태그 온톨로지와 기계학습을 이용한 추천시스템†

(Recommendation System based on Tag Ontology and Machine Learning)

강 신 재*, Ying Ding**
(Sin-Jae Kang, Ying Ding)

요 약 소셜웹은 정보를 공유하고 사용자간 연결 정도를 높이기 위해 현재의 웹을 소셜 플랫폼으로 변화시키고 있다. 본 논문은 여러 소셜웹 사이트에 산재되어 있는 소셜 데이터를 중재하고 연결하는 방법을 제공하기 위해 딜리셔스, 플리커, 유튜브와 같은 대표적인 소셜 태깅 사이트의 태깅 데이터를 분석한다. 그 결과로 서로 다른 태깅 데이터를 통합하고, 서로 다른 소셜 메타데이터를 연결하기 위한 태그 온톨로지를 제안한다. 또한 태깅 데이터의 기계 학습을 통하여 유사 태그 그룹과 사용자 그룹 정보를 획득한 후, 태그 온톨로지를 학습한다. 이의 활용 방안으로는 학습된 태그 온톨로지를 이용하여 모델링한 추천 시스템도 제안한다.

핵심주제어 : 소셜 태깅, 온톨로지, 사용자 프로파일, 추천시스템

Abstract Social Web is turning current Web into social platform for knowing people and sharing information. This paper takes major social tagging systems as examples, namely delicious, flickr and youtube, to analyze the social phenomena in the Social Web in order to identify the way of mediating and linking social data. A simple Tag Ontology (TO) is proposed to integrate different social tagging data and mediate and link with other related social metadata. 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 tag ontology is also suggested as an applying field.

Key Words : Social Tagging, Ontology, User Profile, Recommendation System

1. 서 론

웹 2.0 또는 소셜웹(Social Web)은 정보를 공유하고 사용자간 연결 정도를 높이기 위해 현재의 웹을 소셜 플랫폼으로 변화시킨다. 소셜웹은 [1]에서 웹의 사회적 매개체로써의 기능을 강조하며 처음 제시된 용어인데, 소셜웹 이전의 웹이 일방적인 정보 제공의 형태였다면 소셜웹은 사용자들의 자발적인 참여와 개방성을 통해 블로그 등을 활용하여 정보 및 네트워크를 창조하고 공유하는 특성을 지닌다. 웹 2.0 환경 하에서는 누구라도 정보를 생성하고 웹상에 발행하는 것을 손쉽게 할 수 있다. 이러한 환경은 대다수의 일반 사용자로 하여금 웹 환경에 자발적으로 동참하게 하는 만드는 이유가 되고 있다. 기존의 웹에 정보를 더하는 가장 일반적인 방법 중 하나는 태깅(tagging)이다. 웹 2.0 상에서의 태깅이란 웹 상에 존재하는 객체(하이퍼링크, 사진, 동영상 등)에 키워드를 부여하는 행위로 정의할 수 있는데, 태그와 같이 사용자가 생성한 메타데이터를 통해 웹에 존재하는 정보들이 구조

† 이 논문은 2006년 정부(교육인적자원부)의 재원으로 한국학술진흥재단의 지원을 받아 수행된 연구임(KRF-2006-611-D00030)
* 대구대학교 컴퓨터·IT공학부
** School of Library and Information Science, Indiana University

화되어 지속적으로 성장하고 있는 것이다. 이러한 추세는 시맨틱 웹 커뮤니티에서 정의하여 사용하는 메타데이터들과 일맥상통한다. 예로는 FOAF (친구관계를 위한 메타데이터), SKOS(분류체계를 위한 메타데이터), DOAP(프로젝트를 위한 메타데이터), RSS (뉴스나 블로그를 위한 메타데이터), SIOC(소셜 네트워크를 위한 메타데이터), Dublin Core(문서를 위한 메타데이터) 등이 있다.

반면에 사람에 의한 것이 아니라, 기계에 의해 메타데이터를 자동으로 생성하는 연구도 활발하다. 이는 미리 정의한 온톨로지를 기반으로 기존 데이터를 분석해서 메타데이터를 추가하는 방식이다. Tim Berners-Lee의 오픈데이터 연결 제안[1]에 따르면, 연결된 데이터는 시맨틱 웹으로 가기 위한 필수 과정이며, 이는 서로 다른 개념이나 인스턴스를 연결하기 위해 owl:sameAs 또는 foaf:knows를 사용하는 것에서부터 쉽게 시작할 수 있다고 주장한다. 이를 통해 의미적으로 연결된 데이터가 점점 증가되어 현재의 웹은 모든 정보들이 의미적으로 상호 연결된 미래의 웹이 될 것이라는 생각이다.

폭소노미는 대중(folk)에 의한 분류(taxonomy)를 의미하며, 각 사용자가 관심 있는 웹 자원에 대해 자유 형태의 키워드인 태그(tag)를 자발적으로 부여하고 이를 대중이 함께 공유하는 형태를 가진다. 본 연구에서는 폭소노미에 내재되어 있는 사용자 선호 정보를 마이닝(mining)하기 위해, 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 태그 온톨로지의 학습에 사용하는 방법을 제시하고 있다. 이러한 결과물은 검색 시 질의어(태그) 확장과 태깅 시 연관 태그의 추천, 추천 시스템의 모델링 등에 활용될 수 있다. 태그 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.

본 연구에서는 태깅의 본질적인 속성을 분석하여 태그 온톨로지를 정의하고, 딜리셔스 사이트로부터 폭소노미 정보를 자동으로 추출한 후, 기계학습을 통하여 유사 태그와 사용자 그룹 정보를 획득하여, 태그 온톨로지의 학습에 사용한다. 이의 활용 방안으로 학습된 태그 온톨로지를 이용하여 모델링한 추천 시스템을 제안한다.

2장에서는 관련된 연구를 소개하고, 3장에서는 소셜 태깅 데이터의 모델링 및 수집 과정을 제시하고, 4장에서는 방대한 소셜 태깅 데이터로부터 태그 온톨로지를 학습하는 방법에 대해 설명한다. 5장에서는 태그 온톨로지의 활용 방안으로 추천 시스템을 모델링하고, 6장에서는 결론과 향후 연구 계획을 제시한다.

2. 관련 연구

다수의 사용자에 의해 생성된 대량의 태깅 데이터, 즉 폭소노미에 내재되어 있는 의미 관계를 이끌어내는 것과 관련하여 "떠오르는 시맨틱(emergent semantics)"[2]이라는 용어가 최근 많이 사용되고 있다. 이와 관련된 연구로, Tom Gruber가 태깅 데이터를 모델링하기 위해 온톨로지가 필요하다는 아이디어를 처음 제안하였다[3]. 그가 제안한 태그 온톨로지에는 object, tag, tagger, source, +, - 등이 포함되어 있으며, 협력적인 필터링을 위해 vote 개념도 포함하였다. 본 연구에서 제안하는 태그 온톨로지는 위 온톨로지에 date, source, comment와 같은 추가의 개념과 유사 태그, 유사 사용자 관계를 표현하기 위한 내용을 더하였다. 게다가 기존 소셜 메타데이터들과의 통합을 고려하여 모델링 되었다.

SCOT(Social Semantic Cloud of Tags)[2] 온톨로지는 태그 집합의 의미와 구조를 표현하고, 태그에 기반한 사용자들의 소셜 네트워크를 의미적으로 표현한다. SCOT는 태그와 태그 클라우드(tag cloud)를 구분하며, 리소스에 유일한 태그를 부여하기 위해 URI 메카니즘을 사용하고, 기존 SIOC, FOAF, SKOS에 기반하여 SCOT 온톨로지를 정의한다.

1) http://www.w3.org/DesignIssues/LinkedData.html

2) http://scot-project.org/

Mika[4]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.

Wu[5]는 사용자가 태깅한 데이터로부터 의미관계 정보를 추출하기 위해 "사용자(user), 자원(resource), 태그(tag), 시간(time)"과 같은 네 개의 구성요소로 이루어진 쌍을 기본으로 사용하였으나, 구체적인 온톨로지를 제시하지는 않았다.

Knerr[6]는 시맨틱 웹 기술의 하나인 FOAF[12]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[7]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.

3. 소셜 태깅(Social Tagging)

3.1 소셜 태깅 데이터의 모델링

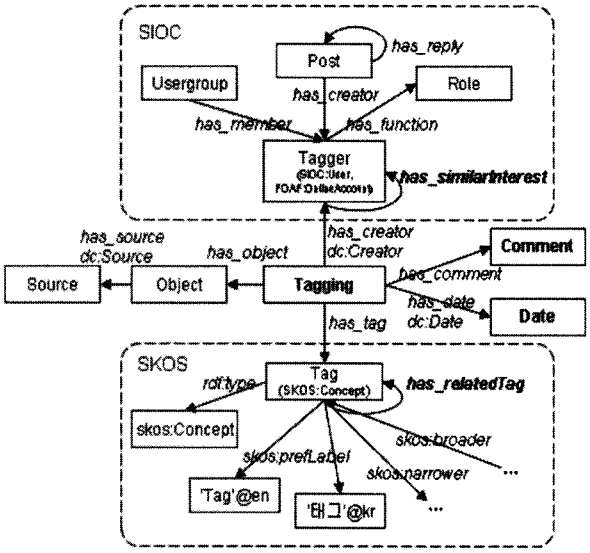
태그(tag)는 웹상의 객체를 분류하기 위해 사용되는 키워드이다. 태깅의 목적은 지속적인 태깅을 통해 검색과 공유가 쉽게 이루어질 수 있는 정보를 만드는 데 있다. 소셜 태깅은 단순한 태깅이 아니라 대중에 의해 생성되는 소셜 메타데이터를 의미하며, 그 결과물은 상향식(bottom-up)으로 폭소노미(folksonomy)라고 불리는 대중에 의한 분류체계를 구축하게 된다.

태깅 서비스를 제공하는 많은 사이트가 있지만, 본 연구에서는 소셜 태깅 데이터를 분석하기 위해 del.icio.us(북마크)와 flickr(사진), youtube(동영상) 사이트를 선택하였다. 분석을 통해 Tom Gruber[3]에 의해 제안된 태그 온톨로지를 개선한 새로운 태그 온톨로지를 제안한다. [3]에서는 object, tag, tagger, source, vote를 온톨로지의 주요 개념으로 제시하였는데, 본 연구에서는 tagging, comment, data 개념을 추가로 제시하였다. 태깅은 tag와 tagger, object가 동시에 관련을 맺는 행위인데, 이를 동시에 묶어줄 수 있는 tagging과 같은 개념이 있어야 상호 연관 정보를 정확히 표현할 수 있고, 추후 태깅 정보의 검색이 용이해질 수 있다. 또한 대부분의 소셜 사이트에서는 객체나 태그에 대해 주석(comment)을 다는 기능을 제공한다. 이는 객체나 태그에 대한 추가의 정보를 제공하는 역할을 해 추후 의미 분석 과정에서 도움이 되므로 comment 개념을 추가하였다. Date 개념은 태깅 작업을 시간적인 측면에서 기술하기 때문에 태깅의 변동 과정 등에서 정보를 추출할 때 도움이 될 수 있다. 이 외에도 태그 온톨로지에 "has_relatedTag"와 "has_ similarInterest" 관계를 추가하여 개념간 의미 정보를 추가로 표현하였다.

TO를 태그 온톨로지라 하면 다음과 같이 표현할 수 있다.

$$TO = (C, \mathcal{R}) \qquad (1)$$

$C = \{c_i, i \in N\}$은 유한한 개념집합이며, $\mathcal{R} = \{(c_i, c_k), i, k \in N\}$은 개념집합에 속한 개념간에 존재하는 관계의 집합이다. 본 연구에서 정의한 태그 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.



(그림 1) 태그 온톨로지

그림 1에 정의된 태그 온톨로지는 FOAF, SIOC, SKOS에서 사용하고 있는 개념 스키마의 클래스들과의 연관도에 따라, 관련 개념을 매핑(mapping)하여 정의한 것으로 이를 통해 기존에 FOAF, SIOC, SKOS를 사용하는 다른 사이트들과 손쉽게 정보를 공유할 수 있게 된다. SCOT 온톨로지와 다른 점은 태그 클라우드를 따로 정의하지 않고 있다는 점이다.

정의한 태그 온톨로지에서 핵심이 되는 8개의 개념과 9개의 관계를 표 1과 표 2에 나타내었다.

<표 1> 태그 온톨로지의 주요 개념

Concept	Description	Value Type
Tagging	Tag, Tagger, Object 개념을 동시에 연결하기 위해 정의된 개념	문자열
Tag	사용자가 객체에 부여하는 키워드	문자열
Tagger	객체를 태깅하는 주체	문자열
Object	태깅의 대상이 되는 웹에 존재하는 객체(링크, 사진 등)	문자열
Source	객체가 존재하는 장소(사이트)	문자열
Comment	태거가 객체나 태그에 부여한 추가 기술정보	문자열
Date	태깅이 일어난 일시	날짜
Vote	해당 객체를 태깅한 서로 다른 태거의 수	정수

<표 2> 태그 온톨로지의 주요 관계

Relation	Domain	Range	Cardinality
has_tag	Tagging	Tag	N
has_relatedTag	Tag	Tag	N
has_creator	Tagging	Tagger	1
has_object	Tagging	Object	1
has_date	Tagging	Date	1
has_source	Object	Source	N
has_comment	Tagging	Comment	N
has_vote	Tagging	Vote	N
has_similarInterest	Tagger	Tagger	N

제시한 태그 온톨로지는 소셜 태깅 행위 자체를 구조적으로 나타내기 위해 정의한 것으로, 대중이 각각 부여한 태그의 의미에 의해 분류되어진 폭소노미와는 차이가 있다. 서로 다른 태깅 데이터를 통합하고, 기존의 소셜 메타데이터를 연결하기 위한 구조를 모델링하는 것이 목적이다.

3.2 소셜 태깅 데이터의 수집

del.icio.us, flickr, youtube 등 주요 소셜 태깅 시스템의 데이터를 수집하기 위해 ST 크롤러(Social Tagging Crawler)를 구현[8]하였는데, 이것은 오픈소스인 웹 크롤러[3]와 정의된 태그 온톨로지 TO를 기반으로 구현되었다. 수집된 정보는 RDF[4] 트리플로 표현되어 Jena[5]를 이용하여 저장되었다. 이는 W3C에 의해 표준화되고 있는 시맨틱웹 기반 기술을 적극 채택하여 OWL로의 확장 및 추론 등의 작업을 수행할 수 있는 토대를 마련하기 위함이다. 수집된 태깅 데이터는 del.icio.us 1.6GB, flick 233MB, youtube 234MB로 총 2.1GB에 이른다. RDF로 표현된 태깅 데이터의 예는 그림 2와 같다.

3) https://crawler.dev.java.net/
4) http://www.w3.org/TR/2004/REC-rdf-concepts-200402 10/
5) http://jena.sourceforge.net/

```
<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/54e1669d-a9aa-4f9c-9288-504045fb217a">
    <j.0:has_object rdf:resource=
"http://www.samspublishing.com/articles/article.asp?p=101373&amp;amp;rl=1"/>
    <j.0:has_date>Feb 07</j.0:has_date>
    <j.0:has_tagger>daveimm</j.0:has_tagger>
    <j.0:has_vote>5</j.0:has_vote>
    <j.0:has_comment></j.0:has_comment>
    <j.0:has_tag rdf:resource="http://del.icio.us/tag/oop_c#_online_tutorial"/>
  </rdf:Description>
    <rdf:Descripton rdf:about=
"http://uto.deri.at/66378c3d-0daf-45c6-bf7d-664f344b1caa">
    <j.0:has_vote>47</j.0:has_vote>
    <j.0:has_object rdf:resource=
"http://www.technologyreview.com/Infotech/18796/"/>
    <j.0:has_tag rdf:resource="http://del.icio.us/tag/ai"/>
    <j.0:has_tagger>crutcher</j.0:has_tagger>
    <j.0:has_tag rdf:resource="http://del.icio.us/tag/cognitive-science"/>
    <j.0:has_tag rdf:resource="http://del.icio.us/tag/pattern-recognition"/>
    <j.0:has_tag rdf:resource="http://del.icio.us/tag/computers"/>
    <j.0:has_tag rdf:resource="http://del.icio.us/tag/cognitive"/>
```

(그림 2) 태그 온톨로지 인스턴스

실제로 ST 크롤러를 통해 del.icio.us 사이트로부터 462,733명의 사용자(태거), 404,388개의 태그, 483,564개의 북마크(객체)가 포함된 총 9,400,029개의 태깅 인스턴스를 추출하였다. 출현빈도 상위 10위의 태그와 북마크는 표 3과 같다.

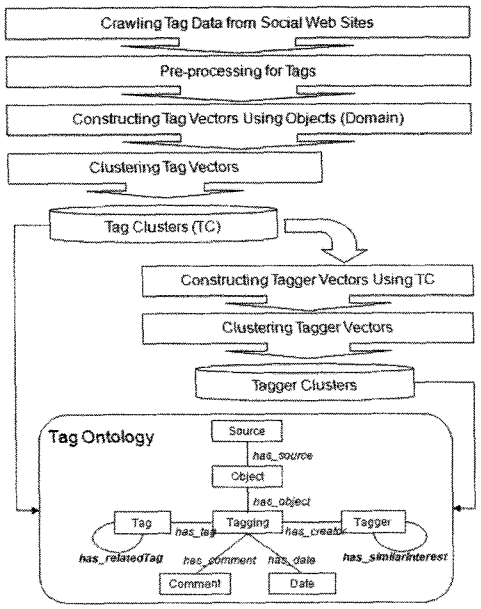
4. 온톨로지 학습

태그 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 의미 관계는 직접 얻을 수가 없다. 따라서 추출된 태깅 정보를 가공하여 기계 학습을 거치면서 해당 의미 관계를 추출하고자 한다.

<표 3> 출현빈도 상위 10위까지의 태그와 북마크

순위	태그	태그 빈도수	북마크	북마크 빈도수
1	blog	141,871	en.wikipedia.org	26,745
2	system	120,673	www.youtube.com	14,990
3	design	109,249	community.livejournal.com	6,594
4	software	87,719	www.google.com	6,376
5	programming	83,665	www.w3.org	6,193
6	tool	83,461	news.bbc.co.uk	5,718
7	reference	74,602	www.flickr.com	5,645
8	web	70,538	java.sun.com	5,538
9	video	65,226	www.nytimes.com	5,222
10	music	61,246	www.microsoft.com	5,219

전반적인 절차는 그림 3에 나타나 있으며, 4.1절에서는 태그 클러스터를 통하여 "has_relatedTag" 관계를 추출하는 과정을, 4.2절에서는 태거 클러스터를 통하여 "has_similarInterest" 관계를 추출하는 과정에 대해 자세히 설명한다.



(그림 3) 태그 온톨로지 학습과정

4.1 태그 클러스터(Tag Cluster)

소셜웹 사이트에서 태깅에 사용된 다양한 태그들을 수작업으로 분류하는 것은 일관성, 비용, 시간 등의 여러 문제로 인해 실용적이지 못하므로, 태그 온톨로지 인스턴스(리소스에 대한 태깅 정보)를 가공하여 자동으로 태그를 분류하고자 한다. 태그를 분류하는데 사용될 수 있는 정보를 태그 온톨로지에서 살펴보면, 하나의 태깅에 관련된 정보로 tagger, object, date, comment 등 여러 가지가 있으나, object의 종류와 내용에 따라 tagger가 해당 object에 tag를 부여하는 것이기 때문에, object가 tag의 특성(쓰임새)을 가장 잘 나타내주는 정보라고 볼 수 있다. 그래서 본 논문에서는 object와 tag의 공기정보(collocation)를 이용하여 태그 벡터를 구성하였다.

태그 클러스터를 생성하기 위해 사용될 태그 벡터를 구성하기에 앞서서, 태그에 대한 전처리를 한다. 이는 다양한 형태의 태그를 정규화(일반화)하기 위함인데, 일반적이지 않은 태그는 제거하고, 형태론적으로 유사한 태그들을 정리하여 정규화된 태그를 선택하는 과정이다. 그 다음 태그가 워드넷(WordNet)[9]에 존재하는지 확인하여 존재하면 후보 태그로 선택한다. 워드넷에 존재하는 태그를 선택하는 이유는 워드넷에 등록된 단어가 일반적으로 자주 사용되고 대표성이 있는 단어로 간주될 수 있고, 또한 워드넷을 이용한 단어 간 유사도 계산이 가능해지기 때문에 추후 클러스터 내 태그의 랭킹이나 태그 클러스터의 랭킹이 가능해지기 때문이다. 이와 같은 과정을 통해 총 26,691개의 태그가 선택되었다.

크롤된 북마크(object)를 정규화하기 위해서는 도메인 서버 이름만 추출하여 사용하였는데, 총 237,273개의 object가 선택되었다.

<표 4> 전처리 후 벡터생성을 위해 선택된 tag & object의 수

	Tagger	Tag	Object
Raw data	462,733	404,388	483,564
Preprocessing	Removing Symbol & Lowering	374,279	-
	Stemming & Checking WordNet	26,691	-
	Domain Server Name Only	-	237,273

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

태그 벡터를 클러스터링하기 위해서는 Witten [10]이 개발한 WEKA (Waikato Environment for Knowledge Analysis) 패키지의 여러 클러스터링 알고리즘을 적용시켜 보았으며, 이 가운데 가장 좋은 성능을 보인 X-means를 선택하여 실험 결과를 얻었다. X-means 알고리즘은 Pelleg과 Moore[11]가 개발하였는데, 기존 K-means 알고리즘의 세 가지 주요한 단점, 즉 느리고 확장이 쉽지 않고, 클러스터의 수 K를 사용자가 정해야 되며, 국소해(local minima)에 빠지기 쉽다는 단점들을 개선한 클러스터링 알고리즘이다. 서버에서 실험이 가능한 크기인 4,676개의 tag와 3,616개의 object를 대상으로 태그 벡터를 구성하였고, 총 98개의 태그 클러스터를 얻을 수 있었다. 하나의 태그 클러스터에 속한 태그들은 유사한 종류의 object를 태깅할 때 같이 사용되는 경우가 많았다는 것을 의미하므로 상호간 "has_relatedTag" 관계를 갖는 것으로 간주할 수 있다.

<표 5> 태그 클러스터 예시

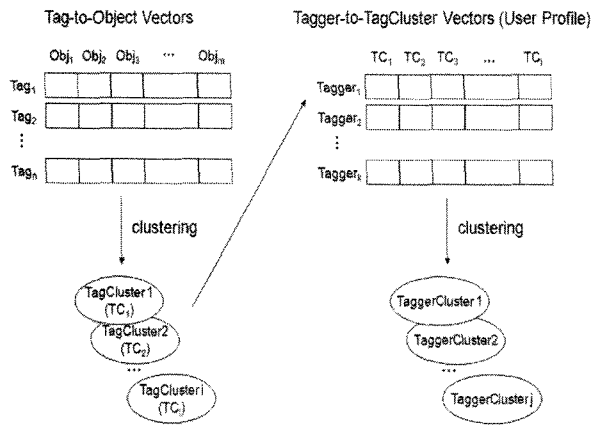
클러스터	태 그
1	ajax, c, code, development, html, java, library, net, python, rails, rudy
2	dictionary, English, language, literature, writing
3	comic, entertainment, film, forum, japan, Japanese, movie, radio, streaming, television, tv
4	calculator, conversion, convert, converter, currency, euro, exchange
5	account, bank, banking, bill, consumer, credit, deal, doctor, financial, healthcare, insurance, loan, medical, medicare, medicine, savings
6	air, apartment, building, cleaning, do, fire, guide, house, housing, move, rental, safety, studio
7	Black, blue, brown, fairy, flower, gratis, leather, line, neo, pink, red, skull, stripes, style, Sweden, Swedish, vintage, white, yellow
8	culture, history, philosophy, politics, religion
9	astronomy, earth, geography, german, map, nasa, space, world
10	font, illustration, inspiration, portfolio, typography

4.2 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. 이 또한 모든 태그를 사용하여 태거 벡터를 구성하기에는 벡터의 차원이 너무 커지기 때문에 현실적으로 기계학습이 어려운 문제점이 있다. 따라서 벡터의 차원을 줄이는 한 방법으로 4.1절에서 획득한 태그 클러스터를 이용하여

태거 벡터를 구성하였다(그림 4).

태거 벡터의 클러스터링을 위해서는 4.1절과 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 군집화한 것으로 태그 온톨로지의 "has_similarInterest" 관계 정보를 담고 있는 것으로 해석할 수 있다.



(그림 4) 태그 및 태거 벡터 구성

사용자 프로파일과 유사 사용자 그룹 정보는 추천 시스템을 모델링하는데 있어 아주 중요한 역할을 한다.

5. 추천시스템 모델링

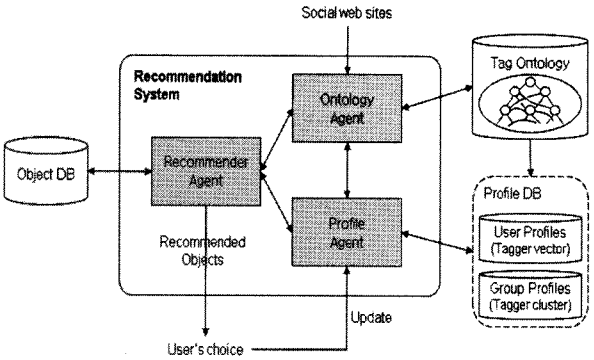
태그 온톨로지를 활용하기 위한 한 분야로 추천 시스템을 모델링하고자 한다. 이는 태그 클러스터를 이용하여 구성된 각각의 태거 벡터는 사용자의 성향(선호정보)을 나타내는 사용자 프로파일의 역할을 할 수 있으며, 태거 클러스터는 유사한 성향을 가지는 사용자 그룹의 정보를 표현하는 그룹 프로파일의 역할을 할 수 있기 때문이다.

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: 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.

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

추천이 어떻게 만들어지느냐에 따라 내용 기반(content-based), 협력 기반(collaborative), 하이브리드(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[12].

본 논문에서는 최적의 추천 정보를 생성하기 위하여 태그 온톨로지로부터 사용자 프로파일과 그룹 프로파일을 추출하여 이용하는 하이브리드 접근법을 제안한다. 추천 시스템의 전체적인 구성은 그림 5에 나타나있으며, 크게 추천 에이전트, 온톨로지 관리 에이전트, 프로파일 관리 에이전트로 나뉜다. 본 논문의 3장과 4장에서 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 DB에서 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 5) 태그 온톨로지를 이용한 추천 시스템

6. 결론 및 향후 연구과제

현재의 웹은 정보, 지식, 사람들을 연결하기 위한 많은 다양한 시도를 하고 있다. 이를 통하여 차세대 웹으로 변환하려는 시도 중 하나가 시맨틱 웹(semantic web)이다. 이것은 기계가 이해할 수 있는 형태로 데이터를 표현함으로써 웹 서비스 간 검색/조정(mediation)/실행이 자동으로 이루어지고 이를 기반으로 지능적인 웹 서비스가 개발/지원될 수 있는 환경을 말한다. 소셜웹은 시맨틱 웹으로 가기 위한 과정으로 설명할 수 있다.

본 논문에서는 소셜웹 사이트에 존재하는 방대한 태깅 정보를 가공하기 위하여 태그 온톨로지를 정의하고, 태깅 정보를 자동으로 추출한 후, 클러스터링 알고리즘을 적용하여 온톨로지를 학습하는 방법론을 제시하였다. 태그 간, 태거 간에 존재하는 연관 관계를 자동으로 추출하였기 때문에 수작업을 배제한 실용적인 방법론이며, 또한 방대한 양의 정보를 사용하여 보다 일반적이고 객관적인 정보를 추출했다고 볼 수 있다.

제시된 태그 온톨로지는 서로 다른 태깅 데이터 간의 통합과 소셜 메타데이터를 중재하는 역할을 하여, 태그를 사용하는 서로 다른 웹 사이트 간 정보의 교환과 처리를 원활히 해주게 된다. 정보검색 시스템에서 태그 온톨로지의 태그 클러스터를 이용하면 질의 확장 등을 통하여 보다 정확한 정보검색 시스템의 구현이 가능하게 된다. 또한 사용자 프로파일과 그룹 프로파일을 태그 온톨로지로부터 생성할 수 있기 때문에 이를 활용한 추천 시스템의 모델링이 쉬워지는 장점이 있다.

향후에는 Google social graph API[6] 등을 활용하여 웹상에 산재되어 있는 개인의 프로파일들을 자동으로 통합하고, URI를 이용하여 프로파일을 표현하는 방법에 대해 연구하고자 한다.

참 고 문 헌

[1] P. Hoschka, "CSCW research at GMD-FIT: From basic groupware to the Social Web", *ACM SIGGROUP Bulletin*, vol.19, no.2, pp.5-9, 1998.

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

[3] 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.

[4] 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.

[5] 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.

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

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

[8] M. Fried, Social Tagging Wrapper. Bachelor Thesis, Institute of Computer Sciences,

6) http://code.google.com/apis/socialgraph/

University of Innsbruck, Austria, 2007.
[9] C. Fellbaum, WordNet: An Electronic Lexical Database (Language, Speech, and Communication), MIT press, 1998.
[10] I. H. Witten, and E. Frank, Data Mining: Practical machine learning tools and Techniques (2nd Edition), Morgan Kaufmann, 2005.
[11] 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.
[12] 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.

강 신 재 (Sin-Jae Kang)

- 종신회원
- 1995년 : 경북대학교 컴퓨터공학과 (공학사)
- 1997년 : 포항공과대학교(POSTECH) 컴퓨터공학과 (공학석사)
- 2002년 : 포항공과대학교(POSTECH) 컴퓨터공학과 (공학박사)
- 1997년 ~ 1998년 : SK Telecom 정보기술연구원 연구원
- 2007년 : 오스트리아 University of Innsbruck, DERI 연구소 방문교수
- 2002년 ~ 현재 : 대구대학교 컴퓨터·IT공학부 부교수
- 관심분야 : Semantic Web, Social Web, Recommender System, Ontology, Natural Language Processing

Ying Ding

- 비회원
- 1993년 : Information Science, Xi'an Electronic Science & Technology University, Xi'an, China (B.Eng.)
- 1996년 : Information Science, The Graduate School of Chinese Academy of Sciences, Beijing, China (M.Sc.)
- 2000년 : Information Science, School of Computer Engineering, Nanyang Technological University, Singapore (Ph.D.)
- 2000년 ~ 2003년 : Senior Researcher, Business Informatics Group, Department of Mathematics & Computer Science, Free University, Amsterdam, Netherlands
- 2003년 ~ 2007년 : Senior Researcher, Digital Enterprise Research Institute, Department of Computer Science, University of Innsbruck, Austria
- 2008년 ~ current : Assistant Professor, School of Library and Information Science, Indiana University
- Interests : Semantic Web, Social Networks Analysis, Webometrics, Knowledge Engineering, and Information Retrieval