팩터그래프 모델을 이용한 연구전선 구축: 생의학 분야 문헌을 기반으로*

Construction of Research Fronts Using Factor Graph Model in the Biomedical Literature

김혜진 (Hea-Jin Kim)**
송 민 (Min Song)***

초 록

연구전선이란 연구논문들 간에 인용이 빈번하게 발생하며, 지속적으로 발전이 이루어지고 있는 연구영역을 의미한다. 연구행위가 집중되는 핵심 연구분야로 발전 가능성이 높은 연구전선을 조기에 예측해내는 것은 학계와 산업계, 정부기관, 나아가 국가의 과학기술 발전에 큰 유익을 가져다 줄 수 있는 유용한 사회적 자원이 된다. 본 연구는 복합자질을 활용하여 연구전선을 추론하는 모델을 제시하고자 시도하였다. 연구전선 추론은 핵심 연구영역으로 발전할 가능성이 높은 문헌들이 포함될 수 있도록 문헌을 복합자질로 표현하고, 그 자질들을 심층학습하여 새로 발행된 문헌들이 연구전선에 포함될 수 있는지 그 가능성을 예측하였다. 서지 자질, 네트워크 자질, 내용 자질 등 복합자질 세트를 사용하여 문헌을 표현하고 피인용을 많이 받을 가능성이 있는 문헌을 추론하기 위해서 확률기반 팩터그래프 모델을 적용하였다. 추출된 자질들은 팩터그래프의 변수로 표현되어 합-곱 알고리즘과 접합 트리 알고리즘을 적용하여 연구전선 추론이 이루어졌다. 팩터그래프 확률모델을 적용하여 연구전선을 추론・구축한 결과, 서지결합도 4 이상으로 구축된 베이스라인 연구전선과 큰 차이를 보였다. 팩터그래프 기반 연구전선그룹이 서지결합 기반 연구전선그룹보다 문헌 간의 직접 연결정도가 강하며 연결 관계에 있지 않은 두 개의 문헌을 연결시키는 매개정도 또한 강한 집단으로 나타났다.

ABSTRACT

This study attempts to infer research fronts using factor graph model based on heterogeneous features. The model suggested by this study infers research fronts having documents with the potential to be cited multiple times in the future. To this end, the documents are represented by bibliographic, network, and content features. Bibliographic features contain bibliographic information such as the number of authors, the number of institutions to which the authors belong, proceedings, the number of keywords the authors provide, funds, the number of references, the number of pages, and the journal impact factor. Network features include degree centrality, betweenness, and closeness among the document network. Content features include keywords from the title and abstract using keyphrase extraction techniques. The model learns these features of a publication and infers whether the document would be an RF using sum-product algorithm and junction tree algorithm on a factor graph. We experimentally demonstrate that when predicting RFs, the FG predicted more densely connected documents than those predicted by RFs constructed using a traditional bibliometric approach. Our results also indicate that FG-predicted documents exhibit stronger degrees of centrality and betweenness among RFs.

키워드: 내용 자질, 네트워크 자질, 서지 자질, 심층학습, 연구전선, 인용분석, 팩터그래프, 확률 그래프 모델
Bibliographic features, content features, factor graph model, network features, probabilistic graphical model(PGM), research front

---

* 이 논문은 박사학위논문의 일부를 요약한 것임.
본 연구는 미래창조과학부 및 한국연구재단의 (재)유전자동의보감사업단(NRF-2013M3A9C4078138) 연구비 지원에 의해 수행되었음.
** 연세대학교 문헌정보학과 박사후연구원(erin.hj.kim@yonsei.ac.kr) (제1저자)
*** 연세대학교 문헌정보학과 교수(min.song@yonsei.ac.kr) (교신저자)
■ 논문접수일자 : 2017년 2월 20일 ■ 최초심사일자 : 2017년 3월 5일 ■ 게재확정일자 : 2017년 3월 6일
■ 정보관리학회지, 34(1), 177-195, 2017. 〔http://dx.doi.org/10.3743/KOSIM.2017.34.1.177〕

1. 서 론

1.1 연구의 필요성 및 목적

의학 및 과학기술의 발전을 주도하는 원동력은 연구개발에 대한 투자이다. Narin과 Hamilton (1996)의 연구에 의하면 기관 및 정부기관이 연구개발을 주도한 경우 연구성과가 특허로 이어지는 비율이 높고, 국가의 기술적인 발전에 대한 기여도 또한 매우 높은 것으로 나타났다. 대학 기관의 연구개발 비용과 특허 신청률 역시 비례 관계로 나타났다(Jaffe & Trajtenberg, 1996). 성과로 연결되는 효과적인 투자를 이끌어내기 위해서는 적절한 투자를 이끌어 낼 수 있는 분야 즉, 연구행위가 집중되어 있는 핵심 영역을 조기에 발견하는 것이 중요하다.

De Solla Price(1965)는 "연구전선(research fronts)"이란 개념을 도입하여 핵심 연구영역의 파악을 시도하였다. 그는 인용분석을 통하여, 연구논문들 간에 인용이 빈번하게 발생하고 있으며 지속적으로 발전이 이루어지고 있는 연구영역을 연구전선으로 정의하면서, 연구자들에게는 가장 최근에 출판된 연구를 인용하려는 경향이 있기 때문에 연구전선은 가장 최신의 연구를 토대로 구성되며, 그 네트워크는 매우 조밀한 양상을 보이게 된다고 설명하였다.

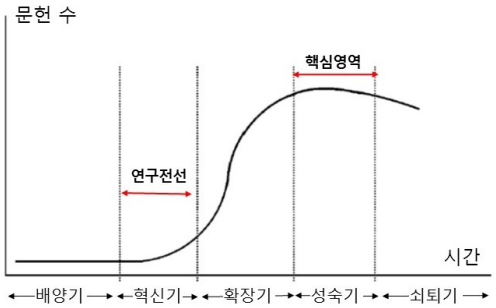
연구전선을 파악하는 대표적인 계량서지학적 인용분석 기법으로는 동시인용법(cocitation analysis)과 서지결합법(bibliographic coupling)이 있다. 동시인용법으로 구축한 연구전선은 문헌 발행 후 피인용빈도가 축적되기까지 상당한 기간이 요구된다는 단점이 있고 서지결합법으로 구축한 연구전선은 피인용빈도가 높은 문헌들의 포함정도를 알 수 없다는 단점이 있다. 실제로 Web of Knowledge(WoK)의 Essential Science Indicators(ESI) 데이터베이스에서 제공하는 연구전선은 피인용빈도가 높은 문헌만을 대상으로 구성된다. ESI는 동시인용법 기반 연구전선을 산출하는데 Science Citation Index (SCI)와 Social Science Citation Index(SSCI)에 색인되는 모든 논문들을 대상으로 해서 피인용빈도 순위가 상위 100,000에 포함되는 문헌들로 한정하여 연구전선 문헌리스트를 산출한다. 그렇기 때문에 결과적으로 ESI에서 산정한 연구전선 문헌리스트에는 발행 직후의 문헌은 배제된 상태로 현재로부터 10년 전의 문헌 데이터를 소급하여 산정한 것이 되어버린다. 서지결합법으로 구축한 연구전선은 인용문헌들(citing papers)을 참고문헌(cited papers)으로 결합하여 참고문헌을 공유하고 있는 문헌리스트를 산출한다. 그렇기 때문에 현재 발행된 문헌들도 포함할 수 있다는 장점이 있지만 이 경우 인용문헌의 피인용빈도는 고려하지 않는다. 기존의 방법론이 연구전선 예측이 아니라 연구전선 파악에 머무를 수밖에 없는 이유는 전통적인 계량서지학적 인용분석 기법이 "인용"이라는 단일자질을 이용하여 지식구조(knowledge structure)를 구조화하고 연구전선을 구축하기 때문이다. 따라서 기존의 현상을 설명하는 연구전선의 파악보다는 향후 활발한 연구 활동으로 이어질 가능성이 높은 연구전선을 예측하는 연구가 필요하다.

본 연구에서 제시한 연구전선 추론모델은 핵심 연구영역으로 발전 가능성이 높은 문헌들이 포함될 수 있도록 문헌을 복합자질로 표현하고, 그 자질들을 학습하여 새로 발행된 문헌들이

연구전선에 포함될 수 있는지 그 가능성을 예측하였다.

〈그림 1〉에서 "혁신기"의 문헌들은 활발한 인용 행위가 발생하는 "확장기"를 지나 "성숙기"에 진입하며 핵심 연구논문이 된다. 본 연구는 Shibata, Kajikawa, Matsushima(2007)가 제시한 지식의 발전 주기 모형에서 확장기와 성숙기로 연결되는 혁신기의 문헌집단이 참 연구전선이라는 전제와 함께 궁극적으로 활발한 인용이 발생할 가능성이 높은 연구전선을 예측한다. 따라서 본 연구에서 예측하고자 하는 연구전선의 가장 큰 특징은 "**최신성(up-to-dateness)**"과 "**미래의 활발한 인용활동(potentially-high-impact)**"이다.

본 연구의 연구문제는 다음과 같다. 첫째, 전통적인 계량서지학 모델인 서지결합 기법을 적용하여 구축한 연구전선은 실제로 피인용빈도가 높은 문헌을 어느 정도 포함하고 있는가? 둘째, 전통적인 계량서지학 모델을 적용하여 구축한 연구전선과 비교하였을 때 확률기반 팩터그래프 추론모델을 활용하여 구축한 연구전선에는 실제 피인용빈도가 높은 문헌을 더 포함하고 있는가?



〈그림 1〉 지식의 발전 주기
(출처: Shibata et al., 2007, p. 873에서 인용함)

1.2 용어의 정리

본 연구에서 언급한 주요 용어에 대한 개념적 정의는 다음과 같다.

(1) 연구전선그룹(Research Front, RF): De Solla Price(1965)의 정의를 바탕으로 논문들 간의 인용이 빈번하게 발생하며 지속적으로 발전하고 있는 연구영역을 형성하는 문헌들로서 향후 지식의 발전 주기에 있어 성숙기에 이르는 핵심 연구논문으로 발전할 가능성이 높은 문헌집합을 의미한다. 따라서 본 연구에서 연구전선그룹이란 문헌 발행 후 활발한 피인용이 계속적으로 발생할 것으로 예측되는 문헌들의 집합을 지칭하는 개념으로 사용하였다.

(2) 연구그룹(Non-Research Front, NRF): 연구전선그룹에 속하지 않은 문헌집합을 의미하는 것으로, 특정 학문영역의 보편적인 지식기반을 이루는 문헌집합을 지칭한다.

(3) 팩터(factor): 팩터그래프상의 팩터는 하나 이상의 변수들 사이에 존재하는 상호작용을 표현하는 함수를 의미한다(Shen & Coughlan, 2007).

본 논문의 구성은 다음과 같다. 제2장 이론적 배경에서는 연구전선 및 팩터그래프 개념을 설명하고 관련 선행연구를 제시하였다. 제3장에서는 팩터그래프 기반 연구전선 추론모델의 실험문헌 집단과 실험을 위한 연구방법론 전반에 대해서 다루었고, 제4장은 팩터그래프 기반 연구전선 추론모델의 연구전선 구축 결과를, 제5장은 결론을 제시하였다.

2. 이론적 배경

2.1 연구전선

기존의 연구전선 파악을 위한 연구들은 계량서지학의 인용기반 매핑기술인 서지결합법과 동시인용법을 이용해 왔다. 서지결합법의 기본 개념은 동일한 참고문헌을 서로 얼마나 많이 공유하고 있는지에 따라 문헌들을 범주화하는 것이다(Jarneving, 2007). 반면에 동시인용법은 학술논문의 참고문헌 목록 안에서 동시인용된 문헌을 범주화하는 것이다(Small, Sweeney, & Greenlee, 1985). 동시인용법은 문헌이 발표된 후 함께 인용되는 문헌들(cited papers)을 연결시키는 것이고, 서지결합법은 문헌들을 인용한 문헌들(citing papers)을 연결시키는 것이다. 그렇기 때문에 시간이 경과함에 따라 동시인용의 강도는 계속 증가할 수 있지만 서지결합의 강도는 시간의 흐름과 무관한 양상을 보인다(Small, 1973). Persson(1994)은 다른 문헌들을 인용하고 있는 인용문헌들, 즉 서지결합으로 연결된 문헌들을 연구전선, 동시인용된 문헌들을 지식기반(intellectual base)라고 명명하였다.

역사적으로 인용이 활발히 일어나는 영역을 파악하기 위해서는 동시인용법이 좀 더 많이 사용되지만, 서지결합법에 비해 동시인용법은 인용빈도를 파악하는 데에 상당한 정도의 시간적 지연이 발생한다. Jarneving(2007)은 서지결합법이 동시인용법에 비해 더 많은 장점을 가지고 있음을 강조하였는데, 그중에서도 서지결합법이 특히 전문분야의 점진적인 발전을 초기단계에서 포착할 수 있는 가능성을 지니고 있다는 점에 주목하였다. 그는 서지결합도가 강한 문헌들은 연구전선 구조에 대한 통찰을 제공할 수 있으며, 학문의 지식구조 생성을 위한 목적을 위해서도 이용될 수 있다고 주장하였다. 그러나 동시인용법과 달리 서지결합도를 이용하여 연구전선 네트워크를 형성하였을 때는 실제로 그 구성 논문들이 얼마나 피인용 되었는지 확인이 불가능하기 때문에 실제 연구전선을 구성하고 있는 개별문헌에 대한 활발한 피인용은 파악할 수 없다.

개별 문헌들의 인용이력 분석을 통해 개별적인 학술문헌들의 수명 패턴(aging pattern)을 알아내는 연구들을 일컬어 통시적 연구라 하는데, 이러한 연구들은 인용빈도의 변화를 조사하기 위해서 시간이 흐름에 따른 저널의 인용/피인용빈도와 개별 학술문헌의 피인용빈도를 이용한다(McCain & Turner, 1989). 유사한 방법으로, 시간 흐름에 따른 성장률을 측정하여 연구전선을 성장(growing), 쇠태(shrinking), 안정(stable), 신흥(emerging), 퇴장(exiting)의 다섯 개 범주로 구분하기도 한다(Upham & Small, 2010).

Shibata, Kajikawa, Takeda, Matsushima (2009)는 연구전선을 파악하기 위한 새로운 방법을 제안하였는데, 동시인용법이 아닌 인용을 서로 주고받는 두개의 문헌들을 이용해서 인용 네트워크를 구축하는 방식을 선보였다. 이것은 동시인용법의 경우 두 문헌들이 동시인용될 때까지는 일정 정도의 시간이 소요된다는 문제점을 극복하기 위해 고안된 것이다. 또한 이것을 바탕으로 군집화된 인용 네트워크 안에서 위상을 측정하는 방법을 이용한 위상적 군집화(topological clustering) 기법도 제안하였는데, 이 기법은 각

군집에서의 문헌의 위상을 분석하여 문헌들 간의 연결밀도가 높은 군집을 발견한다. Shibata et al.(2009)의 연구에서 제시한 위상 측정방법은 학술적인 출판물들의 인용 네트워크 안에서 혁신이 출발하는 지점을 발견하는 데에 유익한 방법이라 할 수 있다.

지금까지 설명한 연구들은 인용 네트워크를 생성하는데 있어 계량서지학적 방법을 이용하고 있다. Song과 Kim(2013)은 생물정보학 분야의 지식 구조를 파악하기 위한 새로운 방법으로 토픽모델링과 같은 텍스트마이닝 기법들을 이용해 핵심 문헌을 분석하였다. 토픽모델 생성을 위해 잠재 디리클레 할당(Latent Dirichlet Allocation, LDA) 방법을 이용했으며 핵심 문헌을 식별하기 위해 PageRank 기법을 이용하였다. 분석결과 생물학 측면에서의 10개 주요 토픽을 식별하였으며, 페이지랭크(PageRank) 값과 단어 동시출현 분석 결과를 이용하여 저자 동시인용 네트워크 안에서 주요 저자 다섯 명을 선정하여 분석하였다.

국내의 문헌정보학・정보학분야의 연구동향은 주로 단어 동시출현 빈도를 이용한 네트워크 분석을 적용하여 수행되었고(서은경, 유소영, 2013; 조재인, 2011), 김조아와 이재윤(2016)의 연구에서는 인용 이미지 구축자 프로파일링 기법을 제안하여 국내 여성학 분야의 연구전선을 제시한 바 있다.

기존의 연구전선에 관한 연구들은 각 학문분야의 지식구조를 파악하여 핵심 연구영역을 제시하는 연구들, 즉 현상을 분석하는 연구들이다. 그러나 본 연구는 지식구조를 분석하기 위하여 연구전선을 제시하는 연구가 아니라, 지식구조를 예측하기 위하여 연구전선을 제시하는 연구로 기존의 연구들과는 그 방향을 달리하고 있다. 즉, 미래에 핵심 주제영역으로 발전 가능성이 높은 문헌들로 구성된 연구전선그룹을 추론하여 제시하기 위해 본 연구를 기획・진행하였다.

2.2 팩터그래프

확률이론에 기반하여 관찰된 데이터의 확률분포를 학습한 후 이를 바탕으로 현재의 데이터를 설명하는 모델을 확률모델이라고 하며 이것을 그래프로 표현한 것이 확률 그래프 모델이다(Sun, Deng, & Han, 2012). 그래프 모델에서 변수는 노드(node)로 표현되고 노드간의 관계는 방향성이 없는 엣지(edge) 또는 방향성을 갖는 아크(arc)로 표현되고, 모든 변수의 분포를 지역함수(local function)의 곱으로 계산할 수 있다(Jordan, 2004; Sutton & McCallum, 2007). 지역함수란 두개의 변수 사이에 있는 함수를 의미한다. 그래프 모델은 복잡한 확률구조를 간결하게 표현할 수 있다는 장점이 있다(Frey, 1998). 팩터그래프는 그래프 모델의 한 종류로서 변수의 곱을 나타내는 결합 확률분포(joint probability distribution)를 인수분해(factorization)하여 확률 공간상의 가장 작은 단위인 인수(factor)로 표현한다. 본 연구에서는 이해를 돕기 위하여 factor를 의미하는 "인수"를 "팩터" 그대로 명명한다. 팩터그래프상의 팩터는 하나 이상의 변수들 사이에 존재하는 상호작용을 표현하는 함수를 의미한다(Shen & Coughlan, 2007). 팩터그래프 모델은 변수들 사이에서 발생하는 팩터들을 분해하고 팩터들 간의 곱을 계산하여 결합 확률분포를 구한다

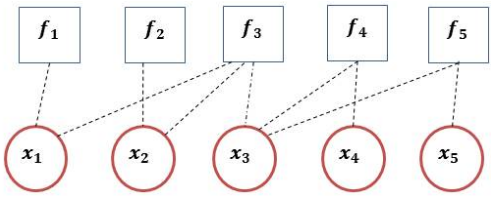
(Frey, 1998; Kollar & Friedman, 2009).

팩터그래프는 변수를 원형노드로 표현하고 팩터를 사각노드로 표현하는 양분 그래프(bipartite graph)이다. 〈그림 2〉처럼 팩터그래프는 원형으로 표시된 각 변수 x_i에 대한 변수노드와 사각형으로 표시된 지역함수 f_j에 대한 팩터노드 등 두 종류의 노드가 있다. 팩터그래프 상의 변수노드 x_i와 인접한 변수노드 x_{i+1}은 팩터노드로 연결된다. 팩터그래프 상의 변수노드 x_i는 변수를 의미하므로 팩터노드는 두개의 변수 사이의 의존성을 의미한다. 다음의 인수분해식 (1)을 통하여 팩터그래프를 간단히 설명하고자 한다.

$$g(x_1,x_2,x_3,x_4,x_5) = f_1(x_1)f_2(x_2)f_3(x_1,x_2,x_3)f_4(x_3,x_4)f_5(x_3,x_5)$$

- 인수분해식 (1)

〈그림 2〉는 위의 인수분해식의 구조를 표현한 것이다. 그래프 g는 각 변수노드 x_1,x_2,x_3,x_4,x_5로 표현된다고 할 때 팩터 f_1은 변수노드 x_1과 연결된 함수이며 f_2는 x_2와 연결된 함수이다. 이런 식으로 f_3은 x_1,x_2,x_3과 f_4는 x_4,x_5과 연결된 함수이다.



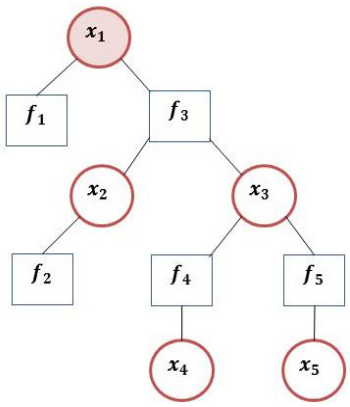
〈그림 2〉 인수분해식을 표현한 그래프 모형

그래프 모델에서 변수의 이웃(neighbor)은 하나의 팩터노드를 통해 연결되는 변수노드를 의미한다. 즉, 팩터로 연결된 두개의 변수사이를 이웃변수라고 명한다. 그러므로 두 이웃변수는 팩터로 연결되어 있고, 각 변수노드와 팩터노드는 엣지로 연결된다. 팩터그래프에서 개별 팩터들과 변수노드를 연결하는 엣지들이 서브 그래프들을 형성하고 이들이 모여 전체 분포를 나타내는 팩터그래프를 형성하게 된다. 두개의 변수 사이에 있는 함수를 지역함수라 하고 그래프내의 모든 지역함수들의 결과물을 전역함수(global function)라고 한다. 〈그림 2〉에서 지역함수는 각 f_1,f_2,f_3,f_4,f_5이며 이를 모두 곱한 전역함수 f는 $g(x_1,x_2,x_3,x_4,x_5)$를 표현하는 함수라고 할 수 있다(Kschischang, Frey, & Loeliger, 2001; Loeliger, 2004; Yeh, Breeden, Yang, Fisher, & Hanrahan, 2013). 팩터그래프는 결합확률을 구성하는 팩터들을 보다 정교하게 나타내고 이러한 구조를 이용하는 추론 알고리즘을 정의하는데 유용하다(Jordan, 2004).

팩터그래프 상에서 이루어지는 전역함수를 계산하는 알고리즘은 합-곱(sum-product) 알고리즘이 대표적이다(Kschischang, Frey, & Loeliger, 2001). 합-곱 알고리즘은 메시지 전달(massage-passing) 알고리즘이라고도 하는데, 이는 변수노드와 이웃 변수노드 사이에 전달되는 함수를 메시지로 표현하기 때문이다(Shen & Coughlan, 2007). 팩터그래프에서 연결된 변수노드들 중에서 상위에 위치한 노드를 부모노드라고 하고 그 밑의 하위에 위치한 노드를 자식노드라고 한다. 메시지가 자식노드에서 상위 부모노드로 전달되면서 팩터노드는 활성화된다(Kschischang, Frey, & Loeliger, 2001).

〈그림 3〉은 앞에서 언급한 인수분해식의 전

역함수를 통해서 전달되는 메시지를 표현하기 위해서 〈그림 2〉의 x_1을 루트노드(root node)로 하는 팩터그래프로 재구성한 것이다. 〈그림 3〉과 같이 루트노드와 자식노드로 구성된 트리구조로 된 팩터그래프로 변형하면 변수들 사이에 전달되어지는 메시지들을 보다 쉽고 빠르게 계산할 수 있게 된다. 팩터그래프 상에 최상위 노드와 자식노드를 제외한 중간에 위치한 노드들은 메시지의 전달 시점에 따라 부모노드가 되기도 하고 자식노드가 되기도 한다(Kschischang, Frey, & Loeliger, 2001).



〈그림 3〉 〈그림 2〉의 x_1을 루트노드로 재구성한 팩터그래프

합-곱 알고리즘은 이러한 메시지 전달에 있어서 자식노드에서부터 시작하여 상위의 부모노드로 메시지를 전달하는 과정을 제일 위에 위치한 최상위 부모노드에까지 이르게 하는데, 각 노드를 지나가는 메시지를 지역함수의 곱으로 계산하고 이것이 상위의 부모노드로 전달되어서 합해지는 반복적인 합과 곱의 연산으로 구성되기 때문에 합-곱 알고리즘이라 명한다. 합-곱 알고리즘의 변형으로 최대-곱(max-product) 알고리즘이 있다(Weiss & Freeman, 2001). 이 알고리즘은 부모노드에서 메시지의 수집이 발생할 때 주변함수값 중에서 최대값을 가진 변수 x를 취하여 메시지를 전달하는 방식을 취한다. 따라서 이 알고리즘을 최소-합(min-sum) 알고리즘이라고도 한다.

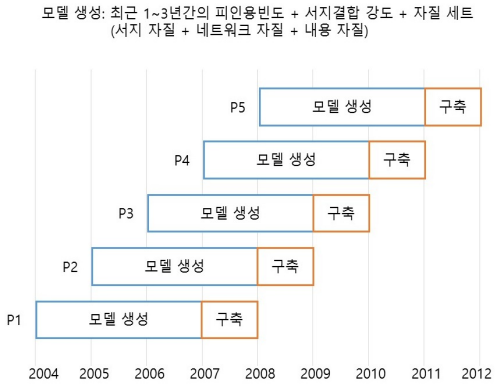
팩터그래프 상의 변수들 간의 메시지 전달이 이루어질 때, 변수 x가 발생했을 때 변수 y가 발생할 확률 즉, $P(y|x)$를 구하는 추론 알고리즘은 크게 두 가지로 나누어 볼 수 있다. 하나는 근사적 추론(approximate inference)이고 다른 하나는 정확한 추론(exact inference) 알고리즘이다. 근사적 추론은 주변확률분포를 근사적으로 샘플링하여 생성하여 추론을 수행하는 것으로 실제 확률분포 대신 이를 근사화하는 근사 확률분포를 통해 해결하는 기법이다. 근사적 추론의 대표적인 알고리즘은 깁스 샘플링(Gibbs sampling) 방법이 있다. 정확한 추론이란 그래프 상의 모든 변수에 대해 정확한 주변확률을 구하여 결합 확률분포를 계산한다. 임의의 부모노드의 결합 확률분포를 계산하고자 할 때, 최하위 자식노드에서부터 시작하여 정해진 부모노드에게 모든 메시지가 전달되어지면 계산이 종료된다. 그러므로 팩터그래프가 선형 체인이거나 트리 구조일 때 효율적으로 수행할 수 있는 알고리즘이다(Sutton & McCallum, 2007). 정확한 추론의 대표적인 알고리즘은 접합 트리(Junction Tree) 알고리즘이다. 접합 트리 알고리즘은 팩터그래프에 존재하는 변수들을 군집화하여 트리형태로 만든 후, 이 트리구조를 이용하여 추론한다. 본 연구에서는 최대-곱 알고리즘과 접합 트리 알고리즘을 사용하여 전역함수를 계산하였다.

3. 데이터 수집 및 연구방법

3.1 데이터 수집 및 모델 평가

본 연구에서는 연구전선을 형성하기 위한 주제분야를 전염병학으로 한정하였다. 전염병학은 공중 보건 및 공공 위생을 다루는 생의학(Biomedical Science)의 한 분야로 보건 위생의 기초 연구를 포함하여 임상 연구와 공중 보건 연구에 적용할 수 있는 다양한 방법론 개발을 다루고 있는 분야이다(Porta, 2014). 전염병학의 연구전선을 파악하기 위해 2004년에서부터 2011년까지 발행된 해당 분야의 연구논문과 2014년도까지 발생한 피인용빈도를 함께 수집하였다. 전염병학과 관련된 논문들을 수집하기 위해 PubMed[1]에서 제공하는 NLM 목록에서 색인되는 저널 중 "epidemiology"를 포함하고 있는 저널명을 검색하여(검색식: "epidemiology"[Journal] AND currently indexed[All]) AND English [lang]) 25종의 저널을 검색하였고, 이 중 영어로 본문을 제공하는 21종의 저널을 본 연구의 수집 데이터 대상으로 한정하였다. 선정된 전염병학 저널에서 발행한 논문들의 서지정보와 초록을 WoK 데이터베이스에서 다운로드 받아 저장하였다. [부록 1]은 본 연구에 사용된 문헌집단의 저널명과 논문 수를 나타낸다. 수집된 논문들의 전처리 및 서지사항 추출은 Bibexcel[2]을 사용하여 파싱하여 처리하였다.

〈그림 4〉는 본 연구에서 제시하는 팩터그래프기반 연구전선 추론모델의 생성과 이것을 가지고 연구전선을 구축하는 단계를 보여준다.



〈그림 4〉 연구전선 추론모델의 생성 및 예측

수집된 문헌의 발행기간은 총 8년이다. 이 기간을 총 다섯 기간으로 분할하여 연구전선의 변화를 실험문헌 집합에 반영하고 다양한 실험문헌 집합을 구성하였다. 일반적인 연구전선에 관한 선행연구는 중첩되는 기간이 없는 방식으로 기간을 설정한다. 그러나 본 연구는 연구전선이 연속성을 가지고 변화하는 것으로 보고, 기간 분할에 있어, 이때 단위 기간은 4년을 한 단위로 하여 2004-2007년(기간1), 2005-2008년(기간2), 2006-2009년(기간3), 2007-2010년(기간4), 2008-2011년(기간5)으로 분할하면서 1년씩 최신 방향으로 진행하게 하였으며 각 기간은 3년의 중첩되는 기간을 가지도록 구성하였다. 4년으로 구성된 한 기간 내에서 처음 3년 동안 발행한 문헌집합을 학습문헌 집합으로 활용하여 추론모델을 생성하였고, 이 모델을 적용하여 마지막 1년 동안의 연구전선을 예측하여 제시하였다. 이와 같은 기간분할을 통하여 얻을 수 있는 효과는 첫째, 수집한 문헌집단내의 연구전선의 연속성을 반영할 수 있어 실험의 신뢰

1) http://www.ncbi.nlm.nih.gov/pubmed
2) https://bibliometrie.univie.ac.at/bibexcel/

도를 높일 수 있다는 것이고, 다음으로 동일한 논문이라 하더라도 속하는 기간에 따라 다른 논문과 서지결합 기반 연결이 형성될 수 있어 다양한 실험문헌 집합 구성이 가능하다는 것이다.

마지막으로, 추론모델을 통하여 예측된 연구전선의 평가는 서지결합 기반 연구전선(베이스라인)과 비교하여 두 개의 서로 다른 방법으로 구축된 연구전선이 문헌 발행 후 2014년까지 축적 피인용빈도를 기준으로 높은 피인용빈도를 받은 문헌을 얼마나 포함하고 있는지를 비교하였다. 베이스라인 연구전선의 구축 방법으로 서지결합법을 선택한 이유는 본 연구에서 제시하는 팩터그래프 기반 연구전선 추론모델이 2007년부터 2011년까지 발행된 문헌들이 연구전선그룹에 속할 가능성을 예측하는 모델이기 때문이다. 만약 동시인용빈도를 기반으로 연구전선을 형성한다면 피인용문헌(cited paper)을 대상으로 연구전선을 형성하게 되어 추론모델을 적용하게 될 대상 문헌의 발행연도와 베이스라인 연구전선의 발행연도가 불일치하게 되므로 동일한 문헌들을 가지고 평가하기 어렵게 된다. 그러므로 본 연구에서는 서지결합법을 활용하여 인용문헌(citing paper)을 대상으로 베이스라인 연구전선을 구축하였고 서지결합강도는 4 이상의 문헌을 대상으로 구축하였다.

3.2 자질 추출

연구전선 추론모델을 만들기 위하여 필요한 자질들은 크게 세 가지로 나누어진다. (1) 서지자질(bibliographic features), (2) 네트워크 자질(network features), (3) 내용 자질(content features).

- **서지 자질:** 서지 자질은 WoK의 필드 태그 정보를 파싱하여 추출한다. 서지 자질은 저자 수준, 논문 수준, 저널 수준을 반영할 수 있는 자질로 구성하였으며 추출된 서지 자질은 총 8개로 〈표 1〉과 같다.
- **네트워크 자질:** 네트워크 자질은 문헌 네트워크에서 해당 문헌이 가지는 위상을 반영하는 자질이다. 문헌 네트워크 내의 문헌들은 인용관계로 연결할 수 있다. 인용관계는 동시인용, 서지결합, 직접인용 등 크게 세 가지로 나누어 볼 수 있다(Boyack & Klavans, 2010). 새로운 문헌이 출판될 때 그 문헌은 문헌이 속한 주제적 범주 혹은 학문영역의 문헌집합에서 특정 위상을 가지게 될 것이다. 이러한 위치는 문헌집합의 개체가 증가하거나 혹은 문헌집합을 규정하는 범위의 변화가 발생함에 따라 변하게 될 것이다. 본 연구는 연구전선그룹을 추론하기 위한 단위 기간을 4년으로, 총 다섯 기간을 설정하였고, 각 기간은 1년씩 문헌 발행기간이 이동한다. 따라서 각 기간에 속하는 개별 문헌들은 최대 세 개의 기간에 포함될 수 있는 가능성을 가지고 있는데, 각 기간별로 다른 문헌 네트워크에 속하게 되므로 매번 다른 위상을 가지게 된다. 네트워크 자질은 바로 이러한 동적인 네트워크 특성을 추론모델에 반영하고자 고안한 자질이다. 네트워크에서 노드의 위상 정보는 중앙성(centrality)을 통해 측정할 수 있다. 중앙성이란 한 노드가 네트워크에서 얼마나 중심에 위치하는지에 대한 정도를 의미하는 것이다. 본 연구에서는 서지결합을 적용하여 문헌 네트워크를 구축하였고 네트워크 중앙성 분석을 통하

〈표 1〉 팩터그래프의 변수로 변환된 자질 값 예시

구분	자질(약어)	자질 값	변수 값
서지 자질	① 저자 수(AU)	7	0.048951
	② 저자의 기관 수(INST)	8	0.075268
	③ 프로시딩 유무(PROC)	0	0
	④ 저자 키워드 수(KW)	6	0.352941
	⑤ 펀딩 유무(FU)	0	0
	⑥ 참고문헌 수(NR)	28	0.073878
	⑦ 페이지 수(PG)	10	0.046511
	⑧ 저널 영향력 지수(JIF)	4.389	0.540937
네트워크 자질	⑨ 연결정도(DG)	18	0.032085
	⑩ 사이중앙성(BW)	868.290797	0.001734
	⑪ 인접중앙성(CL)	0.000038	0.000038
내용 자질	⑫ 추출된 표제 핵심어(TI)	Communication, Communism, Myocardial Infarction, Residence Characteristics	t1 t2 t7 t9
	⑬ 추출된 초록 핵심어(AB)	Confidence Intervals, Heart, Heart Diseases, International Classification of Diseases, Reproducibility of Results, Women	t3 t4 t5 t6 t8 t10

여 문헌의 위상 정보를 반영하는 세 가지 중앙성인 연결정도(degree centrality), 사이중앙성(betweenness centrality), 인접중앙성(closeness centrality)을 네트워크 자질로 추출하여 활용하였다(Bonacich, 2007).

- **내용 자질:** 서지 자질이 문헌의 서지적 특징을 반영하기 위한 자질이라면 내용 자질은 문헌의 내용적 특성을 반영하기 위해 추출되는 자질이다. 기존의 선행연구에서 내용적 특성을 반영하기 위해 단순히 논문의 표제와 초록에 출현한 단어를 사용하였던 것(Castillo, Donato, & Gionis, 2007; Fu & Aliferis, 2008)과 달리 본 연구에서는 키워드 추출 기법(keyphrase extraction algorithm, KEA)을 활용하여 핵심어(구)를 추출한다. KEA는 어휘 기반 핵심어 자동 추출 알고리즘이다. 이 기법은 시스템에서 제공하는 사전과 비교하여 문헌에 등장한 핵심어를 식별하는 모델을 생성하고 새로운 문헌이 입력될 때 핵심어를 자동으로 추출하는 기계학습 기법을 사용한다(Witten, Paynter, Frank, Gutwin, & Nevill-Manning, 1999). 핵심어(구)의 추출을 위해서 MAUI 인덱서[3]를 활용한다. MAUI는 여러가지 온톨로지

3) http://code.google.com/p/maui-indexer/

와 매핑하여 핵심어(구)를 색인하여 추출한다. 본 연구의 주제분야는 전염병학이므로 미국 국립의학도서관(National Library of Medicine, NLM)에서 제공하는 통제어인 의학주제명표 MeSH(Medical Subject Headings)를 매핑하여 핵심어를 추출한다.

- **변수 값 변환:** 추출된 서지 자질과 네트워크 자질, 내용 자질은 팩터그래프 기반 추론모델 알고리즘에 입력 가능한 형태로 표현되어야 한다. 각 자질들은 확률변수노드로 표현되고 그 값은 추출된 자질 값을 정규화된 값으로 변환하여 처리한다. 팩터그래프에서 사용할 수 있는 변수의 유형은 이산형(discrete type)과 연속형(continuous type)이다. 이산형은 정수, 스트링, 배열 등의 데이터를 의미하며 연속형은 실수를 의미한다. 따라서 서지 자질(저자 수, 저자의 기관 수, 프로시딩 유무, 저자 키워드 수, 펀딩 유무, 참고문헌 수, 페이지 수, 저널 영향력 지수)과 네트워크 자질(연결정도, 사이중앙성, 인접중앙성)은 연속형 확률변수 값으로 표현하고, 내용 자질(추출된 표제 핵심어, 초록 핵심어)은 이산형 스트링으로 표현하였다. 서지 자질과 네트워크 자질을 0-1 사이의 값을 가지는 정규화된 값으로 변환하기 위해서 각 자질의 문헌집합 내의 최대값으로 나누어 실수로 표현한다. 네트워크 자질 중에 인접중앙성 값은 이미 정규화되어 나온 값이므로 그대로 사용한다. 〈표 1〉은 팩터그래프 추론모델에 사용된 자질들과 그 자질들을 팩터그래프 기반 추론모델에 사용한 변수 값으로 변환한 예시이다.

4. 팩터그래프 기반 연구전선 구축 결과

〈표 2〉는 서지결합법으로 구축한 베이스라인 연구전선그룹과 팩터그래프 기반 연구전선그룹의 문헌 네트워크 특성을 비교한 것이다.

〈표 2〉 서지결합 기반 연구전선과 팩터그래프 기반 연구전선그룹 비교

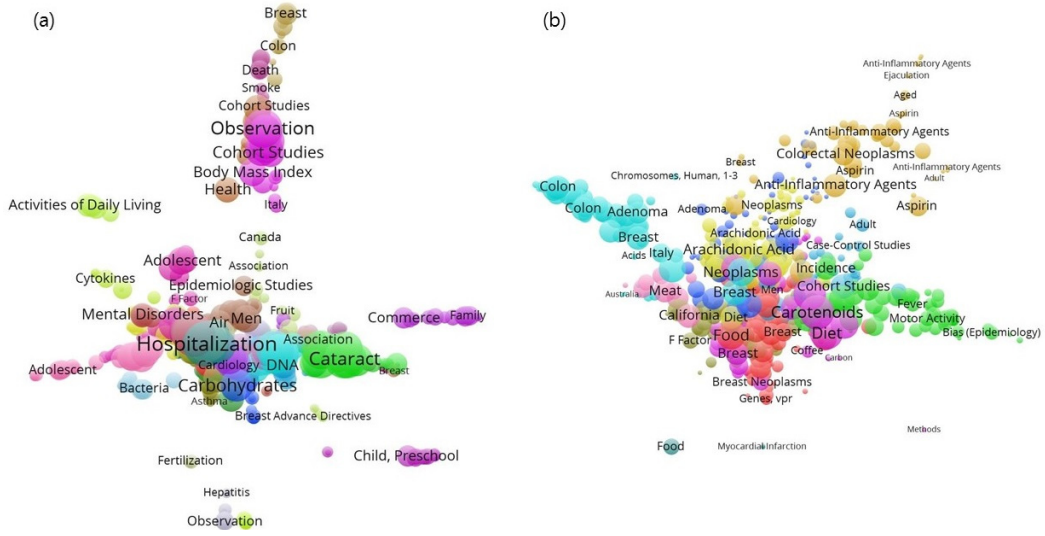
구분		평균 연결정도	평균 사이중앙성	하위 군집 수
서지결합	기간 1	3.16	0.00000064	2,762
	기간 2	3.15	0.00000030	2,854
	기간 3	2.95	0.00000020	3,058
	기간 4	2.09	0.00000024	3,091
	기간 5	2.97	0.00000017	3,196
	평균	2.86	0.00000031	2,992
팩터그래프	기간 1	53.32	0.00017193	195
	기간 2	51.63	0.00017751	174
	기간 3	46.83	0.00019329	157
	기간 4	42.30	0.00022033	132
	기간 5	10.50	0.00008718	335
	평균	40.92	0.00017005	199

각 연구전선그룹의 평균 연결정도와 하위군집 수를 비교하였다. 〈표 2〉에서 보는 바와 같이 각 연구전선그룹을 구성하고 있는 문헌들에는 큰 차이가 있음을 알 수 있다. 베이스라인 연구전선과 비교했을 때 팩터그래프 기반 연구전선은 평균 연결정도가 약 14.3배 정도 크다(2.86, 40.92). 또한 평균 사이중앙성도 팩터그래프 기반 연구전선에서 더 크게 나타나 팩터그래프 기반 연구전선그룹이 서지결합 기반 연구전선그룹 보다 문헌 간의 직접 연결정도가 강하며 연결 관계에 있지 않은 두개의 문헌을 연결시키는 매개정도 또한 강한 집단으로 나타났다. 그리고 두 집단의 하위군집 수 또한 큰 차이를 보인다. 서지결합 기반 연구전선그룹은 평균 하위군집의 수가 2,992개이고 팩터그래프 기반 연구전선그룹은 평균 하위군집 수가 199개이다. 팩터그래프 기반 연구전선그룹이 더 유사한 문헌들의 출현이 많음을 알 수 있다.

〈표 3〉은 서지결합 기반 연구전선과 팩터그래프 기반 연구전선의 최고 연결정도를 가진

〈표 3〉 베이스라인과 팩터그래프 기반 연구전선의 대표 문헌 특징 비교

기간	특징	베이스라인	팩터그래프
기간 1	최고 연결정도 문헌 표제	Randomised by (your) god: robust inference from an observational study design	Methods for pooling results of epidemiologic studies - The pooling project of prospective studies of diet and cancer
	최고 연결정도	54	453
	피인용 빈도	19	106
기간 2	최고 연결정도 문헌 표제	Reduction in incidence of nosocomial methicillin - resistant Staphylococcus aureus (MRSA) infection in an intensive care unit: Role of treatment with Mupirocin ointment and chlorhexidine baths for nasal carriers of MRSA	Increased risk for cervical disease progression of French women infected with the human papillomavirus type 16 E6-350G variant
	최고 연결정도	53	439
	피인용 빈도	59	106
기간 3	최고 연결정도 문헌 표제	Haplotype analysis of common vitamin D receptor variants and colon and rectal cancers	Methods for pooling results of epidemiologic studies - The pooling project of prospective studies of diet and cancer
	최고 연결정도	44	377
	피인용 빈도	39	106
기간 4	최고 연결정도 문헌 표제	Relationship of alcohol consumption and type of alcoholic beverage consumed with plasma lipid levels: Differences between Whites and African Americans of the ARIC study	Statistical methods for the time-to-event analysis of individual participant data from multiple epidemiological studies
	최고 연결정도	27	238
	피인용 빈도	18	28
기간 5	최고 연결정도 문헌 표제	Vitamin D and Melanoma	Correcting "Winner's Curse" in Odds Ratios from Genomewide Association Findings for Major Complex Human Diseases
	최고 연결정도	50	96
	피인용 빈도	23	30



〈그림 5〉 서지결합 기반(a)과 팩터그래프 기반(b)의 하위 대표군집 비교: 기간 1

문헌의 특징을 비교한 것이다. 모든 기간에서 팩터그래프 기반 연구전선에 속한 최고 연결정도 문헌이 베이스라인의 최고 연결정도 문헌보다 연결정도와 실제 피인용빈도가 높음을 알 수 있다. 결과적으로 팩터그래프 기반의 연구전선이 포함하고 있는 피인용빈도가 높은 문헌을 서지결합 기반 연구전선은 포함을 하고 있지 못함을 알 수 있다.

〈그림 5〉는 기간 1의 연구전선 문헌 네트워크에서 연결정도가 가장 큰 문헌(표 3)이 속하여져 있는 하위군집을 대표 하위군집으로 정하고 VOSviewer(Van Eck, Waltman, Dekker, & Van Den Berg, 2010)로 시각화하여 보여준 것이다.

기간 1의 서지결합 기반 연구전선그룹(그림 5a)의 대표 하위군집은 Breast(유방), Association(연관), Adult(성인), Activities of Daily Living(일상생활), F Factor(F 인자), Health(보건), Case-Control Studies(사례-대조군 연구), Adolescent(청소년), Cohort Studies(집단연구), Body Mass Index(체질량 지수), Birth Order(출생 순서), Data Collection(데이터 수집), Disease(질병), Communication(교류), Aged(노화), Arachidonic Acid(아라키돈산), Incidence(사례), Neoplasms(종양), Air(공기), Breast Neoplasms(유방 종양) 등의 단어들이 상위 빈도 핵심어로 나타난다. 반면, 팩터그래프 기반 연구전선그룹의 대표 하위군집에서 나타난 상위 빈도 핵심어는 Breast(유방), Arachidonic Acid(아라키돈산), Cohort Studies(집단연구), Case-Control Studies(사례-대조군 연구), Colorectal Neoplasms(결장 종양), Adult(성인), Association(연관), Adenoma(아데노마, 선종), Antioxidants(산화/노화 방지제), Blood(혈액), Carotenoids(카로티노이드), Colon(결장), Aged(노화), Alcohol Drinking(음주), Diet(식이요법), F Factor(F 인자), Calcium(칼슘), Food(식량), Fruit(과일),

Neoplasms(종양)이다(그림 5b). 서지결합 기반 연구전선(그림 5a)과 팩터그래프 기반 연구전선(그림 5b)은 하위 대표군집에서도 크게 차이가 있음을 알 수 있다.

5. 결 론

본 연구는 핵심 연구논문으로 발전할 가능성이 높은 문헌들로 구성된 연구전선을 구축하기 위하여 진행된 것으로, 확률 그래프 모델인 팩터그래프를 기반으로 하여 연구전선 추론모델을 생성하고 이 모델을 적용해서 특정 문헌이 연구전선그룹에 포함될 것인지 여부를 예측하였다. 어떤 한 문헌이 연구전선그룹에 포함된다는 것은 그 문헌이 미래에 더 높은 피인용빈도를 보이는 핵심논문이 될 가능성이 있다는 것을 의미한다.

본 연구의 의의는 세 가지의 복합적인 자질세트를 적용하여 연구전선을 추론하는 모델을 제시하였다는 것이다. 전통적인 계량서지학적 접근기법을 적용하여 연구전선을 파악하는 것이 아니라 본 연구에서 제시한 팩터그래프 기반 연구전선 추론모델은 향후 활발한 연구영역으로 발전 가능성이 있는 문헌들로 구성된 연구전선을 추론하는 것이다. 그러므로 신규 발행되는 문헌에 대하여 연구전선그룹에 속할 가능성을 예측할 수 있고, 더 나아가 실제 활발한 인용활동이 발생할 가능성이 높은 연구전선을 예측할 수 있기 때문에 조기에 관련 연구영역에 대한 이해관계자들의 투자와 의사결정을 지원할 수 있다는 장점을 가지고 있다. 또한 추가적으로 핵심 연구분야로 발전할 가능성이 높은 연구논문 또는 연구자를 미리 예측하는 시스템과의 연계가 가능하다는 장점도 가지고 있으므로 이 모델을 활용할 경우 학계와 산업계 및 정부에 큰 유익을 가져다 줄 수 있을 것으로 보인다.

그러나 이러한 장점들에도 불구하고 본 연구에는 다음과 같은 한계점이 있다. 연구전선을 서로 인용강도가 높으면서 발전단계에 있는 연구분야로 정의하고 피인용빈도가 높은 문헌을 핵심 연구영역으로 정의하였기 때문에 피인용빈도가 낮거나 아직 발생하지 않은 신흥 주제영역은 포함되지 못할 수도 있다는 한계점을 가지고 있다. 이는 각 학문분야의 고유성으로 인해 발생되는 인용행위의 차이를 반영하지 못하고 있다는 제한적인 속성에서 비롯한 것이다.

후속연구로는 연구전선을 예측함에 있어 본 연구에서 사용했던 여덟 가지 서지 자질들 중에 어떤 자질이 가장 좋은 예측성능을 보이는지 심층적으로 분석하는 연구를 진행할 예정이다. 본 연구는 생의학분야 중 전염병학의 문헌집단을 기반으로 추론모델을 생성하여 연구전선을 구축하였는데, 인문사회과학 계열의 문헌집단에도 팩터그래프 기반 연구전선 추론모델을 적용해 봄으로써 이 분야에도 확률기반 팩터그래프의 활용이 가능한지 그 가능성을 추가적으로 검증하고 타학문분야의 연구전선의 특성을 분석하는 연구 또한 필요하다.

참 고 문 헌

김조아, 이재윤 (2016). 인용 이미지 구축자 프로파일링을 이용한 국내 여성학 분야 연구 전선 분석. 정보관리학회지, 33(2), 201-225. http://dx.doi.org/10.3743/KOSIM.2016.33.2.201

서은경, 유소영 (2013). 국내 정보학분야 연구동향 분석: 2000-2011. 정보관리학회지, 30(4), 215-239. http://dx.doi.org/10.3743/KOSIM.2013.30.4.215

이재윤 (2015). 문헌동시인용 분석을 통한 한국 문헌정보학의 연구 전선 파악. 정보관리학회지, 32(4), 77-106. http://dx.doi.org/10.3743/KOSIM.2015.32.4.077

조재인 (2011). 네트워크 텍스트 분석을 통한 문헌정보학 최근 연구 경향 분석. 정보관리학회지, 28(4), 65-83. https://doi.org/10.3743/kosim.2011.28.4.065

Bonacich, P. (2007). Some unique properties of eigenvector centrality. Social Networks, 29(4), 555-564. http://dx.doi.org/10.1016/j.socnet.2007.04.002

Boyack, K., & Klavans, R. (2010). Co−citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? Journal of the American Society for Information Science and Technology, 61(12), 2389-2404. http://dx.doi.org/10.1002/asi.21419

Castillo, C., Donato, D., & Gionis, A. (2007). Estimating number of citations using author reputation. Proceedings of the String Processing and Information Retrieval, 107-117. https://doi.org/10.1007/978-3-540-75530-2_10

De Solla Price, D. (1965). Networks of scientific papers. Science, 149(4683), 510-515. https://doi.org/10.1126/science.149.3683.510

Frey, B. (1998). Graphical models for machine learning and digital communication. Cambridge, Mass: The MIT Press.

Fu, L., & Aliferis, C. (2008). Models for predicting and explaining citation count of biomedical articles. Proceedings of the American Medical Informatics Association (AMIA), 1, 222-226.

Jaffe, A., & Trajtenberg, M. (1996). Flows of knowledge from universities and federal laboratories: Modeling the flow of patent citations over time and across institutional and geographic boundaries. Proceedings of the National Academy of Sciences, 93(23), 12671-12677.

Jarneving, B. (2007). Bibliographic coupling and its application to research-front and other core documents. Journal of Informetrics, 1(4), 287-307. http://dx.doi.org/10.1016/j.joi.2007.07.004

Jordan, M. (2004). Graphical models. Statistical Science, 19(1), 140-155.

https://doi.org/10.1214/088342304000000026

Kollar, D., & Friedman, N. (2009). Probabilistic graphical models: Principles and techniques. Cambridge, Mass: The MIT Press.

Kschischang, F., Frey, B., & Loeliger, H. (2001). Factor graphs and the sum-product algorithm. IEEE Transactions on Information Theory, 47(2), 498-519.

Loeliger, H. (2004). An introduction to factor graphs. IEEE Signal Processing Magazine, 21(1), 28-41. http://dx.doi.org/10.1109/MSP.2004.1267047

McCain, K., & Turner, K. (1989). Citation context analysis and aging patterns of journal articles in molecular genetics. Scientometrics, 17(1), 127-163. http://dx.doi.org/10.1007/BF02017729

Narin, F., & Hamilton, K. (1996). Bibliometric performance measures. Scientometrics, 36(3), 293-310. http://dx.doi.org/10.1007/BF02129596

Persson, O. (1994). The intellectual base and research fronts of JASIS 1986-1990. Journal of the American Society for Information Science, 45(1), 31-38.

Porta, M. (2014). A Dictionary of Epidemiology (6th ed.). New York: Oxford University Press.

Shen, H., & Coughlan, J. (2007). Grouping using factor graphs: An approach for finding text with a camera phone. Graph-Based Representations in Pattern Recognition, 4538, 394-403. http://dx.doi.org/10.1007/978-3-540-72903-7_36

Shibata, N., Kajikawa, Y., & Matsushima, K. (2007). Topological analysis of citation networks to discover the future core articles. Journal of the American Society for Information Science and Technology, 58(6), 872-882. http://dx.doi.org/10.1002/asi.20529

Shibata, N., Kajikawa, Y., Takeda, Y., & Matsushima, K. (2009). Comparative study on methods of detecting research fronts using different types of citation. Journal of the American Society for Information Science and Technology, 60(3), 571-580. http://dx.doi.org/10.1002/asi.20994

Small, H. (1973). Co−citation in the scientific literature: A new measure of the relationship between two documents. Journal of the American Society for information Science, 24(4), 265-269. https://doi.org/10.1002/asi.4630240406

Small, H., Sweeney, E., & Greenlee, E. (1985). Clustering the Science Citation Index using co-citations. II. Mapping science. Scientometrics, 8(5), 321-340. https://doi.org/10.1007/bf02018057

Song, M., & Kim, S. Y. (2013). Detecting the knowledge structure of bioinformatics by mining full-text collections. Scientometrics, 96(1), 183-201.

http://dx.doi.org/10.1007/s11192-012-0900-9

Sun, Y., Deng, H., & Han, J. (2012). Probabilistic models for text mining. Mining Text Data, 259-295. https://doi.org/10.1007/978-1-4614-3223-4_8

Sutton, C., & McCallum, A. (2007). An introduction to conditional random fields for relational learning. Introduction to Statistical Relational Learning, 93, 142-146.

Upham, S., & Small, H. (2010). Emerging research fronts in science and technology: Patterns of new knowledge development. Scientometrics, 83(1), 15-38. http://dx.doi.org/10.1007/s11192-009-0051-9

Van Eck, N., Waltman, L., Dekker, R., & Van Den Berg, J. (2010). A comparison of two techniques for bibliometric mapping: Multidimensional scaling and VOS. Journal of the American Society for Information Science and Technology, 61(12), 2405-2416. http://dx.doi.org/10.1002/asi.21421

Weiss, Y., & Freeman, W. (2001). On the optimality of solutions of the max-product belief-propagation algorithm in arbitrary graphs. IEEE Transactions on Information Theory, 47(2), 736-744. http://dx.doi.org/10.1109/18.910585

Witten, I. H., Paynter, G. W., Frank, E., Gutwin, C., & Nevill-Manning, C. G. (1999). KEA: Practical automatic keyphrase extraction. Proceedings of the 4th ACM conference on Digital Libraries, 254-255. https://doi.org/10.1145/313238.313437

Yeh, Y., Breeden, K., Yang, L., Fisher, M., & Hanrahan, P. (2013). Synthesis of tiled patterns using factor graphs. ACM Transactions on Graphics (TOG), 32(1), 3. http://dx.doi.org/10.1145/2421636.2421639

• 국문 참고문헌에 대한 영문 표기
(English translation of references written in Korean)

Cho, Jane (2011). A study for research area of library and information science by network text analysis. Journal of the Korean Society for Information Management, 28(4), 65-83. https://doi.org/10.3743/kosim.2011.28.4.065

Kim, Jo-Ah, & Lee, Jae Yun (2016). Analyzing the research fronts of women's studies in Korea using citation image makers profiling. Journal of the Korean Society for Information Management, 33(2), 201-225. http://dx.doi.orf/10.3743/KOSIM.2016.33.2.201

Lee, Jae Yun (2015). Identifying the research fronts in Korean library and information science by document co-citation analysis. Journal of the Korean Society for Information Management,

32(4), 77-106. http://dx.doi.orf/10.3743/KOSIM.2015.32.4.077

Seo, Eun-Gyoung, & Yu, So-Young (2013). Detecting research trends in Korean information science research, 2000-2011. Journal of the Korean Society for Information Management, 30(4), 215-239. http://dx.doi.orf/10.3743/KOSIM.2013.30.4.215

[부록 1] 수집대상 저널명 및 발행 논문 수

번호	저널명	논문 수 (건)
1	Am. J. Epidemiol	2,216
2	Ann. Epidemiol	932
3	Cancer Epidemiol	288
4	Cancer Epidemiol. Biomarkers Prev	3,001
5	Community Dentist. Oral Epidemiol	461
6	Epidemiol. Infect	1,396
7	Epidemiol. Psychiatr. Sci	29
8	Epidemiol. Rev	96
9	Epidemiology	762
10	Eur. J. Epidemiol	794
11	Genet. Epidemiol	545
12	Infect. Control Hosp. Epidemiol	1,575
13	Infect. Genet. Evol	802
14	Int. J. Epidemiol	1,038
15	J. Clin. Epidemiol	1,104
16	J. Epidemiol	400
17	J. Epidemiol. Community Health	1,344
18	Microb. Drug Resist	232
19	Ophthalmic Epidemiol	370
20	Paediatr. Perinat. Epidemiol	525
21	Soc. Psychiatry Psychiatr. Epidemiol	1,067
	총 논문 수	18,977