한국어 위키피디아 카테고리에서의 자동적 지식 획득

김지성○, 김정욱, 최기선
한국과학기술원 전산학과
{jiseong, prismriver, kschoi}@kaist.ac.kr

Automatic Acquisition of Knowledge from Korean Wikipedia Categories

Jiseong Kim○, Jung-uk Kim, Key-Sun Choi
Dept. of Computer Science, KAIST

요 약

하나의 위키피디아 아티클은 보통 여러 개의 카테고리들이 부여되어 있다. 카테고리는 완전한 문장이 아닌 소수의 단어들로 구성된 명사구 형태를 띠고 있다. 카테고리에는 개체(entity)에 대한 개념(concept) 혹은 개체 간의 관계에 대한 정보가 풍부히 포함되어 있다. 제안하고자 하는 것은 이 카테고리에 담긴 정보를 위키피디아와 연동된 디비피디아의 개체와 프로퍼티(property)에 대한 RDF 트리플렛(triplet) 형태로 변환하는 것이다. 이 때 이미 존재하는 RDF 트리플렛 형태의 시드 데이터(seed data)에서 카테고리를 표현하는 문자열의 내부 표현인 어휘 패턴(lexical pattern)을 학습하여 활용하는 지식 획득 방법을 제안한다. 실험에서는 디비피디아 트리플렛을 시드 데이터로 이용해 위키피디아 카테고리에서부터 0.765의 정확도를 갖는 약 54만개의 관계 인스턴스(relation instance)를 추출해 낼 수 있었다. 위키피디아는 계속해서 증가하는 지식의 원천이기 때문에, 여기서 제안하는 방법을 이용하면 꾸준히 갱신되는 위키피디아 카테고리에서부터 지식을 추출하여 디비피디아를 지속적으로 강화할 수 있다.

1. 서 론

위키피디아는 지속적으로 확장하는 지식의 원천으로써 추상적인 개체 혹은 현실 세계의 개체에 대한 다양한 정보를 포함한다. 위키피디아는 개체를 다양한 방법으로(예를 들면 아티클, 인포박스, 카테고리) 설명하며, 이는 개체 그 자체 혹은 개체 간의 관계에 대한 설명을 풍부히 포함한다. 위키피디아에서 주목할만한 지식의 원천 중 하나는 카테고리 시스템이다. 각각의 위키피디아 아티클은 유저에 의해 여러 개의 카테고리가 부여된다. 위키피디아 카테고리는 단순 명사 혹은 다양하게 수식된 명사로써 표현되며, 이로부터 정보 추출 기법을 이용하여 다양한 지식을 추출할 수 있다. 위키피디아 아티클은 보통 하나의 개체를 설명하며, 아티클에 부여되는 카테고리는 그 개체의 카테고리라고 할 수 있다. 따라서 위키피디아 카테고리에서 개체에 대한 개념 혹은 개체 간의 관계에 대한 정보를 추출할 수 있다. 본 논문에서는 시드 데이터를 이용해 자동으로 학습한 어휘 패턴으로 개체에 대한 관계 정보를 추출하는 방법을 제안한다. 먼저 시드 데이터에서 관계에 대한 신뢰도 높은 어휘 패턴을 학습하고, 이를 이용해 위키피디아 카테고리에서부터 지식을 추출해 낸다. 어휘 패턴을 수동으로 제작하던 기존 접근법과는 달리 본 논문에서 제안하는 방법은 기존의 디비피디아[1]를 시드 데이터로 사용하여 자동으로 패턴을 학습하며, 사람이 들이는 수고를 보다 덜어줄 수 있다. 관계의 종류를 다양화하거나 데이터의 양을 거대화 할수록 수동으로 패턴을 만드는 것은 불가능에 가깝지만, 본 논문에서 제안하는 방법을 이용하면 자동으로 패턴을 학습함으로써 보다 다채롭고 풍부한 관계적 정보(relational information)를 추출하는 것이 가능해 진다.

2. 관련 연구

최근에, 위키피디아 카테고리에서부터 관계적 정보를 추출하는 방법에 대한 몇몇 연구들이 선행되어 있다. 예를 들어, Suchanek 외 몇몇 연구진은 2007년에 위키피디아의 카테고리와 인포박스 및 워드넷의 분류적 관계(taxonomic relation)를 이용해 거대한 온톨로지 야고(YAGO)를 개발했다[2]. 야고는 주로 *is-a* 관계를 추출하는데 초점을 맞추었으며, 사람의 손을 빌어 각 관계(예, *locatedIn*)에 대한 어휘 패턴(예, *Rivers in x*)을 만들어 사용하였다. 또 다른 예로, Liu 및 기타 연구진은 2008년에 *Catriple*이라는 방법을 제안하였는데[3], 이 방법 또한 각 관계에 대한 어휘 패턴을 수동으로 만들었고 이를 이용해 위키피디아 카테고리에서부터 지식을 추출하였다. 조금 다른 점이라면 부모 카테고리와 자식 카테고리의 의미적 관계를 이용해 추출 가능한 관계 인스턴스를 증폭시켰다는 점이다. 이와 같은 해에 Nastase 및 기타 연구진은 비슷한 접근법을[4] 제시하였는데, 카테고리들을 관계적 정보가 얼마나 명확히 표현되었느냐에 따라 몇 개의 클래스로 분류하고 각 클래스마다 어휘 패턴을 수동으로 만들어 지식을 추출하였다. 또한 의미적 관계 및 클래스의 속성을 카테고리에서 습득하여 이를 계층 전 구간으로 전파시켜 얻는 지식의 양을 증폭시켰다. 앞서 언급한 세 연구는 모두 관계에 대한 어휘 패턴을 수동으로 만들며 이를 이용해 카테고리에서부터 지식을 추출한다. 이는 다른 언어권의 데이터로 넘어갈 때 패턴을 모두 새로 만들어 주어야 하는 큰 비용이 뒤따른다. 본 논문에서 제안하는 방법은 이런 관계에 대한 어휘 패턴을 자동으로 학습하여 사람에 의한 수고를 줄여줄 수 있다.

3. 선행 사항

개체. 위키피디아는 추상적인 개체 혹은 현실 세계의 개체를 설명하는 백과사전의 일종으로써, 각 아티클마다 부여된 카테고리는 그 아티클이 설명하는 개체의 카테고리라고 볼 수 있다. 직관적인 이해를 위해, 앞으로

위키피디아 아티클 대신 *개체*라는 용어를 사용한다.

관계 인스턴스. 개체 x와 개체 혹은 값 y사이에 관계 r이 있을 때, 우리는 이 것을 혼 문절(Horn clause, 예를 들면 $r(x,y)$), 트리플렛(triplet, 예를 들면 $<x,r,y>$) 등등의 방법으로 표현할 수 있다. 본 논문에서는 관계 인스턴스를 트리플렛으로 표현한다.

카테고리 타입. 카테고리는 포함한 정보의 종류에 따라 몇 가지 타입으로 분류할 수 있다. 그 중 하나는 *개념적 카테고리*로써 개체의 클래스에 대한 정보를 포함한다. 예로, 알버트 아인슈타인은 *미국의 귀화 시민* 카테고리에 속해있다. 또 다른 타입으로는 개체 간의 관계적 정보를 포함하는 *관계적 카테고리*가 있다. 예를 들면, *1989 출생* 카테고리는 개체와 *1989* 간의 관계를 표현한다. 그 밖에도 *물리, 역사* 등과 같이 단순히 테마적 내용을 나타내기 위한 타입이 있다. 본 논문에서 제안하는 방법은 주로 *개념적 카테고리*와 *관계적 카테고리*를 지식의 원천으로 활용한다.

카테고리 문장. 카테고리의 문장은 대게 소수의 전치사, 부사, 형용사 및 특수 문자(예를 들면, 하이픈(-), 콤마(,))를 곁들인 명사구로써 표현된다. 카테고리 계층 상에서 상위 계층에 해당하는 기본 카테고리들은 거의 명사 위주로만 구성된 간단한 문장으로 표현되며, 하위 계층으로 갈수록 보다 세밀한 분류를 위해 복잡하게 수식된 긴 문장으로 표현된다. 카테고리는 완벽한 구성 성분을 갖는 문장보다 더욱 압축적으로 간단하게 표현되기 때문에, 이로부터 어휘 패턴을 학습하고 사용하는 데에 있어 많은 문제점을 동반한다.

목표. 본 연구는 관계에 대한 어휘 패턴을 학습하고 이를 각 개체가 속한 카테고리에 적용하여 개체에 대한 관계 정보, 즉 관계 인스턴스를 추출하는 것을 목표로 한다.

4. 접근 방법

4.1 어휘 패턴 학습

어휘 패턴. 카테고리 문장은 크게 오브젝트 부분과 관계에 대한 어휘 패턴 부분으로 구성이 되어 있다. 예를 들면, 어떤 개체 e가 *제주도 출신 학자* 카테고리에 속해있을 때, 이 카테고리는 오브젝트 부분인 *제주도*와 *출생지*라는 관계의 어휘 패턴인 *출신 학자* 부분으로 구성된다. 즉 이 카테고리에서 $<e, 출생지, 제주도>$ 를 추출해낼 수 있다. 하나의 카테고리 문장은 이런 오브젝트와 어휘 패턴에 대한 쌍을 두 개 이상 포함할 수 있다.

학습 과정. 전체 과정의 첫 단계는 카테고리와 시드 관계 인스턴스로부터 관계에 대한 어휘 패턴을 학습하는 것이다. 패턴 학습 과정은 다음과 같다:

1. 어휘 패턴 학습에 사용될 시드 관계 인스턴스를 준비한다. 어휘 패턴은 오직 시드 데이터에 포함된 관계에 대해서만 추출된다.
2. 모든 개체에 대해 다음을 수행한다: 개체 e가 속한 카테고리 집합 $C_e = \{c_i\}_{i=0}^{n}$ 의 원소들과 시드 관계 인스턴스 집합 $S_e = \{<e, r_i, o_i>\}_{i=0}^{m}$의 원소 간의 모든 쌍의 조합을 비교한다. 만약 어떤 $c_k \in C_e$ 와 부분문자열과 일치하는 오브젝트를 갖는 $<e, r_j, o_j> \in S_e$ 이 존재할 경우, o_j 와 일치하지 않는 c_k 의 남은 부분문자열을 r_j의 어휘 패턴으로 한다. 이런 식으로 각 관계에 대한 모든 어휘 패턴을 추출하고 보관해 둔다.

4.2 빈도 기반 패턴 필터링

패턴의 빈도. 관계 r에 대한 어휘 패턴 p의 빈도수 $f_{p,r}$은 관계 r에 대해 패턴 p를 추출할 때 사용된 시드 관계 인스턴스의 개수로 정의된다.

빈도의 지역 평균(local average). 패턴의 빈도는 다음과 같이 지역적으로 평균화 할 수 있다:

$$\bar{f_p} = \frac{\sum_{r \in R} f_{p,r}}{|R|} \quad (1)$$

여기서 R은 p를 어휘 패턴으로 갖는 관계의 집합이다.

패턴 필터링. 시드 데이터의 오브젝트와 카테고리의 부분 문자열이 우연히 일치함으로써 관련 없는 잘못된 어휘 패턴이 추출될 수 있다. 이런 것들을 걸러주기 위해 빈도 기반 필터링을 실시하였다. 이 때, 빈도의 지역 평균을 빈도 임계치(threshold)로 사용하였다.

4.3 관계 인스턴스 추출

패턴의 모호성(ambiguity). 압축적으로 표현된 카테고리의 특성상 이로부터 추출한 소수의 단어로 이루어진 어휘 패턴은 우연히 여러 관계들에 의해 공유될 확률이 높다. 이는 어떤 카테고리가 패턴과 매칭될 때 그 패턴을 갖는 여러 관계 중 올바른 것을 선택해야 하는 문제를 발생시킨다. 본 논문에서 제안하는 방법은 이 문제를 관계의 레인지(range)를 이용하여 부분적으로 해소한다.

추출 과정. 관계 인스턴스를 추출하는 과정은 다음과 같다:

1. 모든 개체에 대해 다음을 수행한다: 모든 학습된 패턴(P)과 개체 e가 속한 카테고리(C_e)를 비교한다. 만약 어떤 관계 r의 어휘 패턴 $p \in P$가 $c \in C_e$와 부분 매칭하며, 매칭하지 않는 c 의 남은 부분문자열의 길이가 0보다 클 경우, 관계 인스턴스 $<e, r, c-p>$를 추출한다. 여기서 $c-p$는 p와 매칭하지 않는 c의 남은 부분문자열을 나타낸다.
2. 모든 추출된 관계 인스턴스 $<e, r, o>$에 대하여 o가 r의 레인지에 속하는지 검사한다. 만약 속하지 않을 경우 추출 결과에서 제외한다.

5. 실험

5.1 데이터 셋 및 실험 환경

카테고리 데이터. 카테고리 데이터는 4월 말 / 5월 초에 추출된 위키피디아 덤프의 888,105개의 개체-카테고리 쌍을 사용하였다. 순수 카테고리의 개수는 86,560개이다.

시드 데이터. 한국어 위키피디아에서 추출해낸 디비피디아 2014의 696,013개의 매핑 기반 프로퍼티를 시드 관계 인스턴스로 사용하였다.

평가 방법. 이 문제에 있어 완벽한 정답셋을 만드는 것은 불가능하기 때문에, 추출된 결과를 사람이 수동으로 검증하였다. 추출된 관계 인스턴스 중 200개를 무작위로 채집(sampling)하여 두 사람이 동시에 옳다고 한 결과만을 맞는 것으로 간주하였다. 그 기준은 다음과 같다:

1. 추출 시 사용한 카테고리에서부터 해당 관계

인스턴스가 나타내는 정보를 추론해 낼 수 있다.

2. 추출된 관계 인스턴스의 서브젝트 및 오브젝트가 관계의 도메인, 레인지에 속한다.

5.2 실험 결과 및 분석

실험 결과 및 정확도. 실험 결과 총 100개의 관계에 대한 4,510개의 어휘 패턴을 추출해낼 수 있었으며, 이를 이용해 정확도가 0.765로 측정된 총 541,000개의 새로운 관계 인스턴스를 추출해 낼 수 있었다. 전체 추출 결과는 웹사이트[1]를 통해 확인할 수 있다.

고 빈도 관계. 추출한 100가지 종류의 관계 중 가장 빈번한 상위 10개가 표 1에 나타나 있다. 이를 통해 위키피디아 카테고리 시스템이 이런 종류의 관계적 정보를 풍부히 담고 있음을 알 수 있다.

birthPlace	86,581	*genre*	28,574
StartYears	52,806	*location*	25,014
birthYear	48,576	*region*	22,772
deathPlace	31,963	*education*	21,373
nationality	30,413	*youthClub*	19,409

표 1. 상위 10개 최고 빈도 관계들 및 그에 대한 빈도

시드 데이터 크기. 그림 1은 시드 데이터의 크기에 따라 추출되는 결과의 크기가 어떻게 달라지는지를 보여준다. 그림의 위쪽 그래프는 한국어 데이터에 대해 실시한 실험 결과이며, 시드 데이터 크기가 10% 이하일 때 추출된 인스턴스와 커버된 카테고리의 양이 급격히 감소함을 볼 수 있다. 이는 카테고리 개수의 약 80.4%에 해당하는 양이다. 그림의 아래쪽은 영어 데이터에 대해서 실시한 실험 결과이며 한국어 데이터와 비슷한 양상을 보임을 알 수 있다(영어 같은 경우 시드 데이터의 양이 카테고리 개수의 93.7% 이하일 때부터 급격히 감소했다). 우리는 이를 통해서 부분적으로 다음과 같은 결론에 도달할 수 있다: 카테고리의 정보를 최대한 활용하기 위해서는 적어도 시드 데이터의 사이즈가 카테고리의 사이즈와 비슷하거나 크도록 하는 것이 권장된다.

오류 분석. 분석된 주요 오류 원인은 다음과 같다:

1. 카테고리가 포함하는 옳은 관계가 틀린 관계와 밀접히 관련된다. 예를 들면, x 라는 나라에서 태어난(옳은 관계) 사람은 x 에서 죽을(틀린 관계) 확률이 높다.
2. 우연히 시드 데이터의 오브젝트와 카테고리의 부분 문자열이 일치하여 관련 없는 어휘 패턴이 추출돼 사용되었다.
3. 도메인과 레인지 조건을 만족시키지 않는다.

첫 번째 원인이 전체 오류의 51.5%를 차지했다. 실제 세계에서 첫 번째 케이스는 참인 경우가 많으므로, 실제 정확도는 본 연구진이 측정한 것보다 높을 수 있다.

6. 결론 및 향후 연구

본 논문에서는 관계에 대한 어휘 패턴을 자동으로 학습하고, 이를 이용해 관계적 정보를 위키피디아 카테고리에서부터 추출하는 방법에 대해 논의하였다. 여기서 제안한 접근법을 이용하면 계속해서 업데이트되는 위키피디아 카테고리에서부터 자동적으로 패턴 및 지식을 추출하여 지속적으로 디비피디아를 강화할 수 있다. 아직 이 접근법은 많은 면에 있어 개선이 가능하다(예를 들면, 보다 합리적인 빈도 임계치). 본 연구팀은 추후 본 접근법을 보다 확장하고 개선하는 연구를 진행할 계획이다.

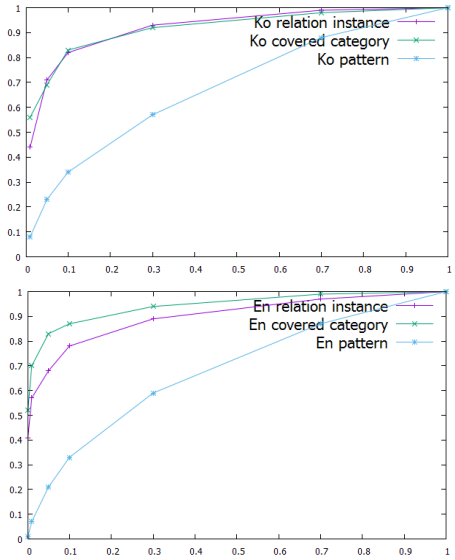

그림 1. 시드 데이터의 크기 비율(x)에 따른 추출된 인스턴스 수, 패턴 수, 커버된 카테고리 수(y)의 정규화된 분포 (위 그래프는 한국어 데이터 대상, 아래는 영어 데이터 대상)

사 사

이 논문은 2015 년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임. (No. R0101-15-0054, WiseKB: 빅데이터 이해 기반 자가학습형 지식베이스 및 추론 기술 개발)

참 고 문 헌

[1] Auer, Sören, et al., “Dbpedia: A nucleus for a web of open data”, Springer Berlin Heidelberg, pp. 722-735, 2007.

[2] Suchanek, Fabian M., et al., "Yago: a core of semantic knowledge", *Proceedings of the 16th international conference on World Wide Web*. ACM, pp. 697-706, 2007.

[3] Liu, Qiaoling, et al., “Catriple: Extracting triples from Wikipedia categories”, In *The Semantic Web*, pp.330-344, 2008.

[4] Nastase and Strube, “Decoding Wikipedia Categories for Knowledge Acquisition”, In *AAAI*, pp. 1219-1224, 2008.

[1] http://elvis.kaist.ac.kr/demos/kcc2015_c2k/