동시출현 단어 분석을 통한 지식 구조의 파악 : 인공지능 분야를 대상으로

Exploration of Intellectual Structure of Artificial Intelligence Field Using Co-word Analysis

이미경, 정영미, 연세대학교 대학원 문헌정보학과

Mi-kyoung Lee, Young-Mee Chung
Dept. of Library and Information Science, Graduate school of Yonsei University

이 연구에서는 통제된 색인어를 이용하여 파악한 지식 구조와 통제되지 않은 키워드를 이용한 지식 구조를 비교하여 두 구조가 어떤 차이점을 보이는지를 살펴보았다. 또한 색인효과가 어떻게 나타나는지, 비통제어를 사용한 경우가 실제적으로 더 상세한 하위 영역을 표현하는지를 확인하고자 하였다. 실험 결과 통제된 색인어인 주제명표목을 사용한 영역지도와 비통제 색인어인 키워드를 사용한 영역지도 둘 다 인공지능 분야의 주요 분야들을 비슷하게 나타냈지만, 주제명표목을 사용한 경우에 색인효과가 일부 나타났다. 그리고 대체적으로 주제명표목에 기반한 영역지도보다는 키워드에 기반한 영역지도가 더 상세하게 나타났다.

1 서 론

계량정보학에서는 동시인용 분석(co-citation analysis)이나 동시출현 단어 분석(co-word analysis)을 이용하여 특정한 주제 분야의 영역 구조를 파악할 수 있다.

동시출현 단어 분석은 해당 주제 분야의 문헌 집합에서 키워드나 분류코드 등을 추출하여 각 단어 쌍의 동시출현빈도(co-occurrence)를 계산한 다음 동시출현빈도를 그대로 사용하거나, 다양한 지수들을 이용해 단어간의 연관도를 구하여 하위 영역들을 매핑하는 것으로 주제 분야의 영역을 시각적으로 표현하는 방법이다.

동시출현 단어 분석은 실험집단 및 사용되는 단어의 유형, 단어들의 정규화 및 연관도 측정 방법, 하위 영역의 매핑 방법 등 다양한 요소에 따라 분석 결과의 질이 달라지기 때문에 적절한 데이터와 실험방법을 결정하는 것이 필요하다.

연구자들은 실험 집단의 범위와 데이터양의 불완전성, 통계적 방법의 적절성과 같은 기술적인 문제와 결과 해석의 어려움 등 동시출현 단어 분석의 몇 가지 문제점들을 제기하였는데, 그 중에서도 동시출현 단어 분석의 가장 중요한 문제는 색인효과

(index effect)라고 하였다.

색인효과는 색인전문가가 통제어를 사용하여 문헌을 색인할 경우에 문헌에서 실제 사용되고 있는 용어들과 다르게 색인어를 부여함으로서 문헌의 주제가 다르게 표현되는 것을 말한다.

따라서 본 연구의 목적은 통제된 색인어인 주제명표목(subject heading)을 이용한 지식 구조와 통제되지 않은 키워드(key phrase identifier)를 이용한 지식 구조를 비교하여 둘의 구조가 어떤 차이점을 보이는지를 살펴보고, 또한 색인효과(index effect)가 어떻게 나타나는지, 비통제어를 사용한 경우가 실제적으로 더 상세한 영역을 표현하는지를 살펴보는 것이다.

이 논문에서는 단어들을 먼저 클러스터링한 후 클러스터들을 다차원축척 지도에 표현하는 접근법(Noyons, and van Raan 1998)을 사용하였는데, 클러스터링을 사용하지 않고 단어들을 직접적으로 다차원축척 지도에 표현한 경우에도 단어들의 그룹이 형성될 수 있지만 클러스터링의 형식을 사용하면 보다 이해하기 쉬운 영역지도가 될 수 있기 때문이다(Peters, and van Raan 1993). 그리고 클러스터링을 이용한 영역지도와 단어들을 직접적으로 이용한 영역지도의 일관성을 예비 실험한 결과에서도 클러스터링을 이용한 방법이 주제 분야의 지식 구조를 왜곡하지 않는다는 것을 확인하였다(이미경, 2003).

2 실험집단 및 실험내용

실험집단 및 단어 유형 선정은 다음과 같이 하였다. 동시출현 단어 분석에서는 적절한 단어 선정이 중요한 요소이기 때문에 저널 단위로 데이터를 선택하기 보다는 문헌 단위로 데이터를 선정하는 것이 바람직하다. 따라서 본 실험에서는 INSPEC 데이터 베이스에서 수집한 1997-2000년에 출판된 인공지능 분야의 영문으로 된 논문 4,555건을 대상으로 실험하였으며, 특정 분류코드와 주제명표목을 부여받은 문헌들을 선정한 후에 주제명표목과 키워드 필드에 출현한 단어들을 각각 추출하여 2개의 실험집단을 만들어 실험을 진행하였다.

우선 'Artificial intelligence'와 관련된 주제명표목으로는 INSPEC 시소러스 중 'Artificial intelligence'의 1단계 하위어(NT)까지 사용하였다. 실험에 사용된 주제명표목은 'Adaptive resonance theory', 'Artificial life', 'Cooperative systems', Generalisation(artificial intelligence)', 'Fuzzy control', 'Knowledge engineering', 'Learning(artificial intelligence)', 'Perceptrons', 'Planning(artificial intelligence)', 'Uncertainty handling' 등 모두 11개다.

그리고 이렇게 수집된 7,221건의 데이터 중에서 Artificial intelligence[C1230]과 하위항목인 Learning in AI[C1230L], Neural nets[C1230D], Reasoning and inference in AI[C1230R] 등 인공지능 관련 분류코드가 부여된 경우로 데이터를 제한하여 최종적으로 4,555개의 문헌을 선택하였다.

키워드에는 구 뿐 아니라 단일어도 포함되어 있다. 실험을 위해 우선 비통제 키워

드에 대해서는 단/복수, 약어, 대시(-), 철자, 띄어쓰기, 품사의 형태 변형 등 단어 통제가 이루어졌다.

주제명표목과 키워드 실험집단을 각각 위와 같이 비통제 키워드의 사전처리 후 두 실험집단에 속한 단어들의 출현빈도를 구하여 이 중 출현빈도가 일정 기준 이상인 단어들을 선택하여 실험하였다. 출현빈도의 기준은 클러스터링 성능과 SPSS 프로그램의 데이터 제한을 고려하여 결정하였는데, 주제명표목은 출현빈도 28 이상의 단어 150개, 키워드는 출현빈도 30 이상인 단어 163개를 선정하였다.

출현 빈도 기준이 너무 높을 경우에는 세부 분야들이 나타나기 어렵고, 너무 낮을 경우에는 클러스터 내에 잡음이 많아지므로 적절한 빈도 수 선정이 필요하다.

선정된 단어들은 단어간의 동시출현빈도를 구한 후, 코사인 유사도에 의해 정규화된 단어-단어 행렬로 변환하고 계층적 클러스터링(hierarchical clustering) 기법인 Ward 방법을 사용하여 클러스터링한다. 그리고 클러스터링 결과 나타난 클러스터들을 각각 다차원축척 기법을 통해 2차원으로 표현하였다.

두 집단에 대해 각각 다차원축척 표현까지 실험한 후에 주제명표목을 사용한 영역 구조와 키워드를 사용한 영역 구조를 비교 분석하였다. 다음은 이 연구에서 동시출현 단어들간의 연관도 측정을 위해 사용한 코사인 유사계수 공식이다.

$$Sim(x,y)=\frac{C_{xy}}{\sqrt{C_x C_y}}$$

Cx : 문헌 내 x의 출현빈도
Cy : 문헌 내 y의 출현빈도
Cxy: 문헌 내 x와 y의 동시 출현빈도

3 실험 결과

3.1 주제명표목의 실험 결과

주제명표목의 동시출현빈도를 코사인계수로 정규화한 150×150 행렬을 Ward 방법을 사용하여 클러스터링을 하였다. 주제명표목의 경우에는 클러스터 수를 늘려도 세부적인 분야를 형성하지 못하였기 때문에 12개의 클러스터를 생성하였다.

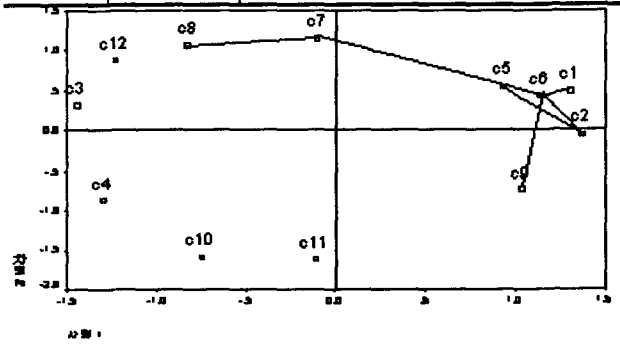
주제명표목의 클러스터링 결과는 〈표 1〉과 같다. 각 클러스터에서 출현빈도가 높은 표목들을 주요표목으로 제시하였다. 클러스터의 크기는 클러스터에 속한 표목들의 출현빈도가 전체 클러스터에서 차지하는 비율을 나타낸 것이다.

〈그림 1〉은 주제명표목 분석에서 클러스터간의 관계를 다차원축척 지도에 2차원으로 표현하고, 0.3 이상의 코사인계수 값을 가진 클러스터들을 다차원축척 지도에 선으로 표현한 것이다

3.2 키워드의 실험 결과

〈표 1〉 클러스터별 주요 주제명표목

클러스터	클러스터명	주요 주제명표목	크기
C1	학습의 최적화 방법	Computational complexity, Pattern recognition, Adaptive systems, Parameter estimation, Stochastic processes, Interpolation, Simulation, Performance evaluation, Maximum likelihood estimation, Minimisation Least squares approximations, Convergence of numerical methods, Iterative methods	11%
C2	제어	Fuzzy control, Neurocontrollers, Adaptive control, Control system synthesis, Nonlinear control systems, Robust control, Optimal control, Control system analysis	9%
C3	동적 계산	Markov processes,State-space methods, Approximation theory, Dynamic programming	2%
C4	인지모델	Artificial intelligence, Psychology, Neurophysiology, Brain models, Philosophical aspects	4%
C5	지능형 에이전트 시스템	Robots, Artificial life, Evolutionary computation, Genetic algorithms, Software agents, Cooperative systems, Intelligence control, Self-adjusting systems, Mobile robots, Navigation, Path planning, Robot vision	12%
C6	신경망 기반 학습	Learning (artificial intelligence), Neural nets, Optimisation, Generalisation(artificial intelligence), Feedforward neural nets, Statistical analysis, Backpropagation, Neural net architecture	31%
C7	추론 및 퍼지모델	Uncertainty handling, Fuzzy logic, Inference mechanisms, Fuzzy set theory, Probability, Bayes methods	13%
C8	지식공학	Knowledge representation, Knowledge based systems, Data mining, Knowledge acquisition, Expert systems	6%
C9	예측 및 동적 시스템	Convergence, Identification, Time series, Prediction theory, Nonlinear dynamical systems, Chaos, Modelling, Nonlinear systems	4%
C10	자연어처리	Natural languages, Speech recognition, Computational linguistics, Hidden Markov models	1%
C11	컴퓨터비전	Computer vision, Object recognition, Image classification, Image recognition, Image segmentation	2%
C12	계획 및 문제해결	Planning (artificial intelligence), Search problems, Problem solving, Heuristic programming, Explanation	74%



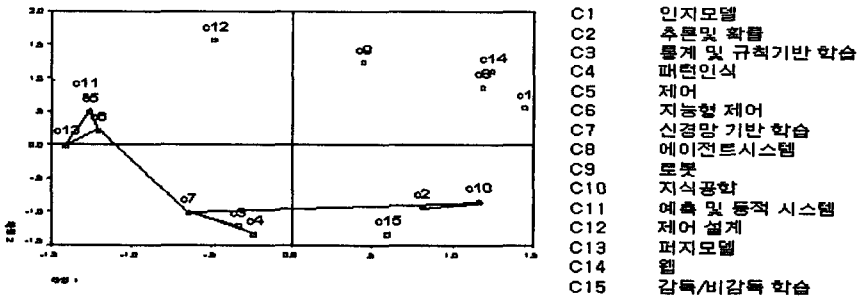
C1 학습의 최적화 방법
C2 제어
C3 동적 계산
C4 인지모델
C5 지능형 에이전트 시스템
C6 신경망 기반 학습
C7 추론 및 퍼지모델
C8 지식공학
C9 예측 및 동적 시스템
C10 자연어처리
C11 컴퓨터비전
C12 계획 및 문제해결

연결선: 코싸인계수 〉 0.2

〈그림 1〉 주제명표목의 다차원축척 지도

〈표 2〉 클러스터별 주요 키워드

클러스터	클러스터명	주요 키워드	크기
C1	인지모델	AI, artificial life, autonomous agent, cognitive science, consciousness, psychology	5%
C2	추론 및 확률	uncertainty, probability, inference, uncertainty handling, knowledge representation, reasoning, approximate reasoning, bayesian method, information theory, Dempster Shafer theory, possibility, incomplete information	7%
C3	통계 및 규칙기반 학습	training data, classification, statistical analysis, clustering, neural network traing, EM algorithm, dataset, incremental learning, reinforcement learning, decision tree, heuristic, computational complexity, rule base, complexity, pruning, bayesian network, mathematical analysis, state space	13%
C4	패턴인식	pattern recognition, pattern classification, feature extraction, character recognition, computer vision, object recognition	4%
C5	제어	fuzzy control, stability, adaptive control, position control, neurocontrol, optimal control,robust control, robot manipulation, PID control, sliding mode control	11%
C6	지능형 제어	genetic algorithm, optimization, computer simulation, intelligent control, soft computing, tuning, adaptive fuzzy control, control system, nonlinearity	6%
C7	신경망 기반 학습	learning, neural network, generalization, feedforward neural network, backpropagation, multilayer perceptron, radial basis function neural network, multilayer neural network, function approximation, iteration, adaptive learning	26%
C8	에이전트 시스템	multiagent, agent, software agent, intelligent agent, distributed AI	2%
C9	로봇	mobile robot, navigation, multirobot, autonomous robot, cooperative system	2%
C10	지식공학	expert system, decision making, knowledge acquisition, problem solving, planning, knowledge based system, intelligent system, domain knowledge, decision support system, case study, case based reasoning, data mining, knowledge discovery	7%
C11	예측 및 동적 시스템	identification, dynamic system, gradient descent, prediction, time series, nonlinear dynamic system, complex system	3%
C12	제어 설계	control, control design, control simulation	2%
C13	퍼지모델	fuzzy logic, fuzzy set theory, fuzzy rule, membership function, fuzzy model, fuzzy neural network, fuzzy inference, fuzzy system, fuzzy reasoning	9%
C14	웹	internet, web, information retrieval	1%
C15	감독/비감독 학습	supervised learning, unsupervised learning	1%



연결선: 코싸인계수 〉 0.2

〈그림 2〉 키워드 다차원축척 지도

키워드의 동시출현빈도를 코사인계수로 정규화한 163×163 행렬을 Ward 방법을 사용하여 클러스터링을 하였고, 15개의 클러스터를 선정하였다. 키워드의 클러스터링 결과는 〈표 2〉와 같다. 각 클러스터에서 출현빈도가 높은 키워드들을 주요 키워드로 제시하였다.

〈그림 2〉는 키워드 클러스터간의 관계를 다차원축척 지도에 2차원으로 표현하고, 0.2 이상의 코사인계수 값을 가진 클러스터들을 다차원축척 지도에 선으로 표현한 것이다.

4 주제명표목과 키워드 분석 결과 비교

키워드를 사용한 영역지도가 주제명표목을 사용한 영역지도에 비해 같은 분야라도 더 상세한 하위 영역을 형성하는 것을 '학습', '에이전트', '제어' 등의 분야를 통해 알 수 있었다. 하지만 주제명표목을 사용한 경우가 항상 일반적인 영역을 형성하는 것은 아니었다.

키워드 실험에서는 주제명표목 실험에서 나타난 '자연어처리'나 '컴퓨터비전'과 같은 하위 분야가 독립된 클러스터로 나타나지 않았다. '컴퓨터비전'은 '패턴인식'에 포함되어 상위 영역으로 흡수되어 나타났고. '자연어처리'는 natural language processing과 같은 단어보다 더 세분화된 단어들이 사용되었기 때문에 출현빈도에서 낮게 나타나 독립된 클러스터를 형성하지 못했다.

따라서 키워드와 주제명표목의 클러스터 경향을 살펴보면 키워드 클러스터링은 비교적 균등한 영역을 형성한 반면 주제명표목을 사용한 클러스터링은 더 일반적인 영역과 상세한 영역을 함께 형성한다는 것을 알 수 있다.

그리고 두 실험 집단의 주제 영역은 전체적으로는 비슷한 양상을 보여주지만 주제명표목을 사용한 경우에 색인효과가 나타나는 부분들이 있었다. '퍼지모델'의 경우 주제명표목에서 '확률', '불확실성 처리(uncertainty handling)', '추론(inference)' 등과 함께 클러스터를 형성했던 것과 비교해 키워드에서는 독립적인 클러스터를 형성하고 있으며, 키워드에서 클러스터 수의 기준을 적게 할 경우에는 '지능형 제어' 클러스터에 속하게 된다. 영역지도 분석에서도 주제명표목을 사용한 영역지도의 경우 퍼지모델이 '지식공학(C8)'과 함께 나타났지만, 키워드를 사용한 영역지도에서는 '지능형 제어'(C6)와 '제어(C5)' 등과 함께 위치한다. 이것은 '퍼지모델'이 이론상 '불확실성 처리', '추론'에 사용될 수 있지만 실제적으로 주로 '제어' 분야에 응용되어 사용되기 때문에 나타난 결과로 보인다.

주제명표목을 사용할 경우 비통제어가 나타내기 어려운 개념들을 나타낼 수 있고, 적은 실험집단을 사용할 경우에도 영역지도의 일관성을 줄 수 있다는 장점이 있다. 실험집단의 양을 다양하게 조절하여 예비실험한 결과 주제명표목을 사용한 경우에는 실험집단의 수가 적더라도 영역지도가 비교적 일관성이 있게 나타나는 반면 키워드의 경우에는 실험집단에 따라 형성된 하

위 영역들간의 일관성이 떨어졌다. 하지만 키워드를 사용한 경우가 클러스터간의 수준이 비슷하고, 클러스터를 구성하는 개념들의 유사성이 더 높았다.

5 결 론

결론적으로 주제명표목을 사용한 영역지도와 키워드를 사용한 영역지도가 인공지능 분야의 주요 하위 영역들을 비슷하게 나타내고 있지만, 주제명표목을 사용한 영역지도의 경우에는 색인효과가 나타날 수도 있다는 것을 알 수 있었다.

또한 대체적으로 주제명표목을 사용하여 클러스터링한 하위 영역보다는 키워드를 사용한 하위 영역이 세분화되어서 나타났다. 하지만 통제어가 일반 단어보다 더 넓은 개념을 가졌다고 하여 항상 더 일반적인 범위의 하위 영역을 형성하는 것은 아니다. 실험 결과 키워드 분석에서 나타나지 않는 상세한 하위 영역들이 주제명표목을 사용한 경우에 나타나고 있다.

그리고 키워드를 사용한 경우가 클러스터 수준이 비슷하고, 클러스터를 구성하는 개념들의 유사성이 더 높았다. 따라서 동시출현 단어 분석에서 클러스터링 방법을 함께 사용할 경우에 비통제어를 사용하는 것이 보다 명확한 하위 영역을 형성하는 데 유리할 것으로 보인다.

참 고 문 헌

이미경. 2003. **동시출현 단어 분석을 통한 지식 구조 파악에 관한 연구: 인공지능 분야를 대상으로**. 석사학위 논문, 연세대학교 대학원.

Law, J., S.Bauin, J.P. Courtial, and J. Whittaker.

1988. Policy and the mapping of scientific chance: a co-word analysis of research into environmental acidification. *Scientometrics*, 14: 251-26

Noyons, E.C.M., and A.F.J. van Raan. 1998. Monitoring scientific developments from a dynamic perspective: self-organized structuring to map neural network research. *Journal of the American Society for information Science*, 49(1): 68-81

Peters, H.P.F., and A.F.J. van Raan. 1993. Co-word based science maps of chemical engineering, Part2: Combined clustering and multidimensionalscaling. *Research Policy*, 22: 47-71