WordNet어휘계층구조 기반의 태그/사용자 분류체계 구축지원도구의 개발

황석형*, 최성희*, 김한수*, 김정래**
*선문대학교 컴퓨터공학과 **선문대학교 영어학과
{shwang, shchoi}@sunmoon.ac.kr, {hi35sojin, mydream2023}@gmail.com

A Development of Tag/User Classification System Based on WordNet Hierarchies

Suk-Hyung Hwang*, Sung-Hee Choi*, Han-Soo Kim*, Jeong-Rae Kim**
*Dept. of Computer Science and Engineering, **Dept. of English
Sun Moon University

요 약

오늘날 인터넷의 발달과 더불어 스마트기기의 보급이 급성장하면서, 다양한 웹사이트에서 데이터가 기하급수적으로 발생되고 있고, 수 많은 다종다양한 데이터를 효율적으로 저장/관리/분석하기 위한 유용한 어노테이션(Anotation) 기법으로서, 리소스에 대한 사용자의 태깅(Tagging)기능이 널리 활용되고 있다. 본 연구에서는, 사용자들의 공통 태그 데이터를 수집하여, WordNet을 기반으로 다양한 수준의 태그/사용자 분류체계를 구축하기 위한 지원도구개발에 관한 연구결과를 보고한다.

1. 서론

오늘날 여러 종류의 대규모 웹사이트와 SNS 등이 기하급수적으로 성장하여, 다종다양한 데이터가 생산되어 이용되고 있다. 다양한 데이터를 효율적으로 저장/관리/분석하기 위한 유용한 어노테이션 기법으로서 리소스에 대한 사용자의 태깅기능이 널리 활용되고 있으며, 웹 상에 내재되어 있는 태그데이터(Tagged Data)를 효율적으로 관리/분석/응용하기 위한 다양한 기법들이 등장하고 있다[1]. 따라서, 인터넷에 대량으로 생성되어있는 태그데이터를 적절하게 수집, 분석하여 다양한 분야(정보검색, 시멘틱웹 등)에 응용하려는 웹 데이터 마이닝(Web Data Mining) 분야의 연구가 최근 큰 주목을 받고 있다[2,3].

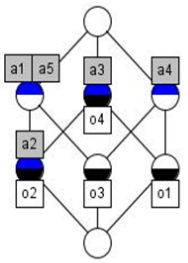
본 연구에서는, 북마크와 관련 정보를 공유하고 태깅 기능을 제공하는 Bibsonomy(Bibsonomy.org)사이트 사용자들의 태그데이터를 수집하고, WordNet[4]과 형식개념분석기법(Formal Concept Analysis)[5]을 기반으로 분석하여 다양한 수준의 태그/사용자 분류체계를 구축하기 위한 지원도구(TagLighter)를 개발하였다.

워드넷(http://wordnet.princeton.edu)[4]은 인간의 어휘지식에 대한 심리언어학적 연구성과를 토대로 1985년부터 프린스턴대학 인지과학연구실이 구축해온 영어어휘 온라인데이터베이스로서, 단순한 단어의 자모순이 아닌 단어의 의미에 의해서 어휘정보를 조직화함으로써 자연언어처리와 정보검색 분야에서 널리 이용되고 있다. 워드넷은 영어단어에 대한 의미뿐만아니라 영어단어들간의 관계정보를 망라하고 있는 사전으로서, 영어단어를 동의어 관계에 있는 어휘들의 집합인 synset으로 묶고 synset간의 내부적 연결을 통해서 구축된 의미연결체계를 제공한다. 워드넷의 synset에 포함된 연결관계에는 synset의 각 요소들에 대한 정의와 더불어서, 동의(synonymy), 반의(antonymy), 상의(hypernymy), 하의(hyponymy), 분의(meronymy), 함의(entailment) 등과 같은 의미관계를 표현한 언어학적인 계층구조정보를 포함하고 있다.

형식개념분석기법[5]은 개념격자라는 수학적 모델을 기반으로 하는 데이터분석기법의 일종으로서, 개념적인 데이터분석과 지식기반처리분야의 제반문제들에 대한 수학적 해법을 제공하고 있다. 형식개념분석기법에서는, 분석대상 데이터를 객체와 속성, 그리고 이들 사이의 포함관계를 이진형태로 나타낸 데이터테이블(formal context라고 부름)을 입력으로 하여, 공통속성을 갖는 객체들을 개념(formal concept)이라는 기본정보단위로 추출하고, 개념들 사이의 상하위관계를 파악하여 계층화함으로써 격자형태의 개념계층구조(Concept Lattice)를 구축한다(그림 1).

	a1	a2	a3	a4	a5
o1			×	×	
o2	×	×	×		×
o3	×			×	×
o4			×		

(a)데이터테이블



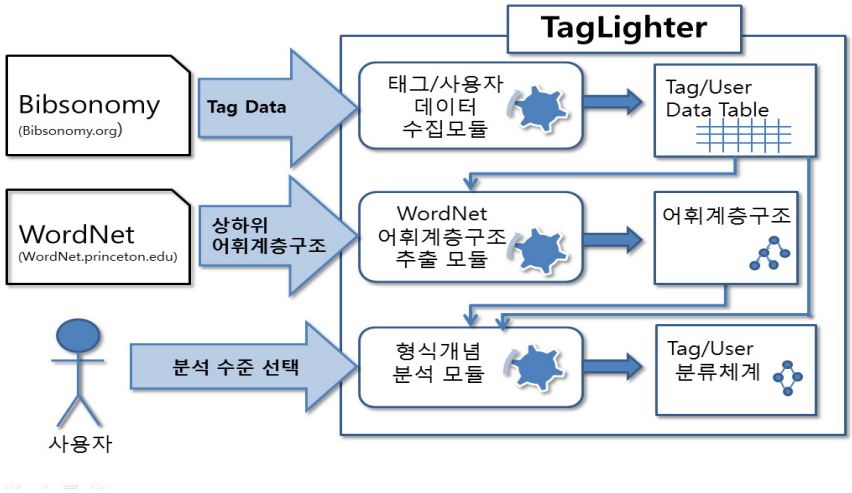
(b)개념계층구조

(그림 1) 데이터테이블과 개념계층구조

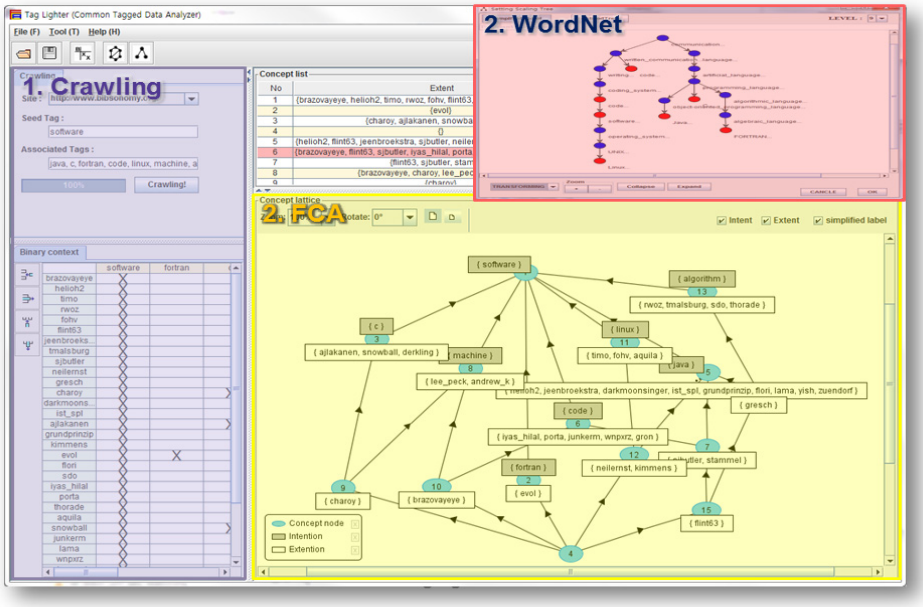
격자형태의 개념계층구조에서는, 각 개념들과 이들 사이의 상하위관계가 링크에 의해 표시되며, 특히, 개념들 간의 링크에 의해 만들어지는 경로에 의해 상위개념으로부터 하위개념으로 속성들이 상속되며, 하위개념으로부터 상위개념으로 해당 객체들이 전파된다. 그림1(b)의 예에서, o2 객체는 고유한 속성으로서 a2를 갖고 있으며, 더불어서, 상위개념들로부터 a1과 a5, 그리고 a3과 같은 속성을 상속받는다. 한편, a3라는 속성을 갖는 객체로서는 o1, o2, o4가 됨을 알 수 있다. 이와 같은 방법을 사용함으로써, 주어진 문제영역의 객체들과 이들이 갖는 속성들을 context형태로 파악하여, 개념을 추출하고 개념격자형태로 나타냄으로써, 도메인 내의 개념들을 분류하고 체계화 할 수 있는 계층적 개념구조를 수월하게 구축할 수 있다.

이상의 연구배경을 토대로 본 논문에서는 WordNet을 기반으로 하는 태그/사용자 분류체계 구축지원도구(TagLighter) 개발에 관하여 보고한다. 태그/사용자 분류체계를 기반으로, 다양한 배경지식과 상황에 따라서 사용자들이 공통적으로 선택한 태그(키워드, 핵심단어)그룹 및 공통태그를 사용하는 사용자그룹을 추출할 수 있다. 따라서, 이와 같은 태그/사용자 분류체계들을 이용하여 다양한 웹 데이터 분류체계가 구축될 수 있으며, 공통적인 태그 검출에 의한 사용자들의 공통관심사의 파악과 사용자 커뮤니티의 추출 및 구성 등에 활용할 수 있다.

이후, 본 논문의 구성은 다음과 같다. 제2장에서는 지원도구(TagLighter)의 시스템구성, 제3장에서는 적용사례, 그리고 제4장에서는 결론에 대해서 설명한다.



(그림 2) TagLighter의 모듈구성



(그림 3) TagLighter의 실행화면

2. 지원도구(TagLighter)의 구성요소

(1)태그데이터 수집모듈

태그데이터 수집모듈에서는 Open API를 제공하고 있는Bibsonomy.org사이트로부터 사용자가 관심있어하는 태그데이터를 수집하여 Tag/User Data Table를 구성한다. 따라서 Tag/User Data Table에는 어떤 User가 어떤 Tag를 사용하고 있는지에 관한 정보가 표현되어 있다.

(2)WordNet 어휘계층구조 추출 모듈

WordNet 어휘계층구조 추출 모듈에서는, Bibsonomy 사이트에서 수집된 Tag/User Data와 wordnet.princeton.edu 사이트(프린스턴대학에서 개발 및 운영하는 온라인 사전. 단어의 뜻과 단어들 사이의 상위어/하위어/등위어 등의 관계를 나타냄)를 기반으로, 사용자의 태그들에 관한 상하위 어휘계층 구조를 추출하여 트리형태의 계층구조를 구성한다.

(3)형식개념분석모듈

형식개념분석기법(Formal Concept Analysis)은, 격자이론(Lattice Theory)을 기반으로 하는 일종의 데이터분석기법으로서, 다양한 데이터들로부터 공통적인 속성을 갖는 데이터요소들을 개념(Concept)이라는 기본 정보단위로 추출하여 개념들 사이의 상하위 관계를 토대로 개념계층구조(Concept Hierarchy)를 구축하기 위한 이론적인 토대를 제공한다. 형식개념분석 모듈은 Tag/User Data와 어휘계층구조를 토대로, 사용자가 선택한 분석 수준에 따라서 형식개념 분석 기법을 적용하여 태그 유저분류체계를 구성한다.

3. 지원도구의 적용사례

TagLighter는 그림 3과 같이 Tag/User Data를 수집하고 그 결과를 보여주는 "①Crawling 실행"기능과 WordNet의 어휘계층정보를 토대로 사용자 태그들에 대한 상하위 어휘계층구조를 나타내고 적용선택을 할 수 있는 "②WordNet어휘계층구조 선택"기능, 그리고 형식개념분석 기법을 적용하여 태그/사용자 분류체계를 출력하는 "③FCA적용"기능으로 구성되어 있다.

이하, 본 연구에서 개발한 TagLighter의 적용사례를 설명한다.

	software	c	java	freeware	linux
helioh2	X		X		
timo	X				X
jeenbroekstra	X		X		
fohv	X				X
flint63	X		X	X	
sjbutler	X		X		
neilernst	X		X		X
gresch	X		X		
charoy	X	X			

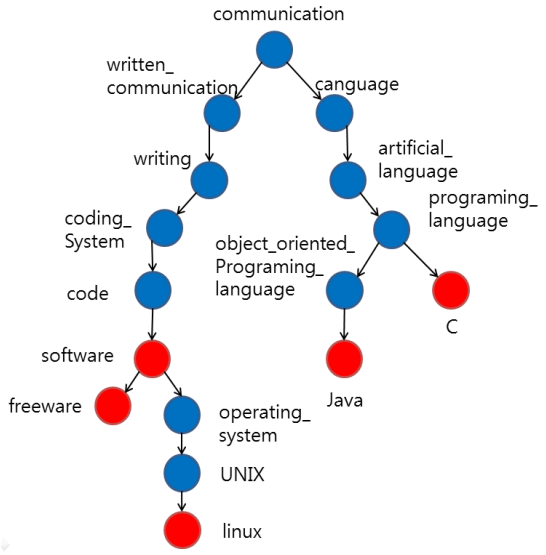
(그림 4) Crawling실행결과 수집된 Tag/User Data Table

(1)태그/사용자 데이터 수집

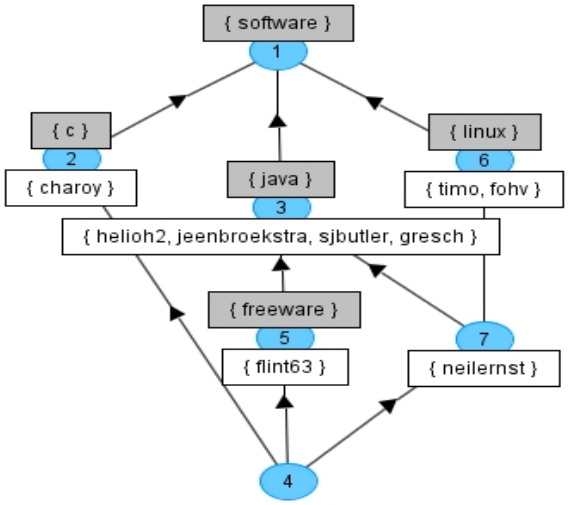
Bibsonomy.org에서 태그/사용자 데이터를 수집하기 위하여, 관심태그들(software, Java, C, linux, freeware)을 Crawling 실행영역에 입력하여 태그데이터 수집모듈에 의해 그림 4과 같이 5개 관심태그들을 사용한 9명의 사용자들에 관한 Tag/User Data Table을 입수할 수 있다.

(2)WordNet어휘계층구조 추출

수집된 태그/사용자 데이터(그림4)의 태그정보를 토대로 WordNet으로부터 다양한 수준의 어휘계층구조들을 추출한다(그림5) 예를 들면, 그림5의 어휘계층구조는 사용자가 입력한 태그들 중에서 Java의 상위 개념어휘로서 object_oriented_programming_language가 존재하고, Java와 C의 공통 상위개념어휘로서 programming_language가 추출되었다.



(그림 5) WordNet어휘계층구조



(그림 6) 공통태그/사용자 분류체계

(3)형식개념분석 적용

그림4의 Tag/User Data Table에 대하여 형식개념분석 기법을 적용하면, 그림 6와 같은 공통태그/사용자 분류체계가 구성된다. 예를 들면 software와 Java 태그를 공통적으로 사용하는 사용자는 총 4명 (helloh2, jeenbroekstra, sjbutler, gresch)임을 알 수 있다.

한편, 그림4의 Tag/User Data Table에 대하여 그림5의 WordNet어휘계층구조를 토대로 형식개념분석을 적용하면, 그림7과 같은 다양한 수준의 태그/사용자 분류체계들을 추출할 수 있다.

4. 결론

본 연구에서 개발한 TagLighter는, Bibsonomy사이트에서 제공하는 태그/사용자 데이터를 수집하고, WordNet을 토대로 어휘계층구조를 파악하여 형식개념분석처리를 수행함으로서 다양한 수준의 태그/사용자 분류체계를 구축하기 위한 지원도구이다. 다양한 배경지식과 상황에 따라서 사용자들이 적합한 수준의 태그/사용자 분류체계를 기반으로, 공통적으로 선택한 태그(키워드, 핵심단어)그룹 및 공통태그를 사용하는 사용자그룹을 추출할 수 있다. 따라서, 이와 같은 태그/사용자 분류체계들을 이용하여 다양한 수준의 웹 데이터 분류 체계가 구축될 수 있으며, 공통적인 태그 검출에 의한 사용자들의 공통관심사의 파악과 사용자 커뮤니티의 추출 등의 분야에 활용할 수 있다.

참고문헌

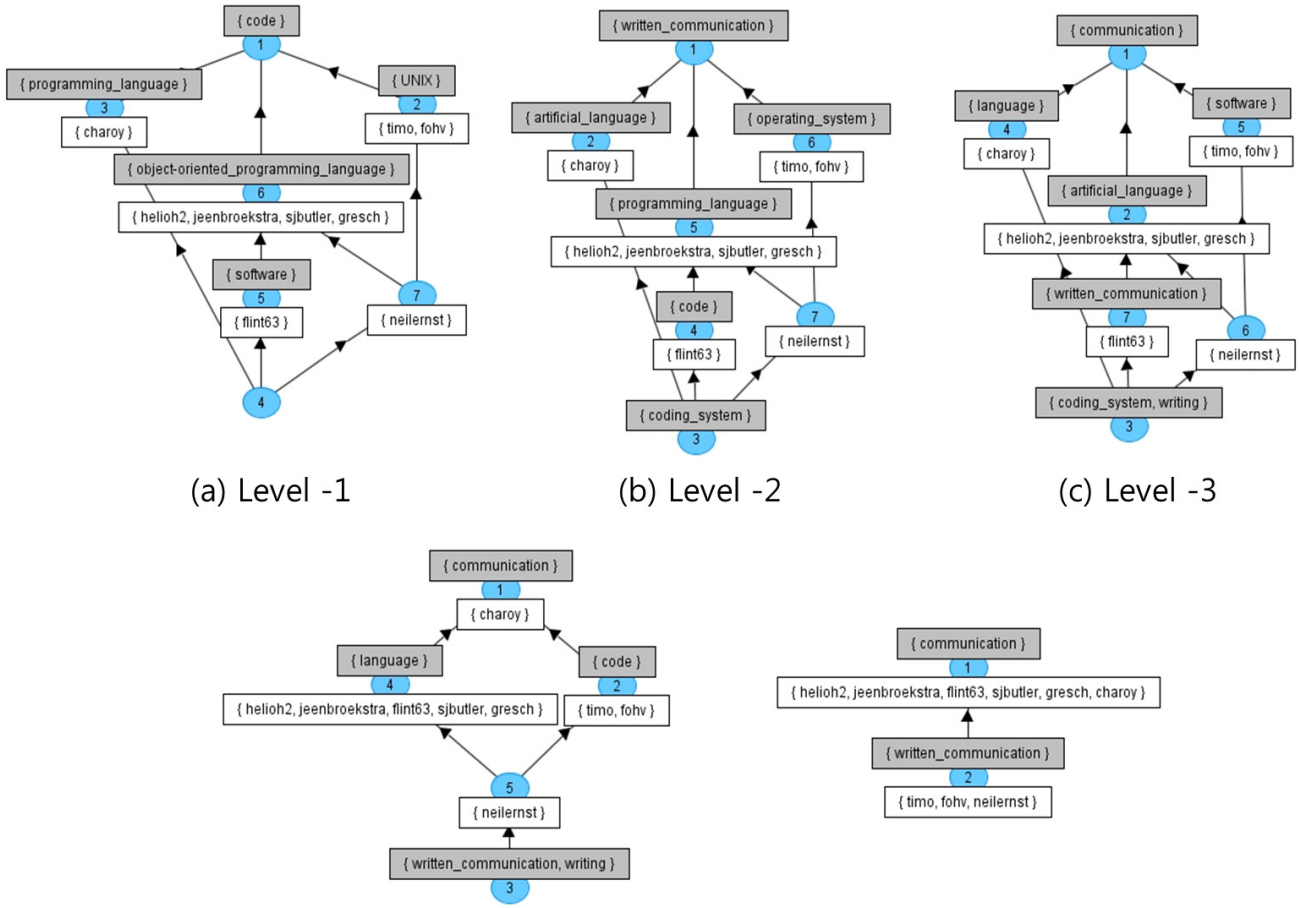
[1] Andreas Hotho, Robert Jäschke, Christoph Schmitz, and Gerd Stumme, "BibSonomy: A Social Bookmark and Publication Sharing System," Proceedings of the Conceptual Structures Tool Interoperability Workshop, pp.87~102, 2006.

[2] Bing Liu, Web Data Mining, Springer, 2010.

[3] 박희진, "폭소노미에 따른 웹 분류 연구:이용자 태깅행위분석을 중심으로," 한국문헌정보학회지, 제45권, 제1호, 2011.

[4] http://wordnet.princeton.edu/

[5] Ganter, Wille, Formal Concept Analysis: Mathematical Foundations, Springer-Verlag, 1999.



(그림 7) 다양한 수준의 태그/사용자 분류체계