위키피디아 카테고리 구조를 이용한 상하위 관계 추출

최동현○ 최기선
KAIST 전산학과 시멘틱웹연구센터
cdh4696@world.kaist.ac.kr, kschoi@cs.kaist.ac.kr

ISA Relation Extraction from Wikipedia Category Structure

DongHyun Choi○ Key-Sun Choi
Semantic Web Research center, Computer Science Department, KAIST

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 위키피디아 카테고리 구조로부터 상하위 관계를 추출하는 방식에 대하여 제안한다. 본 논문에서는 판별하고자 하는 위키피디아 카테고리 구조뿐만이 아닌, 그와 관련된 다른 위키피디아 카테고리 구조까지 고려하여 카테고리 이름에 나타난 토큰들간의 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.

주제어: 상하위 관계, 위키피디아 카테고리, 분류체계

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, 7, 8]. 다수의연구들이 비구조화된 문서로부터 상하위 관계를 얻어내어 분류 체계를 구축하고자 시도하였으나, 다수의 연구들이 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 위키피디아의 카테고리 구조로부터 상하위 관계를 추출하는 방식에 대하여 제안한다. 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 위키피디아 카테고리 구조로부터 상하위 관계를 추출하는 방식에 대하여 제안한다.

1.2 제안하는 모델의 구체적인 예

본 논문에서 위키피디아 카테고리 구조는 <A, B, n>의 형태로써 표현되며, A는 B를 포함하는 카테고리의 이름, B는 A에 의해 포함되는 카테고리 또는 페이지의 이름, n은 A 와 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 iPod의 예시에서 볼 수 있듯이, 모든 카테고리 구조가 상하위 관계로 전환될 수 있지는 않다. 카테고리 구조 <Portable media players, iPod, 0>, <Industrial design examples, iPod, 0>는 상하위 관계로 볼 수 있지만, <2001 introduction, iPod, 0>는 상하위 관계가 아니고, 단지 두 개의 단어가 서로 관계가 있을 뿐이다. 본 논문에서는 어떤 카테고리 또는 페이지 이름 B가 주어졌을 때, B를 하위 카테고리/페이지로 가지는 모든 카테고리 구조 <A, B, n>에 대하여, 각 카테고리 구조가 상하위 관계가 될 수 있는 가능성을 점수로 계산하여 정렬한다. 이 때, 하나의 카테고리 구조에 대한 점수를 매기기 위하여, 페이지 이름 B가 포함된 다른 카테고리 구조에서

1) 2009년 8월 27일 현재

추출된 정보를 이용한다.

표 1은 페이지 iPod에 대하여, 본 논문에서 제시된 방법을 이용하여 카테고리 구조에 점수를 매겨 순위화한 것이다. n ≦ 2로 설정하였다.

표 1 페이지 iPod을 포함하는 카테고리 구조의 순위화 결과

입력	iPod	
입력을 하위로 가지는 카테고리 구조 (n≦2)	<2001_introductions, iPod,0> <2001, iPod, 1> <2000s, iPod, 2> <21st_century_introductions, iPod, 1> <21st_century, iPod, 2> <Introductions_by_year, iPod, 2> <Portable_media_players, iPod, 0> <MPEG, iPod, 1> <ISO_standards, iPod, 2> <IEC_standards, iPod, 2> <Broadcast_engineering, iPod, 2> <Television_technology, iPod, 2> <Digital_audio_players, iPod, 1> <Digital_audio, iPod, 2> <MP3, iPod, 2> <Audio_players, iPod, 2> <Industrial_design_examples, iPod, 0> <Industrial_design, iPod, 1> <Design, iPod, 2> <Industry, iPod, 2> <Applied_sciences, iPod, 2> <iPod, iPod, 0>	
정렬된 카테고리 구조	카테고리 구조	점수
	<Digital_audio_players, iPod, 1>	1.47
	<Portable_media_players, iPod, 0>	1.24
	<Audio_players, iPod, 2>	1.24
	<Digital_audio, iPod, 2>	0.36
	<MPEG, iPod, 1>	0
	이하, 모두 점수 0으로 동일	

2. 관련 연구

위키피디아 카테고리로부터 분류 체계 확장을 위한 상하위 관계를 추출하는 연구 중 대표적인 것으로는 두 가지가 있다. [7]에서는 주어진 카테고리 구조에 포함된 두 카테고리 이름이 같은 중심어를 가지는지, 한 카테고리 이름에서의 수식어가 다른 카테고리 이름에서 중심어로 쓰이는지, 상위 카테고리 이름이 복수형인지 등을 검사하여 주어진 카테고리 구조가 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 위키피디아 카테고리를 기반으로 한 상하위 관계 추출 연구들은 카테고리 이름에서 어떤 특정한 패턴이 나타나야만 주어진 카테고리 구조가 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 구조가 주어졌을 때 관련된 다른 카테고리 구조들을 이용하여 주어진 카테고리 구조가 상하위 관계인지를 판별하는 방법을 제안한다.

3. 제안 모델

본 단원에서는 주어진 카테고리 구조가 상하위 관계인지를 판별하는 점수를 매기기 위해서 사용된 두 가지의 알고리즘에 대하여 설명한다.

본 단원에서는 알고리즘의 용이한 설명을 위하여 다음과 같은 용어를 사용한다.

<A, B, n>: 상하위 관계인지 판별하고 싶은 카테고리 구조.

상위 카테고리: 카테고리 구조에서 첫 번째 항을 지칭

하위 카테고리: 카테고리 구조에서 두 번째 항을 지칭

S(a, b): a를 하위 카테고리로 가지고, n≦b인 카테고리 구조의 집합

U(a, b): S(a, b)에 속한 카테고리 구조의 모든 상위 카테고리 이름의 집합

T(a, b): U(a, b)에서 나타나는 모든 토큰의 집합

Freq(t, u): 토큰 t가 단어 u 안에서 등장한 횟수

$S_{Tok}(u)$: 단어 u에서 등장한 토큰의 집합

Head(u): 단어 u의 중심어

Mod(u): 단어 u의 수식어의 집합

아래에서 제시된 두 방법에서는, 공통적으로 T(a, b)의 각각의 원소에 대하여 점수를 매긴 다음, 그 점수를 이용하여 각각의 카테고리 구조에 점수를 매기게 된다.

3.1. 토큰의 개수에 대한 방법

본 방법에서는 판별하고자 하는 카테고리 구조<A, B, n>에 대하여, A를 이루고 있는 토큰들이 B의 다른 상위 카테고리에서도 많이 등장하고 있을 경우, <A, B>가 상

하위 관계일 가능성이 높다고 가정한다.

본 가정을 토대로 각 카테고리 구조 <A, B, n>가 상하위 관계인지를 나타내는 척도로서의 점수를 매기기 위하여, 먼저 T(a, b)의 각각의 원소 t에 대한 점수를 매긴다. 이를 위해, t가 U(a, b)의 각각의 원소 u 안에서 나타난 횟수를 측정한 후, 그 값을 모두 더한다.

즉, 모든 t ∈ T(a, b)에 대하여,

$$\text{Score}(t, a, b) = \sum_{u \in U(a,b)} Freq(t, u)$$

로 정의한다.

카테고리 구조 <A, B, n>에 대한 점수는 다음의 공식을 이용하여 구한다:

$$\text{Score}(<A, B, n>) = \sum_{t \in S_{Tok}(A)} Score(t, B, b)$$

b는 임의로 지정해 줄 수 있으며, 본 논문에서는 b=2의 값을 사용하였다. 어떤 토큰에 대한 점수가 정의되어 있지 않을 경우, Score(<A, B, n>)은 무조건 0이 된다.

표 2는 표 1의 입력을 본 방법을 사용하여 각 토큰에 점수를 매긴 예시이다. 표 3은 표 2의 점수를 이용하여 표 1의 각 카테고리 구조에 점수를 매긴 방식이다.

표 2 T("iPod", 2)의 각 원소에 개수세기 방식으로 점수를 매긴 예시

t ∈ T("iPod", 2)	Score(t, "iPod", 2)
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

표 3 표 1의 카테고리 구조들 각각에 표 2의 결과를 이용하여 점수를 매긴 예시

카테고리 구조	점수
<Digital_audio_players, iPod, 1>	8
<21st_century_introductions, iPod, 1>	7
<Audio_players, iPod, 2>	6
<Industrial_design_examples, iPod, 0>	6
<2001_introductions, iPod,0>	5
<Introductions_by_year, iPod, 2>	5
<Portable_media_players, iPod, 0>	5
<Digital_audio, iPod, 2>	5
<Industrial_design, iPod, 1>	5
<21st_century, iPod, 2>	4
<ISO_standards, iPod, 2>	3
<IEC_standards, iPod, 2>	3
<Design, iPod, 2>	3
<2001, iPod, 1>	2
<Broadcast_engineering, iPod, 2>	2
<Television_technology, iPod, 2>	2
<Applied_sciences, iPod, 2>	2
<2000s, iPod, 2>	1
<MPEG, iPod, 1>	1
<MP3, iPod, 2>	1
<Industry, iPod, 2>	1
<iPod, iPod, 0>	1

3.2. 수식 관계를 나타내는 그래프에 기반한 모델

위 3.1장에서 제시된 것은 가장 직관적인 방법이나, 많은 오류가 존재한다. 위 표 3에서 두 번째 순위로 랭크된 카테고리 구조 <21st_century_introductions, ipod, 1>의 경우, 상하위 관계를 나타내고 있다고는 보기 힘들다. 이 카테고리 구조가 개수세기 방법에서 상위에 올라선 이유는, "introduction"이라는 토큰이 iPod의 상위 카테고리 이름들과의 관련성에 비하여 나타나는 빈도수가 높기 때문이다.

이 문제를 해결하기 위하여, 먼저 다음의 두 가지 가설을 세운다:

가정 1. 어떤 중심어가 여러 가지 중요한 수식어에 의해 수식받는다면, 그 중심어는 중요하다.

가정 2. 어떤 수식어가 여러 가지 중요한 중심어들을 수식한다면, 그 수식어는 중요하다.

여기서, "중요하다"는 것의 의미는 그것이 하위 카테고리/페이지의 주요한 특징을 나타내고 있음을 의미한다.

위의 가정을 통해 문제를 해결하기 위하여, 본 방법에서는 먼저 수식어 그래프를 구축한다. 어떤 카테고리 구조 <A, B, n>에 대한 수식어 그래프는 다음과 같이 정의된다:

G := (V, E)

V := {v| 각 정점 v는 토큰 t∈T(B, b) 에 대응}

E := {(v1, v2)| v1의 토큰 t1은 v2의 토큰 t2를 U(B, b)의 원소 내부에서 수식}

이 때, 변의 가중치는 U(B, b)에서 해당 수식 관계가 나타난 개수로 정의된다. 중심어는 [11]의 방법을 이용하여 추출되었다.

표 1의 예시를 수식 관계 그래프로 나타내면 그림 2와 같이 표현된다.

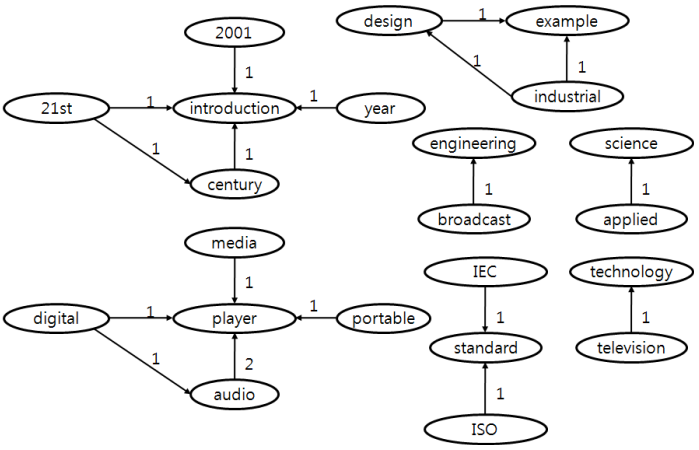

그림 1 표 1의 예시에서 추출된 수식 그래프

수식 그래프에서 위 가정 1과 2에서 서술된 것과 같은 중심어와 수식어의 중요도를 얻어내기 위하여 HITS 링크 분석 알고리즘[12]을 사용한다. HITS 알고리즘을 사용하는 이유는, 하나의 중요도 점수만 반환하는 다른 링크 분석 알고리즘과는 달리 본 논문의 가정에 적합한 두 가지 종류의 중요도 점수 - 얼마나 다른 중요한 페이지에 의해 링크당했는지를 나타내는 Authority 점수와 얼마나 얼마나 다른 중요한 페이지를 링크하는지를 나타내는 Hub 점수 -를 반환하기 때문이다. 본 논문에서는 HITS 알고리즘의 Authority 점수를 중심어의 중요도로, HITS 알고리즘의 Hub 점수를 수식어의 중요도로 가정하였다. 또한, 변의 가중치를 계산식에 넣기 위하여 [13]에서 사용된 변형을 이용하였다.

변형된 HITS 알고리즘의 수식은 다음과 같다.

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

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

위 수식에서, $In(V_i)$는 정점 V_i로 들어오는 변을 가진 정점의 집합을 의미하고, $Out(V_i)$는 정점 V_i에서 나오는 변을 가진 정점의 집합을 의미하며, e_{ij}는 V_i에서 V_j로 가는 변의 가중치를 의미한다.

HITS 알고리즘은 먼저 각 정점의 Authority 및 Hub 점수들을 1로 초기화한 후, 먼저 Authority 점수를 갱신하고 이후 Hub 점수를 갱신한다. 점수 갱신은 각 점수가 어느 일정 수준으로 수렴할 때까지 계속된다.

카테고리 구조 <A, B, n>의 점수는 다음과 같이 계산된다:

Score(<A, B, n>) =

$$Authority(Head(A)) + \sum_{t \in Mod(A)} Hub(t)$$

표 1의 결과는 본 단원에서 제시된 그래프 기반 분석 방법을 토대로 계산된 것이다.

4 실 험

본 논문에서 제시된 그래프 기반 방식의 정확도를 알아내기 위하여, 위키피디아 카테고리 구조 중 그래프 기반 방식으로 0.5점 이상을 얻은 100개의 카테고리 구조를 랜덤하게 추출하였다. 이 100개의 카테고리 구조에 대하여, 2명의 어노테이터가 수작업으로 상하위 관계인지 아닌지를 판별하였다. 표 4는 두 명의 어노테이터의 판별 결과가 얼마나 일치하는지 보여 준다.

표 4 어노테이터 평가 - 컨퓨젼 매트릭스

Annotator 1 \ Annotator 2	O	X
O	57	14
X	3	26

표에서 보는 바와 같이, 두 어노테이터의 평가는 83 % 일치 하였다. 두 어노테이터의 평가가 일치된 83개의 카테고리 구조에 관하여, 68.7 %인 57개의 카테고리 구조에 대하여 두 어노테이터 모두 상하위 관계라고 판단하였고, 31,3 %인 나머지 26개의 카테고리 구조는 모두 상하위 관계가 아니라고 판단하였다.

그래프 기반 방식의 성능을 기존 연구와 비교하기 위하여 [7]의 방법을 적용하여 두 어노테이터의 평가가 일치한 83개의 카테고리 구조의 상하위 관계를 판별하였다. 표 5는 [7]의 어휘적인 규칙을 사용하는 방법의 결과를 나타낸다.

표 5 [7]의 방법 평가

Annotator \ Ponzetto[7]	O	Unknown
O	32	25
X	17	9

표 5에서 볼 수 있는 것과 같이, [7]의 방법으로 상하위 관계인지를 판별하는 것이 불가능한 카테고리 구조들을 상하위 관계가 아니라고 결정하면, 동일한 카테고리 구조에 대하여 [7]의 방법은 단지 41개만이 어노테이터의 평가와 일치하였다. 즉, [7]의 방법은 47.1 %의 정확도를 보였다.

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.

감사의 글

본 논문은 지식경제부 및 정보통신연구진흥원의 정보통신선도기반기술개발사업의 연구결과로 수행되었습니다.

6 참고 문헌

[1] Hotho, A., Staab, S., Stumme, G.: Ontologies improve text document clustering. Proceedings of the IEEE International Conference on Data Mining (2003)

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

[3] Sanderson, M., Croft. B.: Deriving concept hierarchies from text. Proceedings of the International Conference on New Methods in Language Processing (1994)

[4] Cimiano, P., Hotho, A., Staab, S.: Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis. Journal of Artificial Intelligence Research (2005)

[5] Cimiano, P., Pivk, A., Schmidt-Thieme, L., Staab, S.: Learning Taxonomic Relations from Heterogeneous Sources of Evidence. Ontology Learning from Text: Methods, Evaluation and Applications (2005)

[6] Huang, J.X., Shin, J.A., Choi, K.S.: Enriching Core Ontology with Domain Thesaurus through Concept and Relation Classification. OntoLex07 (2007)

[7] Ponzetto, S.P., Strube, M.: Deriving a Large Scale Taxonomy from Wikipedia. Proceedings of the national conference on artificial intelligence (2007)

[8] Nastase, V., Strube, M.: Decoding Wikipedia category names for knowledge acquisition .Proceedings of the 23rd Conference on the Advancement of Artificial Intelligence, pp. 1219-1224. (2008)

[9] http://www.wikipedia.org/

[10] http://dbpedia.org/About

[11] Collins, M.: Head-driven statistical models for natural language parsing. Ph.D. thesis, University of Pennsylvania, Philadelphia (1999)

[12] Kleinberg, J. M.: Authoritative sources in a hyper-linked environment. Journal of the ACM, 46(5):604-632 (1999)

[13] Mihalcea, R., Tarau, P.: A Language Independent Algorithm for Single and Multiple Document Summarization. Proceedings of IJCNLP 2005 (2005)