학술DB에서 SNA(Social Network Analysis) 기법을 이용한 연관검색어 제공방안 연구

김경용[O*], 서정연[**], 선충녕[**]
[*]서강대학교 정보통신대학원, [**]서강대학교 컴퓨터공학과
[*]kky55555[O]@sogang.ac.kr, [**]{seojy , wilowisp}@sogang.ac.kr

A Study on Providing Relative Keyword using The Social Network Analysis Technique in Academic Database

Kyoung-Yong Kim[O*], Jung-Yun Seo[**], Choong-Nyoung Seon[**]
[*]Graduate School of Information & Technology, Sogang University
[**]Department of Computer Science and Engineering, Sogang University

요 약

본 논문은 다양한 주제 분야의 연구 성과물을 제공하는 학술DB에서 주제어(Keyword) 정보를 바탕으로 SNA(Social Network Analysis)기법을 적용해 검색어와 연관도가 높은 연관검색어를 제공하는 것을 그 목적으로 한다. 이를 위해 주제어들 간의 가중치(Weight)를 계산한 뒤 Ego Network 분석을 통해 검색어와 연관된 연관주제어를 추출하고 이를 기존 학술DB에서 제공한 연관검색어와 비교 정리하였다. 그리고 정리된 결과를 연관규칙 마이닝기법, 유사계수를 적용해 연관도측면에서 비교 평가하였다.

주제어: 연관검색어, 사회연결망분석, SNA(Social Network Analysis)

1. 서론

현재 Web과 DB의 발달로 인해 정보량은 지속적으로 증가되고 있으며, 이에 따라 효율적인 정보검색을 위한 방법으로 연관검색어의 역할이 커지고 있다.

하지만 연구 성과물의 집약체라고 일컫는 학술DB에서의 연관검색어는 그동안 기계적인 방식이나 이용자들의 추천방식에 의해서 정제되지 않은 채 제공되어져 왔다. 이에 따라 융·복합 연구가 활발히 진행되고 있는 이 시점에 연구 분야들 간의 상호작용 및 흐름[1]을 반영한 연관검색어를 제공할 필요성이 제기된다.

이를 위해 본 논문에서는 학술 DB의 주제어(Keyword)들의 가중치(Weight) 계산을 통해 SNA(Social Network Analysis)[2] 기법 중의 하나인 Ego Network를 도출한 뒤 검색어와 연관된 연관주제어를 연관검색어로 제공하는 모델을 제시한다.

2. SNA 기법을 이용한 연관검색어 결정 모델

제안 모델의 전체 개요는 그림 1과 같다. 한국교육학술정보원(KERIS) DB에서 주제어 DB를 구축한 뒤 주제어를 빈도순으로 정리해 검색어를 선정하도록 한다. 그리고 검색어가 포함된 주제어 집합을 추출한 뒤 가중치 계산을 통해 Ego Network를 도출한다.

이 네트워크를 통해 검색어(Main Node)와 직접 연결된 연관검색어(Sub Node)를 추출해 RISS에서 제공하는 연관검색어와 비교 정리한다. 그 후 연관규칙 마이닝 기법, 유사계수를 이용해 어떤 방식이 더 유사도가 높은 연관검색어를 제공하는지 평가하기로 한다.

2.1 DB 구축 및 검색어 선정

주제어 DB를 구축하기 위해서 2002 ~ 2011년 한국정보과학회논문지를 대상으로 하였다.

구축 결과 총 3,547편의 논문과 15,068개의 주제어가 구축되었고, 이 중 'Ontology'를 검색어로 선정하여 'Ontology'가 포함된 66개의 주제어 집합을 추출하였다.

2.2 가중치(Weight) 계산

SNA에서 가중치 계산은 한 논문에서 나타난 주제어들은 서로 연관관계가 있다는 전제하에 가중치 계산을 하였다. 가중치 계산 원리는 아래 표 1의 예제DB를 통해 설명하기로 한다.

표 1 가중치 계산을 위한 예제DB

논문	주제어		
P1	A	B	C
P2	B	D	
P3	D	E	

표 1에서 보듯이 논문 P1은 {A,B,C}, 논문 P2는 {B,D}, 논문 P3은 {D,E}의 주제어를 갖고 있다. 한 논문에서 주제어들 간에는 서로 연관관계가 있으며 이때 가중치1의 값을 같게 된다. 이를 전체 행렬관계로 나타나면 아래 표 2와 같다.

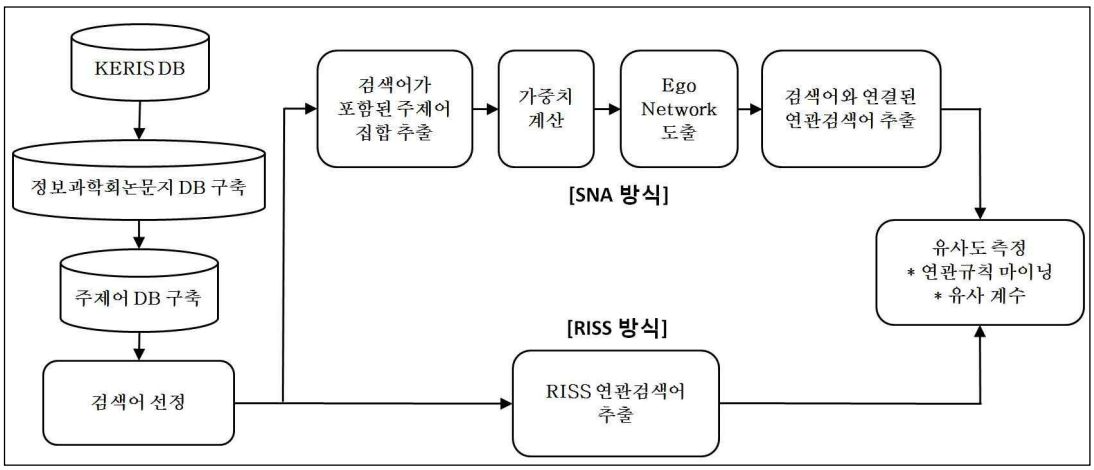

그림 1 제안모델의 전체 개요

표 2 예제 DB의 전체 행렬 계산

	A	B	C	D	E	합계
A		1	1			2
B	1		1	1		3
C	1	1				2
D		1			1	2
E				1		1
합계	2	3	2	2	1	10

예를 들어, 검색어가 주제어 D로 선정되었다고 가정하고, 주제어 D의 가중치 계산 현황만 살펴보면 아래 표 3과 같다.

표 3 주제어 D의 가중치 계산 현황

논문	키워드 관계성	가중치(Weight)
P2	B↔D	1
P3	D↔E	1
합계		2

표 2의 가중치 계산 결과값을 바탕으로 Ego Network를 도출하게 되면 아래 그림 2와 같다.

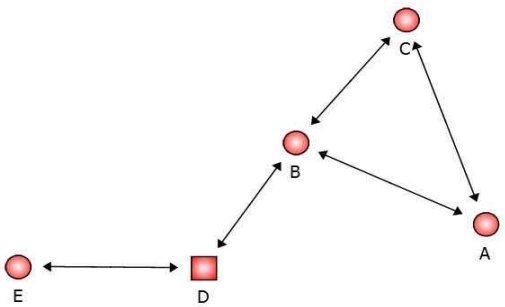

그림 2 예제 DB의 Ego Network 현황

그림 2에서 보듯이 검색어 D는 주제어 E, B와 직접적으로 연결되어 있음을 확인 할 수 있다. 또한 가중치 계산시 검색어와 간접적으로 연결된 노드들을 포함하고 있는 노드에 대해서는 HITS 알고리즘의 Hub-Authority[3] 개념을 적용해 간접노드 수를 가중치에 반영하였다. 그림 2에서 보면 주제어 B는 검색어 D와 직접적으로 연결되어 있으며 주제어 C, A를 검색어 D와 연결시켜 주는 HUB 역할을 하고 있음을 확인할 수 있다.

이를 바탕으로 검색어 D와 연결된 주제어들의 최종 가중치 현황을 정리하면 표 4와 같다.

표 4 검색어 D와 연결된 주제어들의 최종 가중치 현황

No	Source	Target	Weight	간접노드수
1	D	B	3	2
2	D	E	1	

2.3 Ego Network를 통한 연관검색어 추출

검색어 Ontology의 가중치 계산 결과 총 66개의 연관 주제어들이 생성되었으며, 이 중 가중치 2이상의 연관주제어를 정리하면 표 5와 같다.

표 5 'Ontology'의 연관 주제어 가중치 현황

No	Source	Target	Weight	간접노드수
1	Ontology	Semantic Web	12	21
2	Ontology	OWL(Web Ontology Language)	6	10
3	Ontology	Description Logic	4	3
4	Ontology	Ontology Reasoning	4	
5	Ontology	Relation	2	
6	Ontology	Tableaux Algorithm	2	
7	Ontology	Bioinformatics	2	
8	Ontology	WordNet	2	
9	Ontology	Concept	2	
10	Ontology	Gene Ontology	2	
11	Ontology	Knowledge Representation	2	
12	Ontology	Terminology	2	
13	Ontology	Optimized Method	2	

가중치 현황을 바탕으로 그림 3과 같이 Ontology의 Ego Network를 도출할 수 있으며, 링크의 굵기는 가중치 정도를 반영하였으며, 연관도가 가장 높은 3개의 노드들은 다이아몬드형으로 나타냈다.

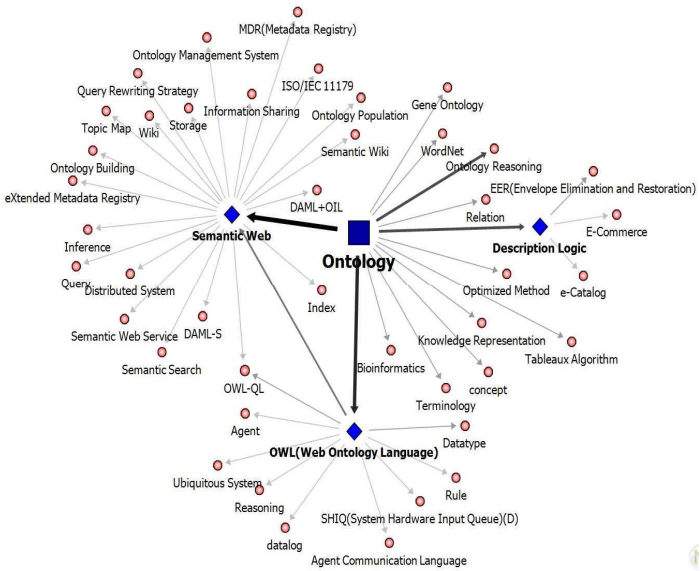

그림 3 Ontology Ego Network

검색어 Ontology의 가중치 계산 결과 간접노드 수 21개를 포함하고 있는 'Semantic Web'이 가중치 12로 가장 연관도가 높게 나타났으며, 그림 3에서 보듯이 다양한 연구주제분야들을 Ontology와 연결시켜주는 핵심적인 HUB 역할을 하는 것을 파악할 수 있다. 그리고 'OWL(Web Ontology Language)', 'Description Logic' 노드들도 다양한 연구주제분야들을 Ontology와 연결시켜 주는 것을 확인 할 수 있다.

이를 바탕으로 RISS에서 제공하는 Ontology의 연관검색어와 SNA 방식으로 도출한 Ontology의 연관검색어 집합 중 상위 3개를 정리하면 표 6과 같다.

표 6 'Ontology'의 연관 검색어 비교 현황

No	RISS	SNA
1	Science	Semantic Web
2	Learning	OWL(Web Ontology Language)
3	Processing	Description Logic

3. 실험 및 결과 분석

3.1 측정방법

유사도 측정의 객관성을 확보하기 위해 실제 RISS에서 제공하는 총 4,419,039편의 학위논문, 국내학술지를 대상으로 하였다. 그리고 검색어(X)와 연관검색어(Y)가 제목, 목차, 주제어, 초록 필드에서 동시에 나타나면 연관도가 있는 것으로 가정했고, 표 6의 Ontology의 연관 검색어 비교 현황을 바탕으로 유사도를 평가하였다.

3.1.1 연관규칙 마이닝 기법

연관규칙 마이닝은 대규모의 데이터 항목집합 사이에서 유용한 연관관계 및 상관관계를 찾는 방법이다. 본 연구에서는 전체트랜잭션수를 RISS의 국내 논문수인 4,419,039편으로 전제하고 지지도, 신뢰도, 향상도의 개념을 적용해서 측정 결과를 분석하였다.

① 지지도

$$\text{지지도}(X \rightarrow Y) = \frac{(X, Y)\text{를 포함하는 트랜잭션의 수}}{\text{전체 트랜잭션의 수}} = P(X \cap Y)$$

② 신뢰도

$$\text{신뢰도}(X \rightarrow Y) = \frac{(X, Y)\text{를 포함하는 트랜잭션의 수}}{X\text{를 포함하는 트랜잭션의 수}} = \frac{P(X \cap Y)}{P(X)}$$

③ 향상도

$$\text{향상도}(X,\ Y) = \frac{P(X \cap Y)}{P(X)P(Y)}$$

3.1.2 유사계수

유사계수에 의한 결과 분석 역시 RISS의 국내 논문수를 대상으로 측정 하였다. 먼저 용어 간 공기 빈도의 4가지 가능한 조합을 나타내기 위해서 표 7과 같은 2x2 분할표를 사용하였다.

표 7 2x2 공기 빈도 분할표

		용어 Y		합계
		출현	미출현	
용어 X	출현	a	b	a + b
	미출현	c	d	c + d
합계		a + c	b + d	N

표 7을 이용해서 실험에 사용할 각 유사계수 공식을 표현하면 표 8과 같다. 이 유사계수들은 0~1사이의 값을 가지며, 1에 가까울수록 유사성이 높음을 의미한다.

표 8 유사계수 정리표

명칭	공식	적용식
자카드 계수	$\frac{N_{a \cap b}}{N_{a \cup b}}$	$\frac{a}{a+b+c}$
다이스 계수	$\frac{2\lvert N_{a \cap b}\rvert}{N_a + N_b}$	$\frac{2\lvert a\rvert}{(a+b)+(a+c)}$
코사인 계수	$\frac{N_{a \cap b}}{\sqrt{N_a} \times \sqrt{N_b}}$	$\frac{a}{\sqrt{(a+b)(a+c)}}$

3.2 연관도 측정 결과 분석

3.2.1 연관규칙 마이닝 기법

실험 결과 연관규칙 마이닝 기법의 지지도, 신뢰도, 향상도 측정값은 아래 표 9 및 그림 4와 같다.

표 9 연관규칙 마이닝 기법 측정 결과

No	연관검색어	지지도	신뢰도	향상도
1	RISS(Science)	0.00005	0.0813	4.415
	SNA(Semantic Web)	0.00012	0.1709	717.883
	편차	▲0.00007	▲0.0896	▲713.468
2	RISS(Learning)	0.00004	0.0664	3.09
	SNA(OWL)	0.00008	0.1207	860.287
	편차	▲0.00004	▲0.0543	▲857.197
3	RISS(Processing)	0.00004	0.0692	5.762
	SNA(Description Logic)	0.00001	0.019	110.767
	편차	▼0.00003	▼0.0502	▲105.005

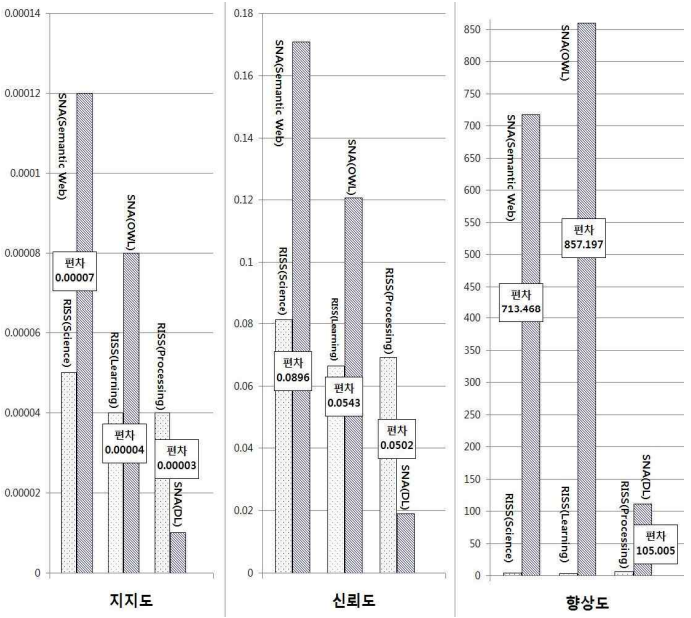

그림 4 연관규칙 마이닝 기법 측정 결과

분석결과 SNA기법을 이용해 제공된 1,2 순위 연관검색어들이 지지도, 신뢰도, 향상도 부분에서 RISS에서 제공한 연관검색어 보다 유사도가 높게 나왔다. 하지만 3위의 연관검색어의 경우 지지도 및 신뢰도 부분에서 RISS 연관검색어가 높게 나왔지만 향상도는 SNA보다 낮게 나왔다. 이는 RISS에서 제공하는 연관검색어가 학술 영역에 특화 되어 있기 보단 매우 일반적인 성격의 단어를 제공하기 때문이다.

3.2.2 유사계수

표 7의 2x2 공기 빈도 분할표를 바탕으로 'Ontology'의 유사계수 측정값은 표 10 및 그림 5와 같으며, 측정결과 SNA 방식으로 제공된 연관검색어가 연관도 측면에서 더 높은 것으로 측정되었다.

4. 결론 및 향후 과제

본 연구는 학술DB에서 SNA(Social Network Analysis) 기법을 적용해 검색어와 밀접한 연관도를 갖는 연관검색어를 제공함에 그 목적이 있다. 실험결과 기존 학술DB에서 제공하는 연관검색어 방식 보다는 SNA기법을 적용해 연관검색어를 제공하는 방식이 연관도 측면에서 더 높게 측정되었다.

표 10 유사계수 측정 결과

No	연관검색어	자카드	다이스	코사인
1	RISS(Science)	0.003	0.006	0.015
	SNA(Semantic Web)	0.146	0.256	0.295
	편차	▲0.143	▲0.25	▲0.28
2	RISS(Learning)	0.002	0.004	0.012
	SNA(OWL)	0.112	0.201	0.272
	편차	▲0.11	▲0.197	▲0.26
3	RISS(Processing)	0.003	0.007	0.016
	SNA(Description Logic)	0.015	0.03	0.038
	편차	▲0.012	▲0.023	▲0.022

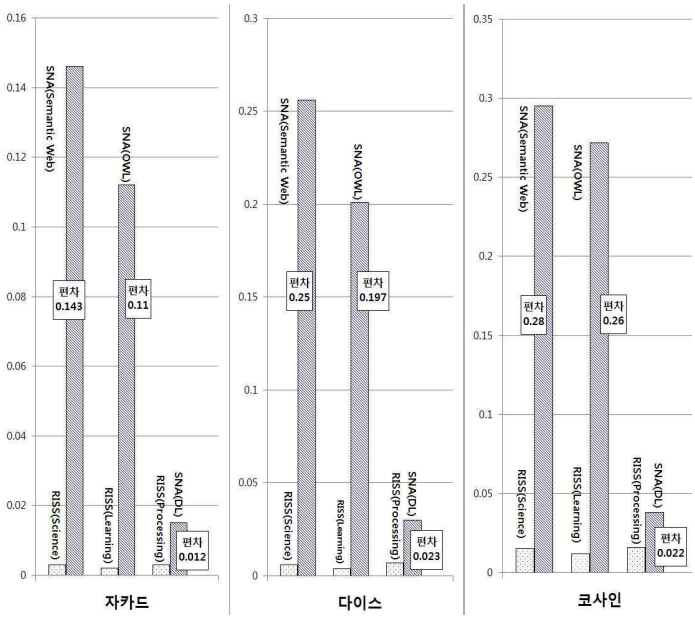

그림 5 유사계수 측정 결과

향후 과제로는 주제어 DB 구축 시 좀 더 다양한 주제 분야의 학술 DB를 수집해 구축하는 것이 필요하다. 제안된 연구에서는 한 주제 분야의 학술지만을 DB로 사용하다 보니 다양한 연구 분야의 연관검색어를 제시하는 측면에서는 한정적이었다.

앞으로 여러 주제 분야의 학술DB를 수집한다면 다양하고 의미 있는 연관검색어들을 제공할 수 있을 것이다.

참고문헌

[1] D. Wilkinson, G. Harries, M. Thelwall, and E. Price, "Motivation for academic Web site interlinking: Evidence for the Web as a novel source of information on informal scholarly communication", Journal of information Science, vol. 29, no. 1, pp. 59~66, 2003.

[2] 김용학, 사회 연결망 분석, 개정판, 서울 : 박영사, 2007.

[3] Taeho Jo, Malrey Lee and Thomas M Gatton, "Keyword Extration from Documents Using Neural Network Model", International Conference on Hybrid Information Technology, Vol.2, pp.194~197, 2006.