단일 문서 기반의 인지적 지식구조 자동 생성 기법 제안 및 검증

김 형 우○ 이 문 용
한국과학기술원 지식서비스공학과
hw_kim@kaist.ac.kr, munyi@kaist.ac.kr

Proposing and Validating an Automated Method of Cognitive Knowledge Structure Creation from Single Documents

Hyungwoo Kim○ Mun Y. Yi
Dept. of Knowledge Service Engineering, KAIST

요 약

본 연구는 단일 문서로부터 문서가 내포하고 있는 지식정보를 지식구조 혹은 인지스키마로 불리는 형태로 자동 생성하는 기법을 제안한다. 제안된 기법을 이용하여 자동 생성된 지식구조는 실제 문서 학습자의 학습 전, 후의 지식구조, 문서의 해당 지식을 명확히 알고 있는 도메인 전문가의 지식구조와의 유사도 측정을 통해 검증하였다. 자동 생성된 지식구조는 학습자의 학습 후 지식구조, 전문가 지식구조와 상당한 유사성을 보이며, 문서의 지식 정보를 인지적인 관점에서 정교하게 표현 하고 있음을 확인하였다. 이는 기존의 단어 기반의 정보 기술들에서 더욱 고차원적인 지식 정보를 활용한 지식구조 기반 정보 기술의 연구 가능성을 제시한다.

1. 서 론

지식사회 또는 정보사회라고 명명되는 현 사회에서는 지식적 업무의 능력이 사회 생산성의 큰 관건이며 이러한 업무 능력의 핵심이 되는 지식정보는 끊임없이 쏟아져 나와 제타바이트(Zeta Byte) 시대가 도래하였다. 이러한 시대의 변화 속에 사람들의 지식 정보에 대한 욕구는 더욱 복잡하고 다양해졌지만, 정보 검색, 문서 추천 등의 기법은 단지 TF-IDF 기반의 핵심 단어 추출 기법에만 의존함으로써 사용자의 요구와 부합되지 않는 수많은 지식정보를 제공하여 정보 과부화(Information Overload)를 발생 시키고, 지식 습득 메커니즘의 효율성을 감소시키는 문제가 있다. 본 연구에서는 단일 문서로부터 문서가 내포하고 있는 지식 정보를 선언적 지식인 핵심 개념 추출 단계, 그리고 핵심 개념간의 관계 정보 추출 단계를 거쳐, 인지적 지식구조 형태로 자동 생성하는 기법을 제안한다. 또한, 생성된 지식구조를 실제 문서 학습자의 지식구조, 문서의 해당 도메인 전문가의 지식구조와 비교함으로써 제안된 기법의 유효성을 인지과학적인 관점에서 검증한다. 본 연구의 결과로써 생성된 지식구조는 기존 TF-IDF 정보로 추출된 문서의 대표 단어 정보보다 실제 사람이 인지하는 문서의 지식 정보에 근접하며, 이는 현재 정보 기술 분야의 한계 극복 및 발전 가능성을 제시한다.

2. 관련연구

2.1 지식구조

일반적으로 지식구조는 인간의 내적 표상을 집합적으로 의미하는 용어로 쓰인다. 어떤 사람이 하고자 하는 일을 커다란 맥락 안에서 이해되도록 표현하며, 사람들이 어떤 행동을 하는가에 대해서 근원적인 이유를 설명하는 모형이다. 보다 협의적으로는 어떤 특정 지식의 도메인에 속하는 핵심개념들 간의 관계를 일정한 형상으로 조직화한 모형을 의미한다[1]. 지식구조는 선언적 지식(Declarative Knowledge), 혹은 절차적 지식(Procedural Knowledge)과는 구별되며 [2] 지식구조가 사람들의 지식 표출화(Externalization) 과정을 통한 결과물의 형성에 중추적인 역할을 한다.

학습이 심화되어 갈수록 학습자의 지식구조 또한 정교하게 변화하는 특징을 갖고 있다. 지식구조는 때로는 멘탈모델 (Mental Model) 혹은 인지 스키마(Cognitive Schema)라고도 불린다. 이러한 관점에서 지식을 갖는다는 것은 단순히 개념, 단어, 규칙들을 알게 되는 것뿐만 아니라 그러한 요소들의 관계들을 정립하는 것을 의미하며, 학습은 학습도메인의 지식구조 및 멘탈 모델을 습득하며 변화시켜가는 일련의 과정으로 볼 수 있다.

Goldsmith & Johnson의 연구[3]에서는 사람으로부터 직접 개인의 지식구조를 추출하는 방법을 연구하였다. [3]의 연구에서는 지식 추출, 지식 표현, 평가의 세 단계로 개인의 지식구조를 측정, 검사한다. 지식 추출 단계에서는 개인에게 개념들 간의 모든 쌍에 대해서 관련성을 설문으로 측정 하고, 이를 패스파인더 알고리즘[4]을 이용하여 지식구조의 형태로 표현한다. 두 지식구조간의 근접성 (Closeness)[6]으로 유사도를

측정할 수 있으며, 해당 도메인의 전문가와 지식구조가 근접 할수록 좋은 학습 성과를 보였다[3]. 또한 단일 지식구조의 개념간 관계의 연관규칙을 이용하여 일관성(coherence)을 측정할 수 있으며[5] 이는 개인의 지식이 얼마나 정교하게 학습되었는지를 보여준다.

2.2 핵심 개념 추출

전통적인 자동 키워드 추출 기법은 크게 문서 콘텐츠의 내용을 이용하는 통계적 기법과 기계 학습 알고리즘 사용 기법으로 나눌 수 있다. 통계적 기법은 전통적이고, 보편적인 기법으로 정보검색 분야에서 사용되는 TF(Term Frequency), TF*IDF(Inverse Document Frequency)가 주로 사용된다.

단어에 가중치를 부여하는 연구[7]가 처음으로 제안되고 난 뒤에, 통계적 기법인 TF*IDF를 이용하여 단어가 들어있는 문서들 중에 비례적으로 빈도수가 높은 단어가 핵심어로 사용될 수 있는 방법[8], 키워드 자동 추출 분야의 전통적인 방법들이 제안되었다. 이후에도 기계 학습이나 문서 내의 단어 위치 등 여러 가지 기법을 이용한 핵심 키워드 추출 기법들이 연구되고 있다. 최근에는 활발한 오픈 프로젝트로 인하여, 전문분야 사전, 시소러스, OpenAPI 서비스 등의 키워드를 추출하고 문서 외부의 정보와 기존 기법을 융합하여 문서의 키워드를 추출하는 연구가 진행되고 있다. 단어 간 의미정보를 제공하는 WordNet, 시소러스의 정보를 이용하거나 전문가들에 의해 특정 도메인의 단어가 정리된 시소러스를 이용하여 키워드 추출의 성능을 향상시키는 방법 등이 제안되어왔다[9].

2.3 개념 간 관계 추출

단어 간 관계 정보 추출의 종류는 크게 의미관계(Semantic Relation) 추출과 연관관계(Association Relation) 추출로 분류할 수 있다. 의미관계란 명사구와 명사구를 연결하는 동사구의 의미적 정보를 말하고, 연관관계란 개념적으로 밀접한 관련이 있으나 동의어나 유사 동의어인 등가관계에 포함되지 않는 두 단어 간 의미적, 심리적 연관 정도를 나타낸다. 본 연구의 핵심 요소인 지식구조자동 생성을 위해서는 핵심 개념간의 관계를 설정하는 단어의 연관관계 정보 자동 추출 연구가 필요하다. 연관관계 추출 기술로서 전통적으로 단어의 공기정보(Co-occurrence)를 이용하는 방법이 있다. 공기정보란 두 단어가 동일한 문서, 문장, 구 등에 같이 발생하는 현상을 말하며, 더 자주 발생할수록 두 단어가 밀접한 관계를 가지고 있다는 전제에 기반하며[10], 키워드 추출의 시초가 된 Salton의 1989년 연구[7]에서 측정 방법이 제시 되었다. 이후 단어 간의 연관성을 구하기 위하여 WordNet과 Wikipedia와 같은 별도의 외부 시소러스 및 코퍼스를 이용하는 연구[11]가 진행되었지만, 이와 같은 방법으로는 미등록단어에 대한 정의가 불가능하고, 고유명사 및 신조어, 전문용어가 많이 사용된 문서를 대상으로 연관관계를 추출하기에는 어려움이 있다.

3. 단일 문서의 지식구조 자동 생성 과정

단일 문서로부터 지식구조를 추출하기 위해서는 첫째, 단일 문서의 핵심 개념 추출 단계, 둘째, 핵심 개념 간의 연관관계 추출 단계, 셋째, 핵심 개념과 관계를 이용한 지식구조 생성 단계로 총 세 단계의 과정을 필요로 한다.

3.1 핵심 개념 추출

단일 문서로부터 핵심 개념을 추출하기 위해서는 먼저 문서에 대한 형태소 분석이 필요하다. 이를 위해 본 연구에서는 루씬 한글 분석기 오픈소스 프로젝트로 개발 된 KoreanAnalyzer[12]를 사용하여 문서의 단어 중 명사만을 추출 한다. 추출 된 명사들 중 자주 나오지만 해당 문서의 핵심 개념이 아닌 경우가 존재하게 된다. 보통 TF*IDF를 이용해 그러한 단어의 가중치를 낮추지만, 단일 문서인 경우에는 IDF 정보를 사용할 수 없다. 핵심 개념으로 볼 수 없는 명사의 불용어 처리를 위해 오픈 사전인 한국어 위키백과[13]를 사용하였다. 한국어 위키백과는 모든 방문자가 적극적으로 편집 가능한 웹 기반의 백과사전으로 다양한 도메인에 대하여 약 28만개의 개념들을 저장하고 있다. 이와 같은 외부 코퍼스의 단어와 매칭되는 명사 중 빈도수 별 상위 N개를 해당 문서의 핵심 개념으로 추출한다. 또한, 한 개 이상의 명사로 구성되는 전문용어(예, 트로이 목마) 인식을 위하여, 연속되는 명사의 전문용어 매칭도 실시하였다.

3.2 개념 간 연관관계 추출

문서의 핵심 개념간 연관관계를 추출하기 위해서, 단어 쌍의 공기정보(Co-occurrence)를 이용한다. 본 연구에서는 공기정보를 두 개념이 같은 문장에서 동시 출현하는 빈도수인 문장 공기정보와 두 개념이 같은 문단에서 동시 출현하는 문단 공기정보로 세분화한다. 다음은 단순 공기정보를 이용하여, 개념간 연관관계 유사도를 측정하는 공식이다. 식 (1)은 문장 간 공기정보를 이용하여 구한 단어간 유사도(Sentence co-occurrences Similarity: SS), 식 (2)는 문단간 공기정보를 이용하여 구한 단어간 유사도 (Paragraph co-occurrences Similarity: PS)를 구하기 위한 공식이다.

$$SS_{ij} = \frac{\{\sum_{1}^{N_S} n(W_i \cap W_j)\}}{Max(C_S)}, \quad (0 \le SS \le 1) \qquad (1)$$

$$PS_{ij} = \frac{\{\sum_{1}^{N_P} n(W_i \cap W_j)\}}{Max(C_P)}, \quad (0 \le PS \le 1) \qquad (2)$$

위의 식에서 N_S와 N_P는 각각, 문서에 나타난 순서에 따른 문장 번호, 문단 번호가 된다. 단어 W_i와 단어 W_j의 유사도는 각 문장 혹은 각 문단에서 동시 출현한 횟수를 총 더한 것을, 문서에서 나타난 각 문서, 문단 공기정보의 최대 값으로 나누어 0과 1사이의 값으로 정규화시킨다.

위의 식으로 공기정보를 이용하여 쉽게 단어 간 유사도를 측정할 수 있지만, 이 방법은 많이 출현한 단어일수록 다른 단어들과 유사관계가 높아지는 문제점을 갖는다. 이러한 문제점을 해결하기 위해 문서 군집화에 널리 쓰이는 코사인 유사도 측정 방법을 변형하여 사용한다.

표 1 문장 번호를 이용한 ISV의 예

	문장 1	문장 2	문장 3	…	문장 N
W_i	3	0	1	…	1
W_j	2	1	0	…	2

표 1에서 처럼 각 문장에 출현하는 개념의 빈도수로 이루어진 ISV(Inverted Senteces Vetcor)를 생성한다.

$$SCS_{ij} = \frac{v_i \cdot v_j}{||v_i|| \times ||v_j||}, \quad (0 \le SCS \le 1) \qquad (3)$$

그 후 식 (3)를 이용하여 단일 문서로부터 각 개념간의 코사인 유사도(Sentence co-occurrences Cosine Similarity: SCS)를 측정할 수 있다.

표 1의 문장 번호를 문단 번호로 바꾸어 동일한 방식으로 개념간의 코사인 유사도(Paragraph co-occurrences Cosine Similarity: PSC)를 측정할 수 있으며, 위의 방식은 단어가 출현한 빈도수에 상관없이 동시 출현한 정도에 따라 유사도가 측정되므로 단일 문서 안에서의 개념 간 연관관계 측정에 적합하다.

3.3 연관관계 정보로부터 지식구조 생성 과정

기존 인지심리학 분야의 지식구조 생성과정에서 주로 사용하는 방법과 동일하게, 식 (4)를 이용하여 각 개념 간 연관관계를 7점 스케일로 변환하였다. (1:매우 관련 있음, 7:전혀 관련 없음)

$$D_{ij} = 7 - S_{ij} \times 6, \quad (1 \le D_{ij} \le 7) \qquad (4)$$

그 후, 각 개념 간의 연관관계 정보로 이루어진 유사도 측정 테이블을 작성하고, 여기에 패스파인더 알고리즘[4]을 적용하여, 각 개념 간을 최단 거리로 연결하여 주는 지식구조를 자동 생성하였다.

4. 실험 및 결과

네이버캐스트[13] 문서를 실험 문서로 사용하여, 본 연구의 결과를 검증하였다. 네이버캐스트는 문화, 생활과학, 인물 등 각 분야의 전문 기자 혹은 교수들이 작성한 총 5,707개의 전문 정보를 보유하고 있으며, 이 중 생물 분야의 문서를 실험문서로 사용하였다.

단일 문서로부터 자동 생성된 지식구조를 검증하기 위하여 학습 실험자 21명(나이:25~32)을 대학교 안에서 모집하였다. 실험자가 문서를 학습하기 전에 자동 추출된 해당 문서의 핵심 개념 11개의 모든 쌍에 대해서 7점 스케일(1:전혀 관련 없음, 7:매우 관련 있음)로 설문을 하였다. 그 후 해당 문서를 학습하게 하였고, 학습 효과를 높이기 위해 충분한 학습시간을 주었다. 학습이 끝난 후 다시 한번 핵심 개념의 모든 쌍에 대한 관련 정도를 측정하게 하였다. 설문 정보로 각 실험자의 해당 문서 학습 전, 학습 후의 지식구조를 기존 인지심리학 분야에서 검증된 방법[3]으로 생성하였다. 또한, 같은 대학의 생명과학과 대학원생 3명을 모집하여 해당 문서의 전문가 지식구조를 생성하였다. 3명의 전문가가 지식구조를 추출하는 과정에서 의견을 공유하기도 했으며, 3명의 개념간 관련 점수의 평균 값으로 전문가 지식구조를 생성하였다.

표 2 지식구조 간의 유사도(Closeness) 수치

	학습 후	전문가	PCS	SCS	PS	SS
학습 전	**0.31**	**0.29**	0.32	0.26	0.33	0.30
학습 후	-	**0.54**	**0.49**	0.42	0.45	0.43
전문가	-	-	**0.67**	0.48	0.57	0.48

표 2의 지식구조 간의 유사도(Closeness) 수치[6]는 그 값이 0(두 지식구조는 상이함)에서 1(두 지식구조는 동일함)까지 변할 수 있다. 표 2에서 각 학습자의 학습 전, 학습 후 지식구조 간 유사도의 평균은 0.31로 낮으며 이를 통해 학습을 통하여 학습자의 지식구조가 변한 것을 확인할 수 있다. 또 한, 학습 전에 비해 학습 후의 지식구조가 전문가의 지식구조와 더욱 유사해졌다는 것을 알 수 있다 (학습 전 유사도: 0.29, 학습 후 유사도: 0.54). 이는 학습이 제대로 되었음을 알려준다. 또한, 문단 공기정보를 이용하여 코사인 유사도 기법으로 문서의 연관관계를 측정하는 PCS를 활용한 결과가 학습자의 학습 후 지식구조와 가장 유사한 성능을 보였다 (유사도: 0.49). 이 값은 전문가들의 지식구조와 학습자의 학습 지식구조를 비교한 값에 근접하다 (유사도: 0.54). 또한 PCS 방법은 비교한 4가지 방법 중에 가장 전문가의

지식구조에 가까운 값을 보여준다 (유사도: 0.67).

표 3 지식구조들의 일관성(Coherence) 수치

학습 전	학습 후	PCS	SCS	PS	SS	전문가
0.24	0.66	0.89	0.73	0.72	0.66	0.73

표 3에서 보여지는 일관성(Coherence)은 각 개념들 간의 관계를 얼마나 일관되게 파악하고 있는지를 보여준다. 예를 들면 A와 B의 관계를 가깝다고 생각하고 B와 C의 관계를 가깝다고 생각한다면 A와 C의 관계 역시 가깝다고 생각해야 일관성이 있게 된다. 표 3에서 학습자의 학습 전 지식구조는 일관성이 부족하지만, 학습 후의 지식구조는 일관성이 현저하게 향상되었음을 보여준다 (학습 전: 0.24, 학습 후: 0.66). 이는 학습이 제대로 이루어졌음을 다시 한번 보여준다. 자동 생성된 지식구조들은 전문가의 지식구조만큼 정교한 형태로 생성되었으며, 특히 PCS 정보로 생성된 지식구조는 전문가의 지식구조 보다 더 일관성을 갖고 있다.

추가적으로, 전문가들 간의 지식구조의 유사도 범위는 0.65~0.73이고, PCS로 자동 생성한 지식구조와 전문자 지식구조 간의 유사도 범위는 0.50~0.68로 유사도가 상당이 근접하고 겹쳐지는 부분도 존재한다. 이는 자동 생성된 지식구조가 마치 또 다른 한 명의 전문가가 생성한 지식구조처럼 이용될 수 있음을 보여준다.

5. 결론 및 향후 연구

본 논문에서는 단일 문서로부터 인지적 지식구조를 자동 생성하는 기법을 제안하였다. 기존의 문서로부터 정보를 추출하는 연구는 기술적인 부분에만 집중하여, 실제 문서 학습을 통해 사람이 얻을 수 있는 지식 정보의 인지적 관점은 간과하였다.

문단 단위에 나타난 개념 간 연관관계 정보를 통해 자동 생성한 지식구조는 실제로 문서로 표현된 지식을 학습한 학습자와 문서의 지식을 정교하게 알고 있는 전문가들의 지식구조와 상당한 유사성을 보이며, 마치 제 3의 전문가로부터 생성한 지식구조와 비슷한 양상을 보였다. 이는 단일 문서로부터 문서가 가지고 있는 인지적 지식정보를 자동으로 생성해 주는 방법의 유효성을 입증한다. 이처럼 자동으로 생성된 지식 정보는 문서추천, 정보검색, 학습 가이드 등 광범위한 분야에 적용될 수 있으며, 기존 단어 기반의 기술에서 단어와 관계를 포함하는 지식구조 기반의 기술로 패러다임을 전환하는 기회가 될 것으로 생각된다.

향후 연구로서는 현재의 단일 문서 기반의 지식구조 자동 생성 기법을 여러 문서들을 활용함으로 도메인 지식구조를 자동 생성하는 기술로 확장 발전 시켜야 할 필요가 있으며, 이러한 지식구조 정보를 적용할 수 있는 다양한 후속 연구가 계속되어야 한다. 그러한 향후 연구를 위해 현재의 연구는 중요한 초석이 된다.

참고문헌

[1] Day, E. A., Arthur, W. J., & Gettman, D.,"Knowledge structures and the acquisition of a complex skill", Journal of applied psychology, vol.86, no.5, 2001.

[2] Davis, F. D., & Yi, M. Y., "Improving Computer Skill Training: Behavior Modeling, Symbolic Mental Rehearsal, and the Role of Knowledge Structures", Journal of applied psychology, vol.89, no.3, 2004.

[3] Timonthy E. Goldsmith, Peder J. Johnson, and William H. Acton, "Assessing Structural Knowledge", Journal of Educational Psychology, vol.83, no.1, pp.88-96, 1991.

[4] Schvaneveldt, R. W., Durso, F.T., and Dearholt, D. W., "Network structures in proximity data", The psychology of learning and motivation , vol. 24, pp.249-294, 1989.

[5] Goldsmith, T., & Kraiger, K., "Structural knowledge a assessment and training evaluation", In J. K. Ford, S. W. J.Kozlowski, K. Kraiger, E. Salas, & M. S. Teachout (Eds.), Improving training effectiveness in work organizations, pp. 73–96, 1997.

[6] Goldsmith, T. E., & Davenport, D.M., "Assessing structural similarity of graphs. In R. W. Schavneveldt (Ed.), Pathfinder associative networks: Studies in knowledge organization (pp.75-87), 1990.

[7] Salton, G., "Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer",. Addison-Wesley, 1989.

[8] Salton, G., "Developments in automatic text retrieval.", Science 253, pp.974 –980. 1991.

[9] A. Hulth, J. Karlgren, A. Jonsson, H. Bostrom, and L. Asker, "Automatic keyword extraction using domain knowledge", Computational Linguistics and Intelligent Text Processing, 2004

[10] Van Rijsbergen, C.J., "A theoretical basis for the use of co-occurrence data in information retrieval", Journal of Documentation, vol.33, no.2, pp.106-119, 1997

[11] F.M. Suchanek, G. Kasneci, and G. Weikum, "Yago: a core of semantic knowledge—unifying WordNet and Wikipedia", in Proc. of the 16th International Conference on World Wide Web, WWW2006, Banff, Canada, 2007

[12] http://cafe.naver.com/korlucene

[13] http://ko.wikipedia.org

[14] http://navercast.naver.com