온톨로지 기반 친족관계 추출 방법

황명권, 최동진, 김판구
조선대학교 컴퓨터공학과
e-mail: mg.hwang@gmail.com, dongjin.choi84@gmail.com, pkkim@chosun.ac.kr

A Method on Relative Relation Extraction based on Ontology

Myung-Gwon Hwang, Dong-Jin Choi, Pan-Koo Kim
Dept of Computer Engineering, Chosun University

요 약

시맨틱 웹의 발전과 더불어 소셜 네트워크 자동 구축에 대한 연구가 활발히 진행되고 있다. 본 논문은 온톨로지를 기반의 소셜 정보 추출에 대한 방법을 다루고 있으며, 특히, 이에 필요한 온톨로지 모델링, 사람들 사이의 관계 추출을 위한 패턴 정의에 대해 기술하고 있다. 온톨로지와 패턴을 기반으로 역사적 인물들의 친족관계를 파악함으로써 소셜 정보의 추출에 대한 가능성을 미리 짐작해 본다.

1. 서론

정보의 의미적 연결을 지향하는 시맨틱 웹과 더불어 사람들의 관계를 추출하고 정의하기 위한 소셜 네트워크 구축 방안에 대한 연구가 활발하게 진행되고 있다. 소셜 네트워크는 정보의 공유, 지역성, 명확성, 신뢰도 등에 초점을 맞추어 진행되고 있다[1]. 본 논문은 웹에 게시되어 있는 문서들의 내용분석을 통해 소셜 정보를 추출하는 방법을 다루고 있으며, 이를 위한 전체 구조를 기술하고, 그에 필요한 세부 모듈들에 대한 설명을 기술한다. 그리고 위키피디아의 역사적 인물에 대한 문서들을 분석함으로써 본 연구의 가능성을 살펴본다. 본 연구의 본론 구성은 친족관계 온톨로지 모델링, 웹 문서에서 소셜 정보와 친족관계를 추출하기 위한 패턴 정의, 그리고 온톨로지와 패턴을 웹 문서로 적용으로 이루어져 있다.

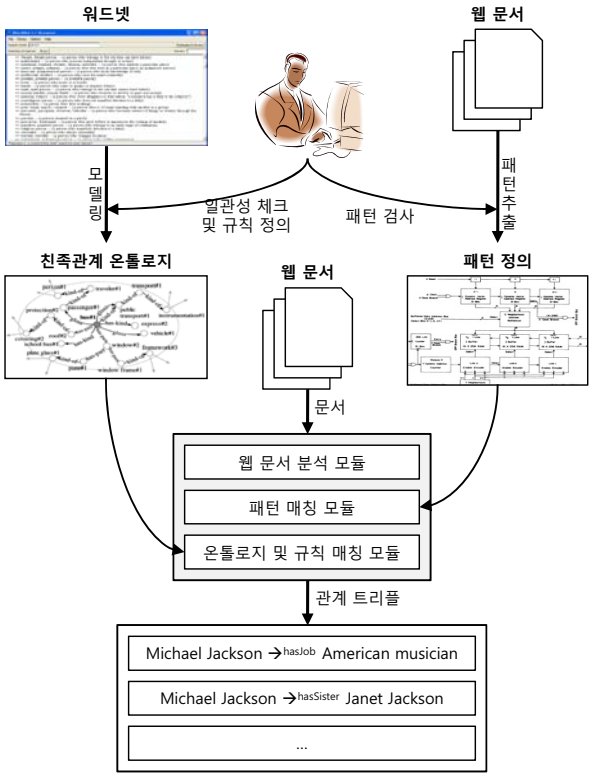
본 논문의 2장은 본 연구의 핵심 내용과 간단한 예제를 다루며, 3장에서 본 연구에 대한 결론과 향후 방향에 대해 기술한다.

2. 본론

다양한 웹 문서에 내포되어 있는 사람들의 친족관계 추출을 위한 전체 구조는 (그림 1)과 같다. 먼저 워드넷으로부터 친족관계를 정의하는 개념들을 추출하여 온톨로지를 모델링한다. 그리고 웹 문서에서 사람과 그에 대한 관계 정보를 추출하여 패턴을 정의한다. 이렇게 정의된 온톨로지와 패턴을 이용하여 다양한 웹 문서에 포함된 친족관계 정보들을 추출하며 트리플 형태로 저장한다. 본 장에서는 각 모듈의 내용에 대해 기술한다.

2-1. 친족관계 온톨로지 모델링

친족관계 온톨로지는 친척들 사이에 존재할 수 있는 관계 개념들을 정의하는 것이며, 그 구성요소로는 'father',



(그림 1) 전체 구조

'mother', 'uncle', 'grandfather', 'aunt', 'father-in-law' 등으로 형성될 수 있다. 워드넷은 이러한 개념들을 'relative#1'와 그 하위 개념으로 총 170개를 정의하고 있다. 워드넷은 일반적으로 사용되는 모든 영어 개념을 포함하고 있지만, 일부 개념들 사이의 관계구조에서 비일관성이 확인되었다. 예를 들어, 'great grandchild'는 'grandchild'의 하위 개념으로 정의되어 있다. 비록 전자가 후자의 자손이지만, 개념적으로 볼 때 둘 사이에는 상/

하위 관계가 형성되는 것은 잘못된 것이다. 이러한 비일관성을 체크하여 총 170개로 구성된 친족관계 온톨로지를 구축하였다. 이를 구성하는 개념들의 모임은 동일 온톨로지 내의 다른 하나의 개념으로 표현될 수 있는 특징이 있다. 예를 들어, B가 A의 parent임을 parent(A,B)로 표현할 때, 이 관계는 child(B,A), father(A,B), grandparent(C,B)^parent(C,A), sibling(A,C)^parent(C,B), mother(A,B) 로부터 각각 형성될 수 있다. 이와 같이, 온톨로지 내의 각 개념이 가질 수 있는 규칙들을 정의하였다. 친족관계 온톨로지 모델링과 각 개념이 가질 수 있는 규칙들은 웹 문서에서 추출된 친족관계를 바탕으로 관계를 확장할 수 있는 기반을 마련한다.

2-2. 패턴 정의

웹 문서에서 소셜 정보 및 친족 관계를 추출하기 위해서는 구문분석을 통해 추출된 고유명사가 사람에 대한 것인지부터 파악해야 한다. 이를 파악하는 다양한 연구들 중에, 우리는 [2]의 방법을 응용하여 적용한다. [2]의 경우는 전문용어의 정의구문을 자동으로 찾아주기 위해 서술패턴에 따른 스타일리스틱 필터(stylistic filter)와 도메인 확률에 따른 도메인 필터(domain filter)를 이용하였다. 본 연구에서는 전문용어의 서술패턴을 사람에 대한 것으로 축소하고, 여러 문서들의 분석을 통해 사람을 기술하는 패턴들을 정의하였다. 가장 대표적 형식은 "사람이름-서술패턴-직업군"이며, 서술패턴으로는 'is-a(n)', 'is-the', 'become a(n)' 등이 있다. 이러한 서술패턴 뒤에 나오는 직업군은 워드넷의 'person#1'의 하위에 있는 개념들 중에서 사람의 직업을 의미하는 것으로 'professor#1', 'doctor#1', 'socialist#1', 'musician#1' 등이 있다. 또한, 추출된 사람들 사이의 친족 관계를 파악하기 위해 2-1에 정의된 온톨로지의 개념들을 적용할 수 있는 패턴을 정의한다. 추출된 유형중의 대부분은 "사람-친족관계 서술패턴-사람(단독 또는 열거)"인 것으로 파악되었다. 예를 들면, "*Stanley Ann Dunham*, mother of *Barack Obama*"에서 'Barack Obama'와 'Stanley Ann Dunham' 사이의 관계를 파악할 수 있다.

2-3. 친족관계 추출

본 연구에서 모델링한 온톨로지와 패턴을 적용하여 친족관계를 추출하기 위해 위키피디아의 문서들을 분석하였으며, POS 태깅은 스탠포트 자연어처리 그룹에서 제작한 pos-tagger 버전 2008-09-28[1]을 사용하였다. 또한 태깅된 결과에서 복합명사를 판단 및 헤드(head) 추출 그리고 단어의 스태밍(stemming)을 파악하기 위해 [3]의 RUN (RUle based Noun term extractor)을 변형하여 적용하였다. 그 이후에 2-1과 2-2의 온톨로지와 패턴을 적용하여 친족관계 트리플을 추출하였다. <표 1>은 현재 미국 대통령인 "Barack Obama"의 친족에 관한 문서들을 분석하여 추출한 결과의 일부를 보이고 있다.

<표 1> 친족관계 추출결과 예제

주어	서술어(관계)	목적어
Stanley Ann Dunham	mother of	Barack Obama
Madelyn Lee Payne Dunham	grandmother of	Barack Obama
Stanley Ann Dunham	grandfather of	Barack Obama
Stanley Ann Dunham	son of	Ralph Waldo Emerson Dunham
Michelle Obama	married person	Barack Obama
Malia Ann	daughter of	Barack Obama
Natasha	daughter of	Barack Obama

<표 1>의 결과에서 2-1의 친족관계에 대한 규칙을 통해 추가적인 정보들도 추론할 수 있다 (예. Malia Ann is a granddaughter of Stanley Ann Dunham). 하지만, 아직 친족관계를 추출하는 패턴의 다양성이 부족하며, 추출된 결과의 필터링이 요구된다.

3. 결론 및 향후 연구

온톨로지 기반의 소셜정보 추출을 위한 전체적인 설계를 다루는 본 논문은, 온톨로지 모델링과 개념들의 규칙을 정의하고 웹 문서에서 사람과 관계 정보를 추출하기 위한 패턴을 파악하였다. 또한 이를 친족관계 추출에 적용하여 본 연구의 가능성을 살펴보았다. 하지만, 패턴의 다양성 및 필터링이 추가적으로 요구되었다. 이는 향후에 지속적인 연구로 이루어질 것이다.

감사의 글

이 논문은 2009년도 정부(교육과학기술부)의 재원으로 한국과학재단의 지원을 받아 수행된 연구임 (No.2009-0064749).

참고문헌

[1] P. Domingos, P. Mika, J. Golbeck, L. Ding, T. Finin, A. Joshi, A. Nowak, R. R. Vallacher, "Social Networks Applied," IEEE Intelligent Systems, 20(1), 80-93, 2005

[2] Velardi, P., Navigli, R., & D'Amadaio, P. "Mining the Web to Create Specialized Glossaries," in IEEE Intelligent Systems, 23(5), 18-25, 2008

[3] Hwang, M. G., Choi, C., and Kim, P.K. Automatic Enrichment of Semantic Relation Network and its Application to Word Sense Disambiguation, IEEE TKDE. (Submitted)

1) The Stanford Natural Language Processing Group: http://nlp.stanford.edu/software/tagger.shtml