>May 07</j.0:has_date>
    <j.0:has_comment></j.0:has_comment>
    <j.0:has_vote>15</j.0:has_vote>
    <j.0:has_tag rdf:resource="http://del.icio.us/tag/magazine"/>
    <j.0:has_tagger>mzarynn</j.0:has_tagger>
    <j.0:has_object rdf:resource="http://www.faeriemagazine.com/"/>
    <j.0:has_tag rdf:resource="http://del.icio.us/tag/faery"/>
    <j.0:has_tag rdf:resource="http://del.icio.us/tag/illustration"/>
    <j.0:has_tag rdf:resource="http://del.icio.us/tag/art"/>
    <j.0:has_tag rdf:resource="http://del.icio.us/tag/fantasy"/>
  </rdf:Description>
  <rdf:Description rdf:about="http://uto.deri.at/038bca3d-9eb9-4a08-9874-5b977c2848a2">
    <j.0:has_tag rdf:resource="http://del.icio.us/tag/tools"/>
    <j.0:has_tag rdf:resource="http://del.icio.us/tag/socialising"/>
    <j.0:has_tag rdf:resource="http://del.icio.us/tag/savvy"/>
    <j.0:has_comment>How to Change the World: The Art of Schmoozing II</j.0:has_comment>
    <j.0:has_tagger>anubiros</j.0:has_tagger>
    <j.0:has_tag rdf:resource="http://del.icio.us/tag/interesting"/>
    <j.0:has_object rdf:resource="http://blog.guykawasaki.com/2007/06/the_art_of_schm.html"/>
    <j.0:has_vote>141</j.0:has_vote>
    <j.0:has_tag rdf:resource="http://del.icio.us/tag/smooze"/>
    <j.0:has_date>Jun 07</j.0:has_date>
    <j.0:has_tag rdf:resource="http://del.icio.us/tag/socializing"/>
  </rdf:Description>
```

그림 2. 태깅 온톨로지 인스턴스
Fig. 2. Instances of the tagging ontology

3. 온톨로지 학습

소셜웹 사이트의 태깅 정보로부터 직접 추출하지 못하는 태깅 온톨로지의 "has_relatedTag"와 "has_ similarInterest" 관계를 학습하는 과정이다. 전반적인 절차는 그림 3에 나타나 있으며, 3.1절에서는 태그 클러스터를 통하여 "has_relatedTag" 관계를 추출하는 과정을, 3.2절에서는 태거 클러스터를 통하여 "has_ similarInterest" 관계를 추출하는 과정에 대해 자세히 설명한다.

3.1 태그 클러스터 (Tag Cluster)

웹 사이트에서 사용되는 방대하고 다양한 태그를 수작업으로 분류하는 것은 일관성, 비용, 시간 등의 여러 문제로 인해 실용적이지 못하므로, 소셜웹 사이트의 태깅 정보(태깅 온톨로지 인스턴스)를 가공하여 자동으로 태그를 분류하고자 한다. 태그를 분류하는데 사용될 수 있는 정보를 태깅 온톨로지에서 살펴보면, 하나의 태깅에 관련된 정보로 태거, 객체, 날짜, 주석 등 여러 가지가 있으나, 객체[7]의 종류와 내용에 따라 태거가 해당 객체에 태그를 부여하는 것이기 때문에, 객체가 태그의 특성(쓰임새)을 가장 잘 나타내주는 정보라고 볼 수 있다. 그래서 본 논문에서는 객체와 태그의 공기정보를 이용하여 태그 벡터를 구성하였다.

태그 클러스터를 생성하기 위해 사용될 태그 벡터를 구성하기에 앞서서, 태그에 대한 전처리 과정을 밟는다. 이는 일반적이지 않은 태그를 제거하고, 형태론적으로 유사한 태그들을 정리하여 후보 태그를 선택하는 과정(그림 4)이다.

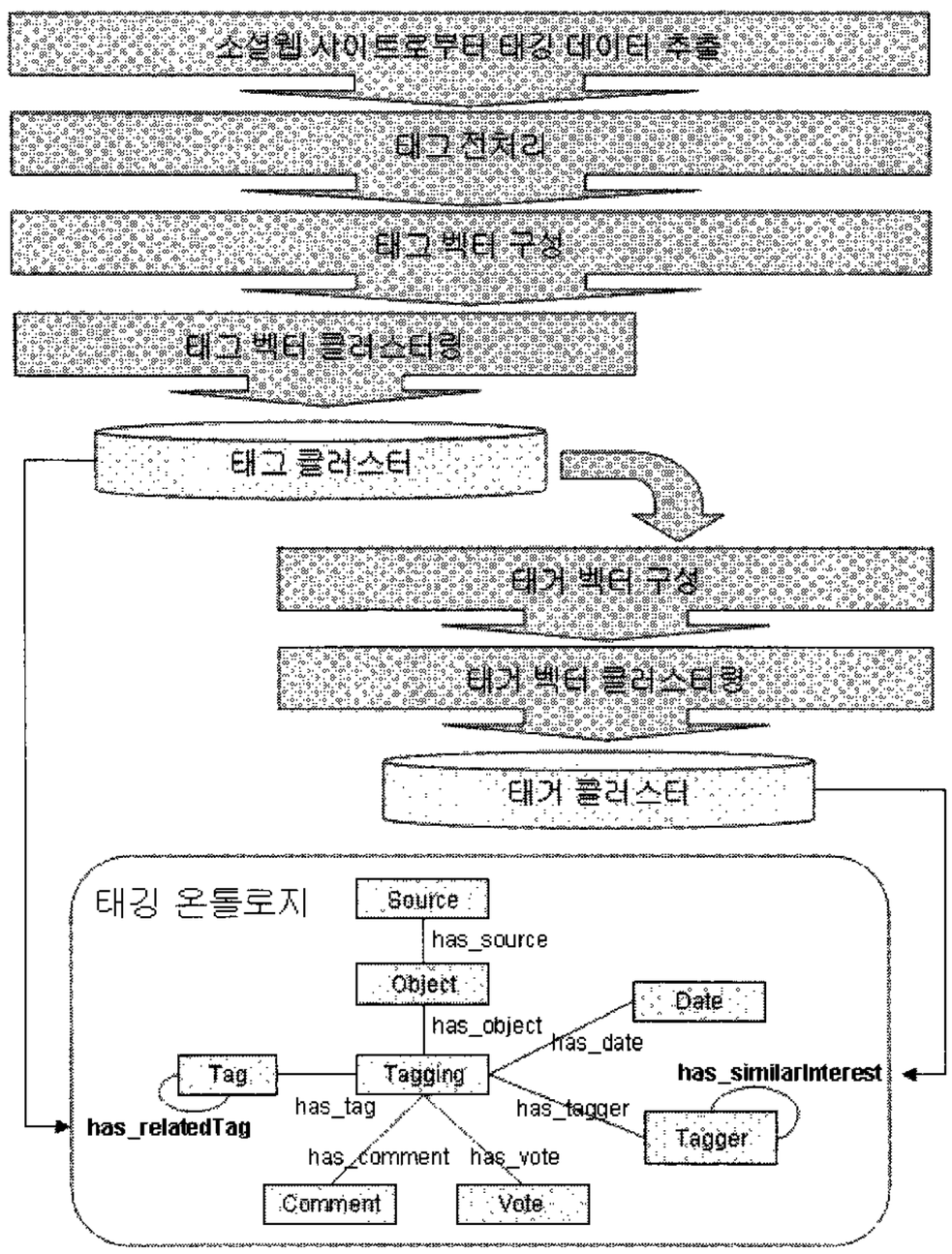

그림 3. 태깅 온톨로지 학습과정
Fig. 3. Learning tagging ontology

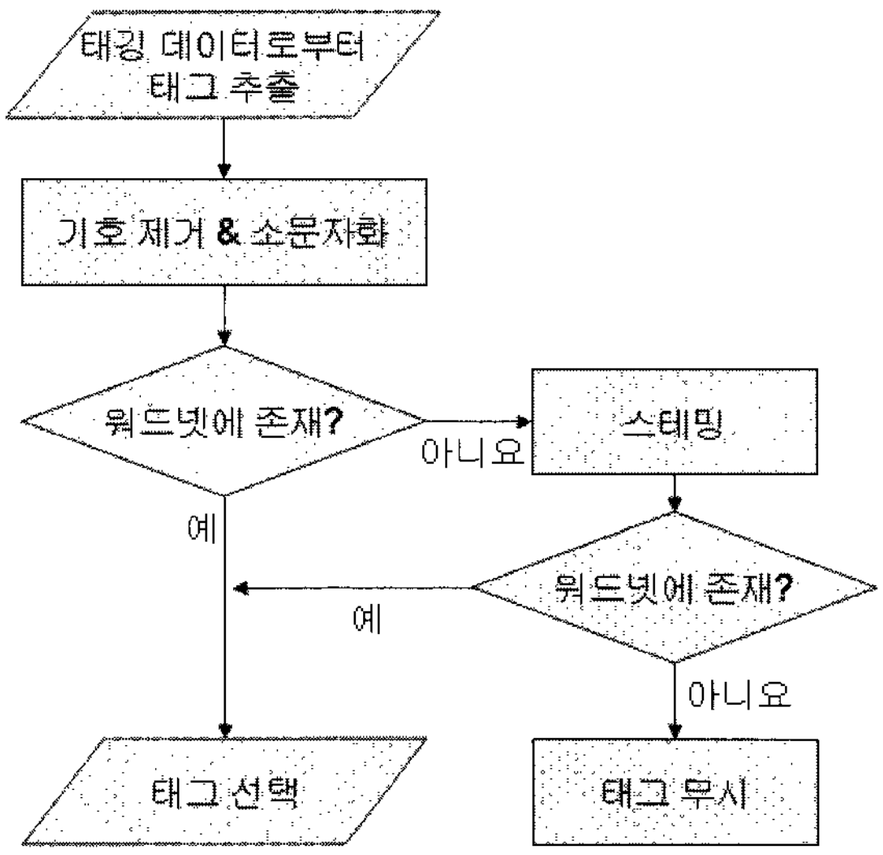

그림 4. 태그 전처리 과정
Fig. 4. Tag pre-processing

소셜웹 사이트의 태깅 정보로부터 추출된 태그들을 대상으로 먼저 !, $, % 등과 같은 기호를 제거하고 모든 대문자를 소문자로 변환한다. 그 다음 태그가 워드넷(WordNet)[7]에 존재하는지 확인[8]하여 존재하면 후보 태그로 선택하고,

5) RDF로 저장된 태깅 온톨로지 인스턴스 파일의 전체 크기는 1.64GB이다.
6) Semantically-Interlinked Online Communities (http://sioc-project.org)
7) 본 논문에서는 딜리셔스 사이트의 태깅 정보를 대상으로 하므로 여기서 객체는 북마크를 의미한다.

존재하지 않으면 스테밍(stemming)[9]을 하게 된다. 워드넷에 존재하는 태그를 선택하는 이유는 워드넷에 등록된 단어가 일반적으로 중요하고 대표성이 있는 단어로 간주할 수 있고, 또한 워드넷을 이용하여 단어간 유사도 계산이 가능해지기 때문에 추후 클러스터내 태그의 랭킹이나 태그 클러스터의 랭킹이 가능해지기 때문이다. 스테밍을 먼저 하지 않은 이유는 스테밍 성능이 완벽하지 않기 때문에 불완전하게 스테밍된 소수의 태그는 워드넷에서 검색할 수가 없기 때문이다. 스테밍을 거친 태그는 다시 워드넷에 존재 여부를 확인하여 최종적으로 후보 태그로 선택된다. 이와 같은 과정을 통해 총 26,691개의 태그가 선택되었다.

선택된 모든 태그와 모든 객체를 대상으로 태그 벡터를 구성하면, 기계 학습을 수행하는 서버 메모리의 용량 제한 때문에 클러스터링 알고리즘을 실행하기가 어려웠다. 그래서 빈도수가 낮은 태그와 객체를 제외하기 위해 일정 빈도 이상을 대상으로, 각 태그와 객체의 공기 빈도수를 정규화하여 태그 벡터를 구성하였다.

태그 벡터를 클러스터링하기 위해서는 여러 기계학습 알고리즘을 적용시켜 보았는데, Witten[8]이 개발한 WEKA(Waikato Environment for Knowledge Analysis) 패키지의 여러 클러스터링 알고리즘 가운데 가장 좋은 성능을 보인 X-mean를 이용하여 실험하였다. WEKA는 실제 응용 프로그램에서 기계학습 알고리즘의 구현을 돕기 위해 만들어진 도구이다. X-means 알고리즘은 Pelleg과 Moore[9]가 개발하였는데, 기존 K-means 알고리즘의 세 가지 주요한 단점, 즉 느리고 확장이 쉽지 않고, 클러스터의 수 K를 사용자가 정해야 되며, 국부해(local minima)에 빠지기 쉽다는 단점들을 개선한 클러스터링 알고리즘이다. 서버에서 실험이 가능한 크기인 4,676개의 태그와 3,616개의 객체를 대상으로 태그 벡터를 구성하였고, 총 98개의 태그 클러스터를 얻을 수 있었다. 하나의 태그 클러스터에 속한 태그들은 유사한 종류의 객체를 태깅할 때 같이 사용되는 경우가 많았다는 것을 의미하므로 상호간 "has_relatedTag" 관계를 갖는 것으로 간주할 수 있다.

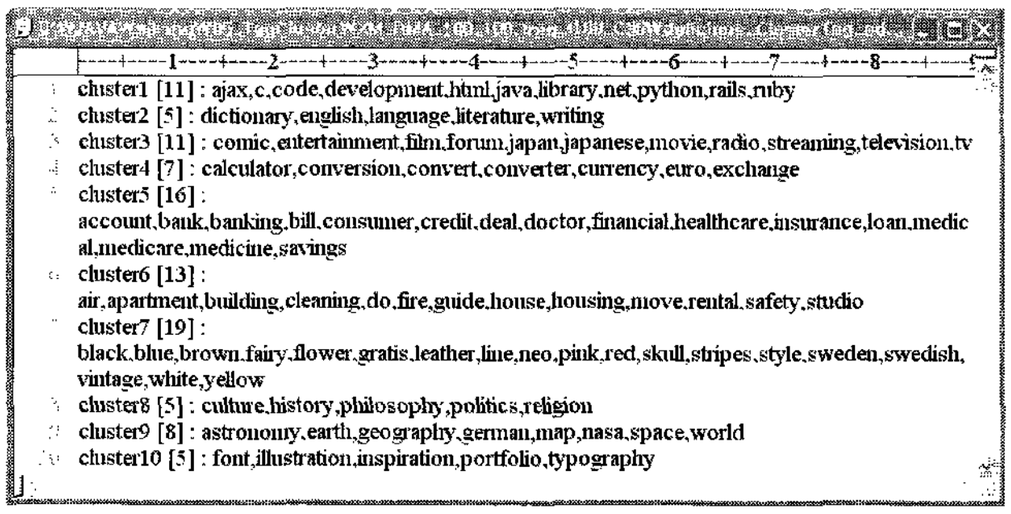

그림 5. 태그 클러스터 예시

Fig. 5. Example of tag clusters

3.2 태거 클러스터 (Tagger Cluster)

사용자가 태깅 시 주로 사용하는 태그가 해당 사용자의 관심 분야를 잘 나타낼 수 있기 때문에 태그를 사용하여 태거 벡터를 구성하였으며, 이는 유사한 성향을 보이는 태거(사용자)의 그룹을 얻기 위해 사용된다. 그런데 모든 태그를 사용하여 태거 벡터를 구성하기에는 벡터의 차원이 너무 커지기 때문에 현실적으로 기계학습이 어려운 문제점이 있다. 따라서 벡터의 차원을 줄이는 한 방법으로 3.1절에서 획득한 태그 클러스터를 이용하여 태거 벡터를 구성하였다(그림 6).

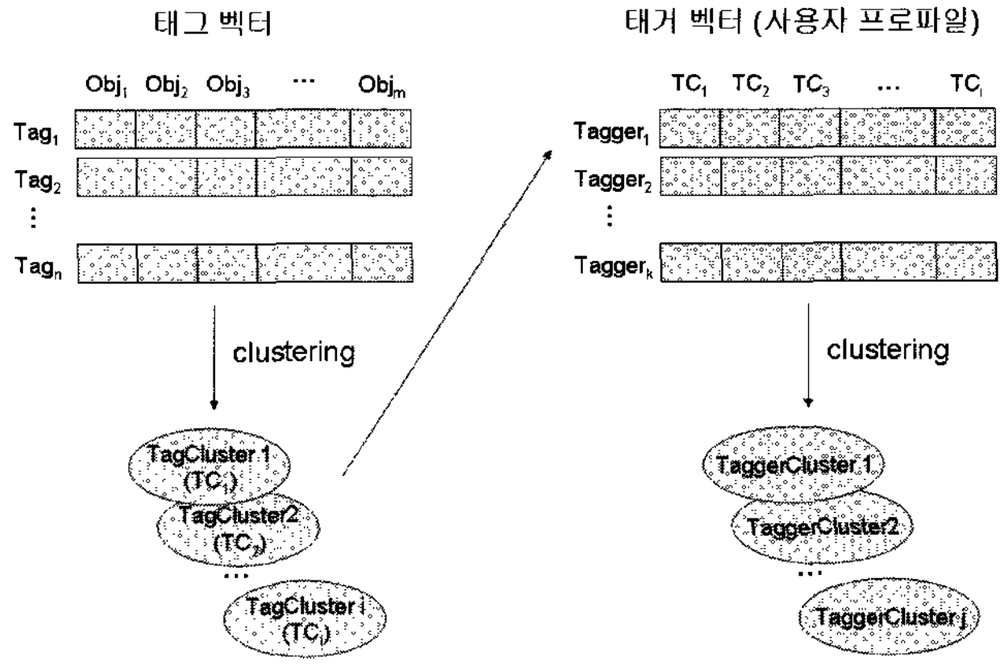

그림 6. 태그 및 태거 벡터 구성

Fig. 6. Construction of tag & tagger vectors

태거 벡터의 클러스터링을 위해서는 3.1절과 같이 X-means 알고리즘을 적용하였으며, 98개의 태그 클러스터와 72,449명의 태거를 대상으로 태거 벡터를 구성하여, 총 1,223개의 태거 클러스터를 얻었다. 태거 벡터는 사용자의 관심분야와 취향을 태그 클러스터를 이용하여 나타낸 것이므로 태깅 정보로부터 자동으로 추출된 사용자 프로파일로 볼 수 있다. 또한 태거 클러스터는 유사한 성향을 갖는 사용자들을 그룹핑한 것으로 태깅 온톨로지의 "has_similarInterest" 관계 정보를 담고 있다. 사용자 프로파일과 유사 사용자 그룹 정보는 추천 시스템을 모델링하는데 있어 아주 중요한 역할을 한다.

4. 추천시스템 모델링

태깅 온톨로지를 활용하기 위한 한 분야로 추천 시스템을 모델링하고자 한다. 이는 태그 클러스터를 이용하여 구성된 각각의 태거 벡터는 사용자의 성향(선호도)을 나타내는 사용자 프로파일의 역할을 할 수 있으며, 태거 클러스터는 유사한 성향을 가지는 사용자 그룹의 정보를 표현하는 그룹 프로파일의 역할을 할 수 있기 때문이다.

추천(recommendation) 문제는 기본적으로 평가(rating) 구조에 기반을 두고 있으며, 사용자가 접하지 못한 새로운 아이템에 대한 평가의 추정 문제로 볼 수 있다. C를 모든 사용자의 집합으로, S를 추천될 수 있는 모든 아이템의 집합(예: 북마크, 책, 영화, 식당 등)으로, u를 사용자 c에 대한 아이템 s의 유용성을 평가하는 함수라 가정할 때, 모든 사용자 c에 대해 각 사용자의 만족도를 최대화할 수 있는 아이템 s′를 찾는 다음과 같은 식으로 추천 문제를 형식화할 수 있다.

$$\forall c \in C,\ s'_c = \arg\max_{s \in S} u(c, s) \quad (1)$$

추천이 어떻게 만들어지느냐에 따라 내용 기반(content-based), 협력 기반(collaborative), 하이브리드

8) 2006년 12월에 릴리스된 WordNet 3.0 버전을 사용하였다. (http://wordnet.princeton.edu/obtain)

9) Snowball Stemmer (http://snowball.tartarus.org)

(hybrid) 형태의 세 가지 접근법으로 분류해 볼 수 있는데, 내용 기반의 추천 시스템은 사용자의 프로파일과 과거 선호도에 따라 유사한 아이템을 추천하는 방식이며, 협력 기반은 한 사용자와 유사한 성향을 갖는 다른 사용자들의 프로파일과 과거 성향에 따라 아이템을 추천하는 방식이다. 하이브리드 형태는 위의 두 가지 방법을 결합한 형태의 접근법이다[10].

본 논문에서는 최적의 추천 정보를 생성하기 위하여 태깅 온톨로지로부터 사용자 프로파일과 그룹 프로파일을 추출하여 이용하는 하이브리드 접근법을 제안한다. 추천 시스템의 전체적인 구성은 그림 7에 나타나있으며, 크게 추천 에이전트, 온톨로지 관리 에이전트, 프로파일 관리 에이전트로 나뉜다. 본 논문의 2장과 3장에서 상세하게 다룬 부분은 온톨로지 관리 에이전트에 해당하며, 프로파일 관리 에이전트는 온톨로지 관리 에이전트와의 정보교환을 통해 각 사용자의 프로파일(태거 벡터)과 사용자가 속한 그룹("has_similarInterest" 관계로 연결된 태거의 집합)의 프로파일 정보를 얻은 후 프로파일을 DB로 만들어 관리한다. 추천 에이전트는 추천 요청이 있을 시 프로파일 관리 에이전트를 통해 해당 사용자의 프로파일과 소속된 그룹 프로파일 정보를 검색하여 추천 템플릿을 생성한 후, 객체 DB[10]에서 추천 대상을 선정한다. 추천된 객체 가운데 사용자가 직접 선택(구매, 저장 등)한 경우에는 해당 정보가 프로파일 관리 에이전트로 피드백되어 기존 프로파일 정보가 갱신된다.

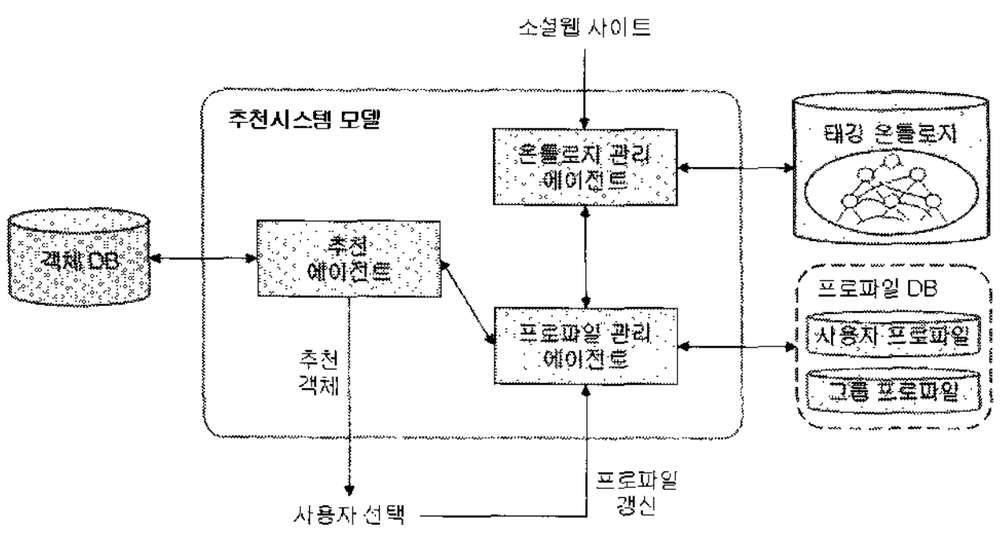

그림 7. 태깅 온톨로지를 이용한 추천 시스템
Fig. 7. Recommender system using the tagging ontology

5. 관련 연구

다수의 사용자에 의해 생성된 대량의 태깅 데이터, 즉 폭소노미에 내재되어 있는 의미 관계를 이끌어 내는 것과 관련하여 "떠오르는 시맨틱(emergent semantics)"[11]이라는 용어가 최근 많이 사용되고 있다. 이와 관련된 연구로, Mika[12]는 시맨틱 웹을 커뮤니티, 시맨틱, 컨텐츠의 세 계층으로 구분하고, 각각 사용자(actor), 태그(concept), 자원(instance)의 클래스로 대응시켜 형식화하였다. 딜리셔스 사이트로부터 이들 간의 공기정보를 가지고 그래프를 형성한 후, 네트워크 분석 기법을 이용하여 경량 온톨로지(lightweight ontology)를 이끌어 내는 방법을 제시하였다.

Gruber[13]는 실제로 태깅 온톨로지를 제시하지는 않았지만, 태그에 대한 선택정도(vote)를 표현할 수 있는 부정적 태깅(negative tagging) 개념과 태깅이 어느 시스템/사이트에서 이루어졌는지를 나타낼 태깅의 소스(source) 개념 등 태깅 온톨로지에 포함되어야 하는 내용을 제안하였다.

Wu[14]는 사용자가 태깅한 데이터로부터 의미 관계 정보를 추출하기 위해 "사용자(user), 자원(resource), 태그(tag), 시간(time)"과 같은 네 개의 구성요소로 이루어진 쌍을 기본으로 사용하였으나, 역시 구체적인 온톨로지를 제시하지는 않았다.

Knerr[15]는 시맨틱 웹 기술의 하나인 FOAF[6]를 이용하여 사용자 프로파일을 표현하고, 각 사용자의 태깅 데이터를 따로 관리하는 구조를 제안하였다. 온톨로지의 주요 클래스로는 "시간(time), 사용자(user), 도메인(domain), 가시/접근성(visibility), 태그(tag), 자원(resource), 유형(type)"을 정의하고 사용하여 소셜 시스템간 상호호환이 이루어질 수 있게 하였다.

Specia[16]는 시맨틱 웹 환경에서 기존 폭소노미들을 통합하기 위한 방법론을 제시하였다. 소셜웹 사이트로부터 추출한 태그를 공기정보를 이용하여 클러스터링한 후, 태그 간에 내재하고 있는 관계정보를 얻기 위해 위키피디아(Wikipedia)나 구글(Google), 시맨틱 웹 검색 엔진(Swoogle)을 이용하여, 기존 온톨로지와 지식베이스에 존재하는 개념과 태그를 매핑하고 의미 관계를 검색하였다. 아직은 초기단계의 연구이며 클러스터링 알고리즘의 개선 및 폭소노미 통합 전과정을 자동화하기 위해서는 추가의 연구가 필요하다.

6. 결론 및 향후계획

폭소노미를 사용하는 소셜웹 사이트에 존재하는 방대한 태깅 정보로부터 자동으로 의미 관계 정보를 추출하기 위하여, 태깅 온톨로지를 정의하고, 태깅 정보를 자동으로 추출한 후, 클러스터링 알고리즘을 적용하여 온톨로지를 학습하는 방법론을 제시하였다. 태그 간, 태거 간에 존재하는 연관관계를 자동으로 추출하였기 때문에 수작업을 배제한 실용적인 방법론이며, 또한 방대한 양의 정보를 사용하여 보다 일반적이고 객관적인 정보를 추출했다고 볼 수 있다.

기존 정보검색 시스템은 기본적으로 단어기반의 검색을 하므로, 동형이의어나 다의어를 질의어로 사용할 경우, 의미상 관련이 없음에도 표제어만 같다는 이유로 서로 다른 의미의 정보들이 혼재되어 검색되는 문제가 있다. 그러나 태깅 온톨로지의 태그 클러스터를 이용하면 질의 확장 등을 통하여 보다 정확한 정보검색 시스템의 구현이 가능하게 된다. 또한 사용자 프로파일과 그룹 프로파일을 태깅 온톨로지로부터 생성할 수 있기 때문에 이를 활용한 추천 시스템의 모델링이 쉬워지는 장점이 있다.

현재의 연구는 영어를 대상으로 하고 있으나, 워드넷과 같은 역할을 할 수 있는 자원만 있다면 한국어 사이트에의 적용도 어려움이 없다. 향후에는 대용량의 학습 데이터를 처리할 수 있는 클러스터링 알고리즘을 개발하여 모든 태깅 데이터를 대상으로 실험을 계속할 계획이며, 워드넷을 이용한 태그간 랭킹과 태그 클러스터간 랭킹을 활용할 방안을 연구하고자 한다.

10) 추천 시스템을 적용하는 분야에 따라 도서와 같은 물품이나 북마크, 사진과 같은 다양한 객체를 사용할 수 있다.

참 고 문 헌

[1] http://ko.wikipedia.org/wiki/Del.icio.us

[2] S. Gloder, and B. A. Huberman, "Usage Patterns of Collaborative Tagging Systems." *Journal of Information Science*, Vol.32, No.2, pp. 198-208, 2006.

[3] T. V. Wal, Folksonomy Explanations, http://www.vanderwal.net/random/ entrysel.php?blog=1622, 2005.

[4] T. R. Gruber, "Towards Principles for the Design of Ontologies used for Knowledge Sharing", *International Journal of Human-Computer Studies*, Vol.43, pp.907-928, 1995.

[5] F. Manola, and E. Miller, *RDF Primer*, W3C, http://www.w3.org/TR/ rdf-primer, 2004.

[6] L. Miller, and D. Brickley, Friend of a Friend project, http://www.foaf- project.org, 2000.

[7] C. Fellbaum, *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*, MIT press, 1998.

[8] I. H. Witten, and E. Frank, *Data Mining: Practical machine learning tools and Techniques* (2nd Edition), Morgan Kaufmann, 2005.

[9] D. Pelleg, and A. W. Moore, "X-means: Extending K-means with Efficient Estimation of the Number of Clusters", *In 17th International Conference on Machine Learning*, pp.727-734, 2000.

[10] G. Adomavicius, and A. Tuzhilin, "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions", *IEEE Transactions on Knowledge and Data Engineering*, Vol.17, No.6, June 2005.

[11] K. Aberer and et al., "Emergent Semantics Principles and Issues", *Proceedings of Database Systems for Advanced Applications (DASFAA2004), LNCS 2973*, pp.25-38, 2004.

[12] P. Mika, "Ontologies Are Us: A Unified Model of Social Networks and Semantics", *Proceedings of the 4th International Semantic Web Conference (ISWC2005), LNCS 3729*, pp.522-536, 2005.

[13] T. R. Gruber, "Ontology of Folksonomy: A Mash-up of Apples and Oranges," *International Journal of Semantic Web and Information Systems*, Vol.3, No.1, pp.1-11, 2007.

[14] X. Xu, L. Zhang, and Y. Yu, "Exploring Social Annotations for the Semantic Web", *Proceedings of the 15th international conference on World Wide Web (WWW2006)*, New York, USA, pp.417-426, 2006.

[15] T. Knerr, "Tagging Ontology - Towards a Common Ontology for Folksonomies", http://code.google.com/p/tagont, 2006.

[16] L. Specia, and E. Motta, "Integrating Folksonomies with the Semantic Web", *Proceedings of the 4th European Semantic Web Conference (ESWC2007)*, Innsbruck, Austria, 2007.

저 자 소 개

강신재(Sin-Jae Kang)

1995년 : 경북대학교 컴퓨터공학과 공학사

1997년 : 포항공과대학교 컴퓨터공학과 공학석사

2002년 : 포항공과대학교 컴퓨터공학과 공학박사

1997년~1998년 : SK Telecom 정보기술연구원 주임연구원

2007년 : 오스트리아 U. of Innsbruck, DERI 연구소 방문교수

2002년~현재 : 대구대학교 컴퓨터·IT공학부 부교수

관심분야 : 시맨틱 웹, 소셜 웹, 온톨로지, 자연어처리

Phone : 053-850-6584

E-mail : sjkang@daegu.ac.kr