그래프 구조를 이용한 카테고리 구조로부터 상하위 관계 추출

(Graph-based ISA/instanceOf Relation Extraction from Category Structure)

최 동 현 † 최 기 선 ††
(DongHyun Choi) (Key-Sun Choi)

요 약 상하위 관계 자동 추출은 분류체계를 자동 구축하는 데 있어서 핵심적인 내용이며, 이렇게 자동으로 구축된 분류 체계는 정보 추출과 같은 여러 가지 분야에 있어서 중요하게 사용된다. 본 논문에서는 카테고리 구조로부터 상하위 관계를 추출하는 방식에 대하여 제안한다. 본 논문에서는 판별하고자 하는 카테고리 구조뿐만이 아닌, 그와 관련된 다른 카테고리 구조까지 고려하여 카테고리 이름에 나타난 토큰들 간의 수식 그래프를 구축한 후, 그래프 분석 알고리즘을 통하여 각 카테고리 구조가 상하위 관계일 가능성에 대한 점수를 매긴다. 실험 결과, 본 알고리즘은 기존의 연구로 상하위 관계임을 판별할 수 없었던 일부 카테고리 구조에 대하여 성공적으로 상하위 관계인지를 판별하였다.

키워드 : 상하위 관계, 카테고리, 분류체계

Abstract In this paper, we propose a method to extract isa/instanceOf relation from category structure. Existing researches use lexical patterns to get isa/instanceOf relation from the category structure, e.g. head word matching, to determine whether the given category link is isa/instanceOf relation or not. In this paper, we propose a new approach which analyzes other category links related to the given category link to determine whether the given category link is isa/instanceOf relation or not. The experimental result shows that our algorithm can cover many cases which the existing algorithms were not able to deal with.

Key words : Taxonomic relations, Category, Taxonomy

1. 서 론

1.1 문제 정의

분류 체계는 문서 클러스터링[1], 데이터베이스 검색[2] 및 정보 추출[3]과 같은 작업들에 있어서 중요하게 사용된다. 따라서, 분류 체계를 자동으로 구축하기 위해서 수많은 연구들이 이루어져왔다. 일부 연구는 비구조화된 일반 문서로부터 분류 체계를 구축하고자 하였고[4,5], 일부 연구는 위키피디아 카테고리와 같이 구조화된 정보로부터 분류 체계를 얻어내고자 시도하였다[6-8]. 많은 연구들이 비구조화된 문서로부터 상하위 관계를 얻어내어 분류 체계를 구축하고자 시도하였으나, 대부분 낮은 정확률과 재현율을 보였다. 반면에, 구조화된 데이터로부터 상하위 관계를 얻어내고자 하는 시도는 상대적으로 높은 성능을 보였으나, 대규모의 분류 체계를 얻어내기 위해서는 그보다 훨씬 더 큰 크기의 구조화된 데이터를 요구하는 문제가 있다. 최근 들어 위키피디아[9]나 DBPedia[10]와 같은 구조화된 대용량의 데이터가 사용 가능해짐으로써 이 문제는 어느 정도 해결되었다.

본 연구에서는 주어진 카테고리 구조로부터 상하위 관계를 추출하는 방식에 대하여 제안한다. 즉, 본 연구에서는 어떤 임의의 카테고리 구조에 속하는 카테고리

· 이 논문은 2009 한글 및 한국어 정보처리 학술대회에서 '위키피디아 카테고리 구조를 이용한 상하위 관계 추출'의 제목으로 발표된 논문을 확장한 것임

† 학생회원 : KAIST 전산학과
cdh4696@world.kaist.ac.kr
†† 종신회원 : KAIST 전산학과 학과장
kschoi@cs.kaist.ac.kr

논문접수 : 2009년 11월 24일
심사완료 : 2010년 4월 19일

링크 각각에 대하여, 해당 카테고리 링크가 상하위(ISA/instanceOf) 관계를 의미하고 있는지, 또는 단순히 광의어(BT)/협의어(NT)/관련어(RT) 관계를 의미하고 있는지를 판별하는 알고리즘을 제안한다.

본 연구의 방법은 그 유효성을 입증하기 위하여 위키피디아 카테고리 구조에 적용되어, 위키피디아 카테고리 구조로부터 상하위 관계를 얻어내는 데 사용되었다. 위키피디아의 카테고리 구조는 카테고리와 페이지, 그리고 그것들간의 포함 관계로 이루어져 있다. 페이지는 위키피디아의 문서 하나를 의미하며, 카테고리는 이러한 페이지들과 다른 카테고리들을 무기명 다수의 일반인들이 임의로 분류한 후 이름을 붙인 것이다.

위키피디아 카테고리 구조는 전문가가 아닌 사람들에 의하여 구축되었지만, 다수의 사람들에 의하여 정제됨으로써 어느 정도의 신뢰성을 확보할 수 있고, 또한 위키피디아 카테고리 구조는 764,581개의 카테고리와 6,801,594개의 페이지 간의 35,904,116개의 포함 관계를 보유[1]하고 있어 대량의 상하위 관계를 추출하기 위한 좋은 자료가 된다.

본 논문에서는 2장에 관련 연구를 서술하고, 3장에 본 연구에서 제안하는 알고리즘에 대하여 구체적으로 서술한다. 4장에서는 본 연구의 알고리즘을 위키피디아의 카테고리 구조에 적용한 결과를 보이고, 5장에서 결론을 맺으며 추후 연구에 대하여 제안한다.

1.2 제안하는 모델의 구체적인 예

본 논문에서 임의의 카테고리 구조에 속하는 카테고리 링크를 <A, B, n>의 형태로써 표현하였을 때, A는 B를 포함하는 카테고리의 이름, B는 A에 의해 포함되는 카테고리 또는 페이지의 이름, n은 A와 B 사이에 존재하는 카테고리의 개수이다. 예를 들어, Wikipedia의 *iPod* 페이지는 *"2001 introduction", "portable media player", "industrial design examples", "IPod"* 라는 4개의 카테고리에 속해 있는데, 이를 위의 표현 방식으로 나타내면, *<2001 introduction, iPod, 0>, <Portable media players, iPod, 0>, <Industrial design examples, iPod, 0>, <IPod, iPod, 0>*이 된다.

위의 *iPod*의 예시에서 볼 수 있듯이, 모든 카테고리 구조가 상하위 관계로 전환될 수 있지는 않다. 카테고리 구조 *<Portable media players, iPod, 0>*는 카테고리 구조를 이루는 두 카테고리/페이지 이름이 상하위 관계를 이루는 것으로 볼 수 있지만, *<Industrial design examples, iPod, 0>, <2001 introduction, iPod, 0>*는 카테고리 링크를 이루는 두 카테고리/페이지의 이름이 상하위 관계를 이루는 것으로 볼 수 없으며, 단지 광의어/협의어 관계 또는 관련어 관계에 있을 뿐이다.

본 논문에서는 어떤 카테고리 링크 <A, B, 0>가 주어졌을 때, B를 하위 카테고리/페이지로 가지는 카테고리 이름들을 이용하여 주어진 카테고리 링크를 이루는 두 카테고리/페이지 이름이 상하위 관계인지를 나타내는 점수를 계산하고, 이 점수가 정해진 값 이상이 되면 두 카테고리/페이지 이름 A, B가 상하위 관계를 이루는 것으로 판별하고, 정해진 값 미만이 되면 두 카테고리/페이지 이름 A, B가 상하위 관계를 이루지 않는 것으로 판별하는 방법을 제안한다.

2. 관련 연구

카테고리로부터 분류 체계 확장을 위한 상하위 관계를 추출하는 연구 중 대표적인 것으로는 두 가지가 있다. [7]에서는 주어진 카테고리 링크에 포함된 두 카테고리 이름이 같은 중심어를 가지는지, 한 카테고리 이름에서의 수식어가 다른 카테고리 이름에서 중심어로 쓰이는지, 상위 카테고리 이름이 복수형인지 등을 검사하여 주어진 카테고리 링크가 상하위 관계인지 아닌지를 판별하였다. [8]에서는 카테고리 이름에서 나타나는 특정 패턴 5개(members of X, X [VBN IN] Y, X [IN] Y, X Y, X by Y)를 정의하고, 정의된 패턴이 카테고리 이름에서 발견되었을 시 정해진 규칙을 기반으로 하여 상하위 관계 및 기타 다른 관계들을 추출하였다.

지금까지 이루어진 카테고리를 기반으로 한 상하위 관계 추출 연구들은 카테고리 이름에서 어떤 특정한 패턴이 나타나야만 주어진 카테고리 링크가 상하위 관계인지를 판별할 수 있었다. [7]의 논문에서 보고된 바에 의하면, [7]의 방법으로는 349,263개의 카테고리-카테고리 링크 중 81,564 개에 대하여, 그것이 상하위 관계인지를 판별할 수 없었고, 또한 카테고리 - 페이지 링크에 관해서는 연구가 진행되지 않았다. 본 연구에서는 기존 연구의 이러한 한계를 극복하기 위하여, 카테고리 링크가 주어졌을 때 관련된 다른 카테고리 링크들을 이용하여 하위 카테고리/페이지의 본질 속성을 이루는 토큰들을 파악함으로써, 주어진 카테고리 링크에 포함된 두 카테고리/페이지 이름이 상하위 관계인지를 판별하는 방법을 제안한다.

3. 제안 모델

본 단원에서는 주어진 카테고리 링크가 상하위 관계인지를 판별하는 점수를 매기기 위해서 사용된 그래프 기반 알고리즘에 대하여 설명한다.

본 단원에서는 알고리즘의 용이한 설명을 위하여 다음과 같은 용어를 사용한다.

1) 2009년 8월 27일 현재

<A, B, n>: 상하위 관계인지 판별되어야 할 카테고리 링크. A가 상위 카테고리, B가 하위 카테고리/아티클이고, n은 A와 B 사이에 존재하는 카테고리의 개수이다.

U(A, n): 페이지 A의 상위 n 단계의 상위 카테고리들의 집합. 예를 들어, 주어진 카테고리 구조내에 카테고리 링크 *<portable media player, iPod, 0>*과 *<media player, portable media player, 0>*이 존재한다면, *portable media player*는 *U(iPod, 1)*에 속하지만, *media player*는 속하지 않는다.

본 알고리즘에서 제시된 방법은, 대상 카테고리 링크 *<A, B, 0>*이 주어졌을 때, *U(A, n)*을 이용하여 *<A, B, 0>*을 이루는 두 단어 A, B가 상하위 관계인지를 나타내는 점수를 매기는 것으로 간략하게 서술될 수 있다.

3.1 기본 원리

[14]에 따르면, 만약에 인스턴스 x가 개념 X와 *instanceOf*관계를 가지면, X의 정의는 x의 본질 속성을 포함한다. 카테고리화하는 것은 어떤 객체들을 그들의 공통된 속성을 이용하여 그룹화하는 것이다. 본질 속성은 어떤 객체의 정의와 밀접한 관련이 잇는 속성이므로, 대부분의 카테고리 구조에서 본질 속성이 객체들을 그룹화하는데 가장 많이 사용될 것이라고 가정할 수 있다:

가정 1. 본질 속성은 객체들을 카테고리화할 때 다른 속성들에 비해 많이 사용될 것이다.

즉, 어떤 아티클/카테고리를 나타내는 이름 B의 본질 속성을 나타내는 토큰은, B의 상위 카테고리의 이름들에서 많이 발견될 것이다.

그러나 가정 1이 본질 속성을 나타내지 않는 토큰이 상위 카테고리 이름에서 자주 나타나지 않으리라는 것까지 의미하는 것은 아니다. 예를 들어서, 다음 표 1에서 주어진 *U(iPod, 2)*의 원소의 명단에서 토큰 "*introduction*"은 무려 3회나 나타나는데, 이는 *iPod*의 본질 속성을 나타내는 토큰 "*player*"나 "*audio*"와 같은 횟수이다.

따라서, 어떤 객체의 본질 속성을 나타내는 토큰(지금부터 본질 토큰이라 명한다)을 얻어내기 위해서는 추가적인 가정이 필요하다:

가정 2. 본질 토큰은 다른 본질 토큰과 같이 나타날 것이다.

가정 2에 따르면, 어떤 토큰 A가 본질 토큰이고, 토큰 B가 어떤 카테고리 이름 내에서 A와 같이 등장할 경우, 토큰 B 또한 본질 토큰일 가능성이 높다. 이는, 서로 같이 자주 등장하는 두 토큰은 의미적으로 연관이 있을 가능성이 높다는 가정의 연장이다.

본 알고리즘에서는 '같이 등장하다'는 것의 정의를, 어떤 한 단어내에서 두 토큰이 각각 단어의 중심어와, 그 중심어의 수식어로 등장하는 것으로 정의하였다. 따라서, 가정 2는 다음의 두 갈래의 가정으로 세분화된다:

가정 2-1. 만약 어떤 수식어가 본질 속성을 나타내는 중심어들과 함께 자주 등장한다면, 그 수식어는 본질 속성을 나타낸다.

가정 2-2. 만약 어떤 중심어가 본질 속성을 나타내는 수식어들과 함께 자주 등장한다면, 그 중심어는 본질 속성을 나타낸다.

3.2 제안 알고리즘

위 3.1장에서 제시된 가정들을 토대로, 본 장에서는 각 토큰이 본질 속성이 될 가능성을 나타내는 점수(본질 속성성)를 얻어내는 그래프 기반 알고리즘을 서술한다. 3.1장의 가정에 따르면, 알고리즘에서는 각 토큰의 점수를 그 토큰이 *U<A, n>*에서 나타나는 빈도수와, 그 토큰이 같이 나타나는 중심어/수식어를 고려하여 계산되어야 한다. 본 알고리즘에서는 위 두 가지 고려 사항을 만족시키기 위하여, HITS page ranking algorithm[12]에 기반한 방식을 사용한다. 즉, 주어진 카테고리 링크 *<A, B, 0>*에 대하여, 먼저 *U<B, n>*을 이용하여 수식 그래프를 구축한 후, 변형된 HITS algorithm을 이용하여 *U<B, n>*에 등장한 각 토큰의 점수를 계산한다. 이후, 각 토큰의 점수를 이용하여 A와 B가 상하위 관계를 나타낼 가능성에 대한 점수를 얻는다. 만약 이 점수가 높으면, A와 B 사이에는 상하위 관계가 성립한다.

3.2.1 수식 그래프의 구축

수식 그래프는 각 노드가 *U<B, n>* 원소들에 포함된 각 토큰을 나타내고 각 변은 변의 두 정점이 각각 수식

표 1 U<iPod, 2>에 나타나는 토큰의 분석

U<iPod, 2>	2001_introductions, 2001, 2000s, 21st_century_introductions, 21st_century, Introductions_by_year, Portable_media_players, MPEG, ISO_standards, IEC_standards, Broadcast_engineering, Television_technology, Digital_audio_players, Digital_audio, MP3, Audio_players, Industrial_design_examples, Industrial_design, Design, Industry, Applied_sciences, iPod	
U<iPod, 2>에 나타나는 토큰들의 개수	introduction, player, audio, design	3
	2001, 21st, century, standards, digital, industrial	2
	2000s, by, year, portable, media, MPEG, ISO, IEC, broadcast, engineering, television, technology, MP3, example, industry, applied, science, iPod	1

어-중심어 관계로 *U<B, n>*의 원소들 가운데 등장하였었는지를 나타내는 방향성이 있는 그래프이다. 만약 카테고리 이름이 단 하나의 토큰만 포함하고 있을 경우, 먼저 새로운 더미 노드를 하나 추가한 다음 그 더미 노드로부터 해당 토큰으로 수식 관계를 만들어 준다. 예를 들어, *U<IPod, 2>*의 부분 집합인 *{2001 introductions, Portable media players, digital audio player, audio player, industrial design example, IPod}*의 경우, 그림 1과 같은 수식 그래프가 얻어진다:

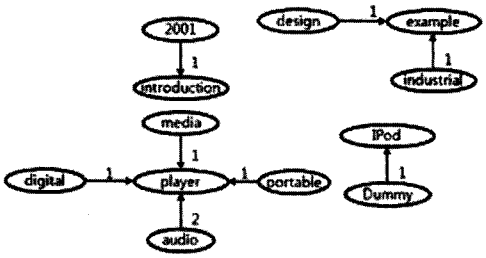

그림 1 *U<IPod, 2>*로부터 얻어진 수식 그래프

3.2.2 각 토큰의 본질 속성성 계산

수식 그래프를 구축한 후, 수식 그래프에 HITS 알고리즘을 적용하여 각 노드의 점수를 얻어낸다. 원래 HITS 알고리즘은 변의 가중치 값을 반영하지 않기 때문에, 이를 반영하도록 수정된 HITS 알고리즘[13]을 사용하였다:

$$Authority(V_i) = \sum_{V_j \in In(V_i)} e_{ji} \cdot Hub(V_j)$$

$$Hub(V_i) = \sum_{V_j \in Out(V_i)} e_{ij} \cdot Authority(V_j)$$

여기서, $In(V_i)$는 정점 V_i로 나가는 변이 있는 정점의 집합, $Out(V_i)$는 정점 V_i에서 들어오는 변이 있는 정점의 집합, 그리고 e_{ij}는 정점 V_j로부터 V_i로 가는 변의 가중치 값을 나타낸다.

HITS 알고리즘 적용 후 수식 그래프에서, *Authority score*는 중심어로서 그 노드의 본질 속성성을, 그리고 *Hub score*는 수식어로서의 그 노드의 본질 속성성을 나타낸다.

3.2.3 카테고리 링크의 점수 계산

위 모든 과정을 거친 뒤 얻어진 결과를 토대로, 카테고리 링크 <A, B, 0>의 점수는 다음과 같이 계산된다:

$$Score(<A,B,0>) = Authority(h) + \sum_{a \in mod(A)} Hub(a)$$

이 때, *Score(<A, B, 0>)*은 카테고리 링크 *<A, B, 0>*의 최종 점수, *h*는 *A*의 중심어, *mod(A)*는 *A*의 수식어들의 집합을 나타낸다. 이 때, *Score(<A, B, 0>)*이 높을수록 카테고리 링크 *<A, B, 0>*을 이루는 두 단어 A와 B가 상하위 관계를 이룰 가능성이 높다.

4. 실 험: 위키피디아 카테고리의 경우

본 논문에서 제시된 그래프 기반 방식의 정확도를 알아내기 위하여, 위에서 서술된 위키피디아 카테고리 구조를 사용하였다. 위키피디아 카테고리 링크 중 1,214개의 카테고리-카테고리 링크와 78개의 아티클-카테고리 링크를 랜덤하게 추출하여 각 링크가 상하위 관계를 나타내는지에 대한 여부를 어노테이션하였다. 어노테이션을 위해 두 명의 어노테이터가 동시에 작업한 후, 모순이 생길 시에는 서로 토론하여 올바른 결과를 선택하는 방법을 사용하였다. 1214개의 카테고리-카테고리 링크 중 848개가 상하위 관계로 판별되었으며, 78개의 아티클-카테고리 링크 중 52개가 상하위 관계로 판별되었다.

1,214개의 카테고리-카테고리 링크를 각각 600개와 614개의 링크를 가진 두 덩어리로 나누어, 600개의 링크를 가진 덩어리를 개발 세트로 사용하고 나머지 614개의 카테고리-카테고리 링크와 78개의 아티클-카테고리 링크를 이용하여 시스템의 성능을 테스트하였다. 파라미터는 정확률을 최대한 높이는 방식으로 설정되었는데, 이는 위키피디아는 이미 굉장히 방대한 양의 데이터를 보유하고 있기 때문에 재현률이 조금 낮아도 정확률이 높으면 대량의 양질의 데이터를 얻을 수 있기 때문이다.

4.1 사용 자질에 따른 비교

본 시스템에서 카테고리 링크의 점수를 계산하는 데 사용된 자질들 각각의 영향을 살펴보기 위하여, 각 자질만을 이용하여 점수를 계산하는 실험을 수행하였다. 표 2는 이 실험의 결과를 보여준다:

표 2 카테고리-카테고리 링크에 대한 실험 결과: 서로 다른 자질을 이용

	Precision	Recall	F-measure
Authority	0.7143	0.0341	0.0651
Hub	0.3333	0.0034	0.0068
Auth+Hub	0.7978	0.5105	0.6226

위 표에서 **Authority**는 점수 계산 시 *Authority score*만을 사용한 결과, **Hub**는 *Hub score*만 사용한 결과, **Auth+Hub**는 두 가지 점수를 모두 사용한 결과로서, 본 시스템에서 제시된 방법에 의하여 얻어진 결과이다. 실험 결과에서 보이듯이, 각 한 가지 자질만 사용하였을 경우 시스템은 매우 낮은 성능을 보이고, 두 가지 자질이 같이 사용될 경우에만 시스템은 정상적인 성능을 보이는 것을 알 수 있다.

4.2 그래프의 크기가 본 시스템의 성능에 미치는 영향

본 시스템에서 카테고리 링크의 분석에 사용되는 수식 그래프의 크기에 따른 성능을 분석하기 위하여, 수식

그래프를 구축하기 위한 상위 카테고리 수집 과정을 변경하여 각각 n 개 이상의 상위 카테고리를 수집하면 카테고리 수집 작업을 멈추도록 하였다.

표 3은 수집된 상위 카테고리 개수에 따른 시스템성능의 분석 결과를 나타낸다.

표 3 수식 그래프의 크기에 따른 성능의 비교

n	Precision	Recall	F-measure
10	0.8165	0.4403	0.5721
15	0.8074	0.4437	0.5727
20	0.7978	0.5105	0.6226
25	0.8182	0.3072	0.4467
30	0.8240	0.3515	0.4928

위 실험 결과에 따르면, 시스템은 n=20 정도일 때 가장 균형잡힌 결과를 보여 준다. 수집한 상위 카테고리의 수가 적으면 사용하기 위한 정보가 모자라 재현율이 떨어지는 것을 볼 수 있고, 반대로 수집한 상위 카테고리의 수가 많아져도 판별 대상 카테고리 링크와 상관이 없는 카테고리들이 수식 그래프에 많이 포함이 되기 때문에 재현율이 크게 떨어지는 것으로 보인다.

4.3 타 시스템과의 비교

본 시스템의 결과를 [7]의 실험 결과와 비교하였고, 표 4와 표 5는 이 실험의 결과를 보여준다. 개발 셋을 이용하여 파라미터는 0.8로 설정하였고, n=20으로 설정하였다.

표 4 카테고리-카테고리 링크에 대한 실험 결과

	Precision	Recall	F-measure
Baseline1	0.6987	1.0	0.8226
Baseline2	0.6479	0.9795	0.7799
A	0.7978	0.5105	0.6226
A + Head	0.8047	**0.7203**	**0.7601**
A + B	0.8204	0.7028	0.7571
B	**0.8944**	0.5629	0.6910

표 5 아티클-카테고리 링크에 대한 실험 결과

	Precision	Recall	F-measure
Baseline1	0.6706	1.0	0.8028
Baseline2	0.6667	0.9333	0.7778
A	0.8182	**0.5192**	**0.6353**
A + Head	0.8182	**0.5192**	**0.6353**
A + B	0.8148	0.4231	0.5570
B	**1.0**	0.0577	0.1091

A는 본 시스템의 결과, B는 [7]의 실험 결과, Baseline 1은 모든 링크를 상하위 관계로 판별했을 때의 결과, Baseline 2는 노드의 Authority score를 들어오는 변의 개수로 설정하고 Hub score를 나가는 변의 개수로 설정했을 때의 결과이다. A+Head는 먼저 주어진 링크 양 끝단의 중심어가 동일하면 상하위 관계로 판별하고 아닐 경우 본 시스템을 적용한 결과이며, A + B는 B를 이용하여 판별이 실패한 링크에 한해 A를 적용한 결과이다.

실험 결과에서 보이는 바와 같이, B의 시스템은 어휘적 특징을 많이 사용하기 때문에 카테고리-카테고리 링크에서는 잘 동작하지만, 아티클-카테고리 링크에서는 매우 낮은 재현률을 보인다. 반면에, 본 논문에서 제안된 시스템은 카테고리-카테고리 링크와 아티클-카테고리 링크에 대해 거의 동일하게 좋은 성능을 보이고 있음을 알 수 있다.

또한, 실험 결과에 따르면 본 논문 결과의 F-measure가 Baseline보다도 낮은 것을 알 수 있는데, 위키피디아의 카테고리 구조는 그 자체로 엄청난 양의 데이터를 보유하고 있기 때문에 재현률과 정확률의 비중을 동일시하는 F-measure는 본 연구에서 시스템간 직접적인 비교에 사용되기에는 적절치 않다. 다만 정확률이 재현률에 비해 시스템 성능 평가에서 어느 정도나 더 중요한 비중을 차지해야 하는가에 대해서는 확실하지 않은 면이 있기 때문에, 올바른 평가 척도에 대한 추가적인 연구가 필요하다.

5. 결 론

본 논문에서는 그래프 분석을 통하여 위키피디아 카테고리 구조에서 상하위 관계를 얻어내는 새로운 방법에 대하여 서술하였다. 다른 알고리즘과의 비교 결과, 본 논문에서 제시된 방법은 기존에 제시된 방법[7,8]을 이용하여 알아낼 수 없었던 상하위 관계들을 얻어낼 수 있었다. 현재는 단순히 카테고리의 이름과 동일한 하위 카테고리 구조를 가진 다른 카테고리 구조들을 사용하여 주어진 카테고리 구조가 상하위 관계인지 아닌지를 알아낼 수 있지만, 주어진 카테고리 구조에 포함된 페이지의 내용 등을 추가적인 자질로 이용할 수 있을 것이다. 이 부분은 추후 연구가 필요하다.

참 고 문 헌

[1] A. Hotho, S. Staab, G. Stumme, "Ontologies improve text document clustering," *Porceedings of the IEEE International Conference on Data Mining*, pp.541-544, 2003.

[2] S. Chakrabarti, B. Dom, R. Agrawal, P. Raghavan, "Using taxonomy, discriminants, and signatures for navigating in text databases," *Proceedings of the international conference on very large data bases*, 1997.

[3] M. Sanderson, B. Croft, "Deriving concept hierar-

chies from text," *Proceedings of the International Conference on New Methods in Language Processing*, 1994.
[4] P. Cimiano, A. Hotho, S. Staab, "Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis," *Journal of Artificial Intelligence Research*, vol.24, pp.305-339, 2005.
[5] P. Cimiano, A. Pivk, L. Schmidt-Thieme, S. Staab, "Learning Taxonomic Relations from Heterogeneous Sources of Evidence," *Ontology Learning from Text: Methods, Evaluation and Applications*, pp.59-73, 2005.
[6] J. X. Huang, J. A. Shin, K. S. Choi, "Enriching Core Ontology with Domain Thesaurus through Concept and Relation Classification," *OntoLex07*. 2007.
[7] S. P. Ponzetto, M. Strube, "Deriving a Large Scale Taxonomy from Wikipedia," *Proceedings of the national conference on artificial intelligence*, 2007.
[8] V. Nastase, M. Strube, "Decoding Wikipedia category names for knowledge acquisition," *Proceedings of the 23rd Conference on the Advancement of Artificial Intelligence*, pp.1219-1224. 2008.
[9] Wikipedia, http://www.wikipedia.org/
[10] DBpedia, http://dbpedia.org/About
[11] M. Collins, "Head-driven statistical models for natural language parsing," Ph.D. thesis, University of Pennsylvania, Philadelphia, 1999.
[12] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM*, vol.46, no.5, pp.604-632, 1999.
[13] R. Mihalcea, P. Tarau, "A Language Independent Algorithm for Single and Multiple Document Summarization," *Proceedings of IJCNLP 2005*, 2005.
[14] R. Mizoguchi, "Part 3: Advanced cource of ontological engineering," *New Generation Computing*, vol.22, no.2, pp.193-220, 2004.

최 동 현
2007년 2월 KAIST 전산학과 학사 취득 2007년~ KAIST 전산학과 석박사통합과정. 관심분야는 자연어 처리, 시맨틱 웹, 온톨로지

최 기 선
1978년 서울대 수학과 학사. 1980년 KAIST 전산과 석사. 1986년 KAIST 전산과 박사. 1988년~ KAIST 전산과 정교수. 2006년~ Semantic Web Research Center 센터장. 1998년~2006년 Korea Terminology Research Center for Language and Knowledge Engineering 센터장. 관심분야는 기계 번역, 자연언어처리, 정보 검색, 온톨로지, 시맨틱 웹