자아 중심 네트워크 분석과 동적 인용 네트워크를 활용한 토픽모델링 기반 연구동향 분석에 관한 연구*

Combining Ego-centric Network Analysis and Dynamic Citation Network Analysis to Topic Modeling for Characterizing Research Trends

유소영 (So-Young Yu)**

초 록

이 연구에서는 토픽 모델링 결과 해석의 용이성을 위하여, 동적 인용 네트워크를 활용하여 LDA 기반 토픽 모델링의 토픽 수를 설정하고 중복 배치된 주요 키워드를 자아 중심 네트워크 분석을 통해 재배치하여 제시하는 방법을 제안하였다. 'White LED' 두 분야의 논문 데이터를 이용하여 분석한 결과, 동적 인용 네트워크 분석을 통해 형성된 분석대상 문헌집단에 혼잡도에 따른 토픽수를 사용하고 중복 분류된 토픽 내 주요 키워드를 자아중심 네트워크 분석 기법을 적용하여 재배치한 결과가 토픽 간의 중복도가 가장 낮은 것으로 나타났다. 따라서 동적 인용 네트워크 및 자아 중심 네트워크 분석을 적용함으로써 토픽모델링에 의한 분석 결과를 보완하는 다면적인 연구 동향 분석이 가능할 것으로 보인다.

ABSTRACT

The combined approach of using ego-centric network analysis and dynamic citation network analysis for refining the result of LDA-based topic modeling was suggested and examined in this study. Tow datasets were constructed by collecting Web of Science bibliographic records of White LED and topic modeling was performed by setting a different number of topics on each dataset. The multi-assigned top keywords of each topic were re-assigned to one specific topic by applying an ego-centric network analysis algorithm. It was found that the topical cohesion of the result of topic modeling with the number of topic corresponding to the lowest value of perplexity to the dataset extracted by SPLC network analysis was the strongest with the best values of internal clustering evaluation indices. Furthermore, it demonstrates the possibility of developing the suggested approach as a method of multi-faceted research trend detection.

키워드: 자아중심 네트워크, 토픽 모델링
Ego-centric network, SPLC, Topic Modeling, DBI, Index S

* 이 논문은 2013학년도 한남대학교 학술연구조성비 지원에 의하여 연구되었음.
** 한남대학교 문헌정보학과 조교수(soyoungyu201@gmail.com)
■ 논문접수일자 : 2015년 2월 24일 ■ 최초심사일자 : 2015년 2월 26일 ■ 게재확정일자 : 2015년 3월 9일
■ 정보관리학회지, 32(1), 153-169, 2015. 〔http://dx.doi.org/10.3743/KOSIM.2015.32.1.153〕

1. 연구의 배경 및 목적

1.1 연구의 배경

다양한 연구동향 분석 방법 중 인용 네트워크와 단어 동시 출현 네트워크를 사용하는 계량정보학적 접근 방식은 데이터에 기반한 양적인 접근 방법이라고 할 수 있다. 따라서 분석 대상이 되는 연구 분야에 대한 주제 전문 지식을 충분히 갖추고 있지 않더라도 분석 데이터의 특성을 이해하면 분석 결과를 생산할 수 있다. 그리고 이와 같은 분석 결과는 주제 분야 전문가의 해석이나 확인을 통해 보다 정교화되며, 실제로 다양한 연구동향 분석에는 계량적인 분석 방법과 전문가의 질적 분석을 혼합하여 사용하고 있는 추세이다(정우성, 양현재, 2013). 따라서 연구동향 분석의 전반적 경향을 살펴볼 때, 계량정보학적인 분석 결과는 해당 주제 분야 전문가가 개인적인 주제 전문성을 확장하여 해석할 수 있도록 해석이 용이한 형태로 분석 결과가 제시되어야 할 것으로 보인다. 그러므로 연구 동향 분석 기법 개발과 함께 분석 결과를 제시하는 방법에 대한 고민이 함께 이루어져야 할 것으로 생각된다.

최근 들어 연구 동향 및 지적 구조 분석에 자주 적용되는 토픽 모델링 결과는 주요 용어들의 클러스터 형태로 제시되고 있다. 따라서 적절한 수의 용어 클러스터를 설정하기 위한 방법과 관련된 연구들이 수행되었는데, 최근 이와 관련된 한 연구에서는 여러 가지 경우의 토픽 수에 따른 토픽모델링을 실시한 후 용어분류의 정확성이 가장 높은 결과를 선택하는 방법(Griffiths & Steyvers, 2004), 비교적 많은 수의 토픽을 설정하여 LDA 분석을 실시한 후 유사한 토픽들을 결합하여 최종 결과를 생성하는 방법(Song, 2010; Yu, 2014), 그리고 토픽 모델링의 기반이 되는 확률모델을 최적화하여 적절한 수의 토픽을 비모수적으로 추정하여 토픽 모델링을 수행하는 방법(Teh, Jordan, Beal, & Beli, 2006)의 세 가지 유형으로 크게 분류하였다(Ding & Chen, 2014).

특정 주제 분야에 대한 이해가 충분히 없는 상태에서 토픽의 수를 정확히 정하기에는 어려움이 있기 때문에(Ding & Chen, 2014; Mccallum, Mimno, & Wallach, 2009; Ramage, Rosen, Chuang, Manning, & McFarland, 2009b), HDP(Hierarchical Dirichlet Process, Teh, Jordan, Beal, & Beli, 2006)나 LLDA(Ramage, Hall, Nallapati, & Manning, 2009a)와 같은 정교한 토픽모델링 기법을 개발하는 것과 함께, 토픽 모델링의 결과를 정련하여 제시함으로써 이용자 해석의 용이성을 높이는 방법도 다양한 시도가 필요하다고 할 수 있다.

1.2 연구의 목적 및 방법

따라서 이 연구에서는 토픽 모델링의 토픽 수를 결정하고 토픽 모델링 결과를 정련하는데 동적 인용 네트워크 분석과 자아중심 네트워크를 적용하여 연구 동향을 다면적으로 분석하였으며 적용된 네트워크 분석기법의 성능 평가를 수행 하였다. 구체적인 연구 문제는 다음과 같다.

첫째, 동적 인용 네트워크 분석 결과를 토픽 모델링 내 토픽 수 선정 기준으로 사용함으로써 단순 토픽 모델링을 적용한 것과 다른 분석 결

과를 도출할 수 있을 것이다. 둘째, 토픽 모델링 후 토픽 내 주요 용어를 네트워크 분석 기법을 적용하여 재배치하였을 때, 각 토픽의 주제적 응집성이 증가할 것이다.

이와 같은 연구 문제를 살펴보기 위하여 이 연구에서는 적절한 토픽 수 선정에 동적 인용 네트워크 분석 기법과 전통적인 자연언어처리 개념인 혼잡도(perplexity)를 적용하여 비교하였다. 그리고 LDA 기반 토픽 모델링 결과로 나타난 토픽 클러스터 내 주요 키워드를 자아 중심 네트워크 분석 기법을 적용하여 재배치함으로써 토픽 모델링의 결과를 다양하게 제시하여 비교하고 각 토픽의 주제적 응집성을 높일 수 있는지 클러스터링 내적 품질평가를 통해 살펴보았다.

2. 네트워크 분석을 결합한 토픽 모델링 연구

인용 네트워크 분석을 포함한 네트워크 분석과 토픽 모델링 결합 관련 연구들은 추천 시스템 및 연구동향 파악 연구 및 응용 분야에서 최근 들어 활발하게 이루어지고 있다(Ding & Chen, 2014; Huang, Wu, Liang, Mitra, & Giles, 2015; Jiang, 2015; Park & Song, 2014; Seo & Yu, 2013; Yu, 2014).

이 외에도 주요 연구동향 파악을 위해 인용 네트워크 분석과 텍스트 마이닝 기법을 적용하는 실무 사례들을 국가 단위 연구성과 분석에서 살펴볼 수 있다. 일본 NISTEP(National Institute of Science and Technology Policy)의 Science Map은 2002년부터 매 짝수년마다 'NISTEP REPORT' 형태로 발표되고 있는데, 발표 시기를 기준으로 최근 6년간의 주요 연구 분야를 동시인용 네트워크 분석을 기본으로 텍스트 마이닝 기법을 조합하여 파악하고 시각화하여 제시하고 있다. 여기서 주요 연구 분야(hot research topic)는 전통적인 학문 분류에 의한 연구 분야가 아닌 비교적 최근 급부상하고 있는 연구 분야를 조작적으로 정의한 것이다(Saka & Igami, 2014). NISTEP의 Science map은 Web of Science에 색인된 최근 6년치 논문 중 인용 빈도 기준 상위 1% 논문(core papers)을 이용하여 동시인용 네트워크를 형성하여 이를 시각화한 것이다. Science Map을 위한 core papers 간의 동시인용 네트워크 추출 시, 인용 빈도 기준 상위 10%의 인용 문헌(citing papers)만을 이용함으로써 주요 연구 경향만을 분석에 반영할 수 있도록 하고 있다. 그리고 텍스트 마이닝 기법은 주요 연구 분야의 주요 키워드를 해당 분야의 논문 제목과 초록에서 추출하는데 사용한다. 최근에는 축척된 Science Map 분석 결과를 활용하여 유망 및 지속 가능 연구 분야 분석도 시도하고 있다(Saka & Igami, 2014). 이와 유사하게 한국의 한국과학기술정보원은 매년 미래유망기술을 선정하여 발표하고 있는데 여기에 논문 및 특허를 대상으로 한 계량분석 결과를 활용하고 있다(한국과학기술정보연구원, 2013, 2014).

이를 통해 살펴볼 때, 현재 연구동향 및 주요 연구 분야를 파악하는데 의사결정자 또는 해당 주제전문가의 해석과 판단을 도와줄 수 있는 형태로 분석 결과를 제시하기 위하여 다양한 기법을 적용하고 주요 키워드 및 주요 논문, 그리고 주요 저자들을 선정하여 제공하는 것을

알 수 있다. 또한 보다 해석이 용이한 형태로 분석결과를 도출하기 위하여 인용 빈도 등의 기준을 활용하여 데이터 분석의 정확성을 높이고자 하는 것을 알 수 있다. 따라서 이 연구에서는 토픽 모델링의 결과로 제시되는 토픽 내 주요 키워드의 중복 출현을 감소시킴으로써 해석의 용이성을 도모하고자, 토픽 수 선정 및 토픽 모델링 후 키워드 간의 동시출현 유사도에 의한 주요 키워드의 재배치에 따른 토픽 간 유사도 변화를 실험적으로 살펴보고자 하였다.

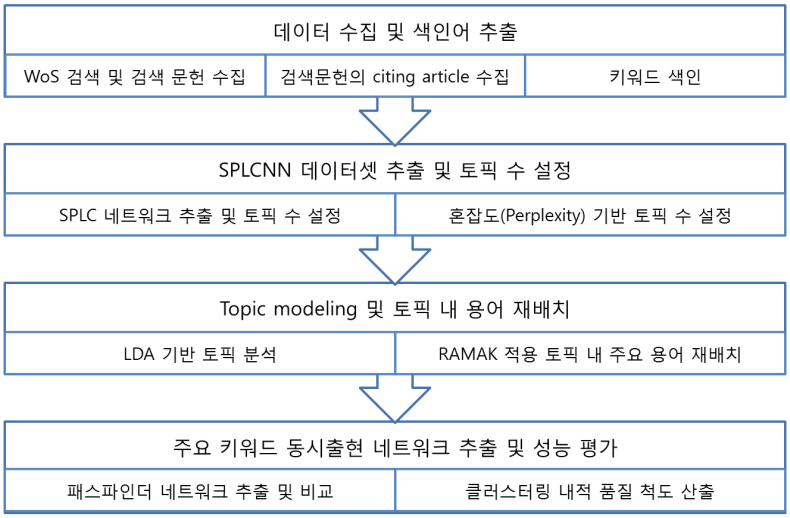
3. 연구 방법

3.1 연구 개요

이 연구는 데이터 수집 및 SPLC 분석을 적용한 동적 인용 네트워크 분석, 토픽 수 선정 및 LDA 기반 토픽 모델링 수행, 자아 중심 네트워크를 이용한 토픽 내 주요 키워드 재배치, 주요 키워드의 단어 동시출현 네트워크 시각화 및 내적 품질 평가의 순서로 수행되었다. 연구 개요도는 〈그림 1〉과 같다.

3.2 데이터 수집 및 색인어 추출

이 연구에서는 SPLC 문헌 네트워크를 추출하기 위하여 Web of Science에 색인된 'White LED' 분야의 논문 데이터를 이용하였다. 2014년 7월 1일 ~ 2014년 7월 14일 사이에 LED 분야 전문가가 생성한 검색식(이재윤, 김판준, 강대신, 김희정, 유소영, 이우형, 2011)을 이용하여 검색된 문헌 레코드들과 이를 인용하는 문헌 레코드 중 동일 출판년도별 수집 당시 현재 인용빈도 기준 상위 10%에 해당하는 레코드를 수집하였다(유소영, 2013). 수집된 7,427개 논문으로 전체 문헌 데이터셋(이하 ALL)을 구성하였다.



〈그림 1〉 연구 개요

토픽모델링 및 단어 동시출현 네트워크 생성을 위한 키워드 색인은 제목, 저자 키워드, 그리고 Keyword Plus®[1]를 대상으로 하였다. 키워드 색인에는 Stanford NLP에서 개발한 Stanford Parser를 이용하였으며 불용어는 제거하고 스테밍은 실시하지 않았다. 색인 결과 ALL 데이터셋 내 7,427개 문헌에서는 3,058개의 고유 키워드가 색인되었다.

3.3 SPLC 분석을 통한 SPLCNN 데이터셋 추출 및 토픽 수 설정

비교 데이터셋으로 ALL의 인용 네트워크에서 파악된 25개의 SPLC 문헌 및 이들을 직접 인용하거나 인용하고 있는 최근접 이웃(Nearest Neighbor) 문헌을 확인하여 850개 문헌 레코드로 구성된 SPLCNN 데이터셋을 추출하였다(이하 SPLCNN).

데이터셋 추출에는 HistCite™와 PAJEK v3.12를 이용하여 히스토리오그래프(Garfield, 2001.9.19; 2001.11.27)와 Search Path Link Count(이하 SPLC, Hummon, & Doreian, 1989) 기법을 적용하였다(de Nooy et al., 2011). 또한 ALL 데이터셋과 동일한 색인방법을 적용하여 562개의 고유 키워드를 색인하였다.

SPLC 네트워크 분석 결과를 활용하여 SPLC 네트워크 내 출현한 주요 문헌 수 25를 토픽수로 선정하였다. 그리고, 비교를 위하여 10번의 사전실험을 통해 혼잡도가 낮게 나타난 토픽수를 분석에 사용하였다. ALL 데이터셋은 5개 토픽, 그리고 SPLCNN 데이터셋은 2개 토픽을 사용하는 것이 혼잡도가 가장 낮은 것으로 나타났다. 사전실험에는 Stanford TMT(Topic Modeling Toolbox) version 0.4.0을 사용하였다(Ramage, Rosen, Chuang, Manning, & McFarland, 2009b).

3.4 LDA 실시 및 토픽 내 주요 용어 재배치

토픽모델링은 LDA 기법에 기반하였으며, Stanford TMT version 0.4.0을 사용하였다. 총 4개의 LDA 기반 토픽 분석이 실시되었는데, ALL 데이터셋에 SPLC 분석에서 파악된 25개 토픽을 적용하여 분석한 모델링(이하 SA), ALL 데이터셋에 혼잡도 분석에서 파악된 5개 토픽을 적용하여 분석한 모델링(이하 PA), SPLCNN 데이터셋에 SPLC 분석에서 파악된 25개 토픽을 적용하여 분석한 모델링(이하 SS), 그리고 SPLCNN 데이터셋에 혼잡도 분석에서 파악된 2개 토픽을 적용하여 분석한 모델링(이하 SP)이 이에 해당한다. 토픽모델링 후, 각 LDA 기반 토픽 분석 결과에 자아중심 네트워크 분석에 기반한 RAMAK(Re-Assigning Multi-Assigned Keyword, Yu, 2014)을 적용하여 두 개 이상의 토픽 클러스터에 적재된 주요 용어를 재배치하였다.

RAMAK은 여러 토픽에 동시에 상위 키워드로 출현한 단어의 토픽을 하나로 결정하여 토픽 모델링 결과의 해석이 쉽도록 제시하기 위하여 고안되었다(Yu, 2014). 따라서 이는 토픽 모델링 결과를 제시하기 위해 추출한 단어 동시 출현 네트워크에 적용된다. 이 네트워크

1) Thomson Reuters가 부여한 것으로, 색인되는 문헌 레코드의 저자가 인용한 문헌 제목을 기반으로 생성한다. 〈2014.8.10. 검색. Web of Science 핵심 컬렉션 도움말〉

의 노드는 토픽 모델링 결과로 제시된 상위 키워드이고, 각 단어는 한 개 이상의 토픽 번호를 가지게 된다. 그리고 각 키워드들 간에는 유사도 공식에 의해 동시 출현 강도가 계산된다. RAMAK에서는 단어 동시출현 네트워크 내에서 두 개 이상의 토픽 번호를 가지고 있는 단어의 토픽을 결정하는데, 그 단어와 동시출현한 출현 키워드들의 토픽을 고려하도록 설계되었다. 이 연구에서는 RAMAK의 평균 동시출현 가중치 대신 동시출현 가중치의 총합을 사용하였다. 왜냐하면 평균은 동시출현 키워드 수에 의한 정규화의 의미를 갖는데, 동시출현 키워드가 많고 적음은 Ego가 갖는 특성이므로 정규화할 필요가 없다고 판단되었기 때문이다.

3.5 토픽 내 주요 키워드 동시출현 네트워크 추출 및 성능 평가

주요 키워드의 재배치 후 토픽 내 주요 키워드 동시출현 코사인 유사도를 이용하여 패스파인더 네트워크를 추출하여 주요 연결 구조를 비교함으로써 문헌집단 및 토픽 수에 따른 네트워크 간 차이를 비교하였다. 시각화에는 NodeXL을 사용하였다(Hansen, Shneiderman, & Smith, 2010).

그리고, 토픽 모델링 결과 중 각 토픽과 이에 적재된 주요 키워드를 이용하여 분석 결과의 품질에 대한 내적 평가를 실시하였다. 연구 동향 분석 결과가 개별 이용자가 특정 논문만으로 파악하기 어려운 전체적인 연구 동향을 제시한다고 하였을 때, 이용자 피드백에 의한 성능 평가가 아닌 내적 성능 평가가 적절한 것으로 보인다. 또한 기존의 토픽 모델링 관련 연구들에서도 토픽 수의 변화에 따른 토픽 모델링 결과의 품질 평가를 시도하였으므로(Ding & Chen, 2014), 같은 맥락에서 이 연구에서도 성능 평가를 수행하였다.

이 연구의 성능평가에는 클러스터링의 내적 품질 평가와 동일하게 '토픽 간 유사도가 낮고, 동일한 토픽 내 주요 키워드들 간의 유사도가 높으면 토픽 모델링 성능이 좋다'라는 가정에 기반하여 여러 가지 클러스터링의 내적 품질 평가 척도 중 해석이 용이하고 비교적 많이 사용되는 Davies-Bouldin Index(이하 DBI, Davies, & Bouldin, 1979)와 Silhouette index(이하 Index S, Rousseeuw, 1987)를 사용하였다.

DBI는 위의 가정을 따라서 개발된 지수이다. 특정 용어 k가 속한 클러스터인 Ci를 가정할 때, 클러스터 내 다른 용어 w들과의 거리는 가깝고, 다른 클러스터와의 거리는 멀 때 클러스터링이 잘 이루어졌다고 평가한다. 모든 클러스터 Ci에 대하여 그 값을 구한 후 평균을 내어 전체적인 클러스터링 성능을 평가하게 된다. DBI는 거리에 근거하기 때문에 그 값이 작을수록 클러스터링 성능이 더 좋음을 나타낸다. 계산 과정 및 공식은 다음과 같다.

$$R_{i,j} = \frac{S_i + S_j}{M_{i,j}}$$

먼저 클러스터 간의 관계성(거리값)을 구한다. Ri,j는 두 클러스터 Ci와 Cj 간의 거리이며, Si는 Ci의 클러스터내 거리, Sj는 Cj의 클러스터내 거리이다. 그리고 Mi,j는 Ci와 Cj 간의 거리를 말한다. 이 때, Ri,j의 값은 Rj,i와 같게 된다.

Ri,j를 구한 후, Ci 클러스터가 다른 클러스터

들과 갖는 거리값 중 최대 거리값을 Di로 선정한다. 여기서 '≡'는 '대응된다'라는 뜻이다.

$$D_i \equiv \max_{j:i\neq j} R_{i,j}$$

DBI는 Di를 모든 클러스터에 대해서 구한 후 이를 합하고 평균하여 계산하게 된다. 아래 공식에서 N은 클러스터수를 말한다.

$$\mathrm{DBI} \equiv \frac{1}{N}\sum_{i=1}^{N} D_i$$

Index S는 클러스터링의 내적 품질 척도의 기본 가정에 근거한 지수로 1에서 -1 사이의 값을 갖는다. 이 지수값이 1에 가까우면 분석된 문헌이나 용어가 적절하게 클러스터에 속하게 된 것을 말하며, -1에 가까울수록 그렇지 않다는 것으로 해석된다(Rousseeuw, 1987). 이 지수값은 각 데이터 i에 대한 실루엣 범위(silhouette width) 값인 S(i)를 계산한다. 그리고 이를 각 클러스터 별로 평균하여 특정 클러스터의 평균 실루엣 범위를 계산한다. 그 후, 모든 클러스터에 대한 평균 실루엣 범위를 합한 후 평균하여, 전체 데이터셋의 값인 Index S가 최종적으로 계산된다. 이 때, 특정 데이터 i의 실루엣 범위 공식 S(i)는 다음과 같다. a(i)는 클러스터내 평균 거리, b(i)는 데이터 i와 다른 클러스터와의 거리 중 최소 평균 거리를 말한다.

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

이 연구에서는 DBI와 Index S를 통계 프로그램인 R 3.0.2 기반으로 구현된 ClusterSim v 0.44패키지를 사용하여 산출하였다(Walesiak & Dudek, 2010).

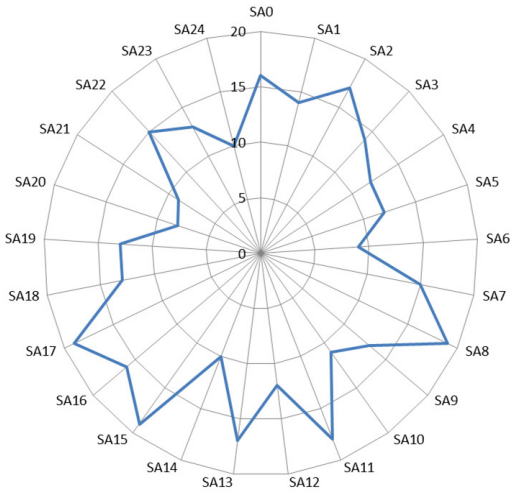
4. 분석 결과 및 내적 성능 평가

4.1 토픽 수에 따른 전체문헌 집합 내 LDA 기반 토픽 분석 결과

PA에서는 5개 토픽 내 주요 키워드 88개를 확인하였고, SA에서는 25개 토픽 내 주요 키워드 344개를 확인하였다. 그리고 두 개 이상의 토픽 클러스터에 중복 출현한 주요 키워드를 RAMAK을 적용하여 재배치한 결과는 〈그림 2〉와 같다. 재배치 전에는 모든 토픽 클러스터에 확률 분포 기준 상위 20개 키워드가 동일하게 제시되었으나, 재배치 후 각 토픽 클러스터를 통해 제공되는 주요 키워드의 수가 변화하였다. 재배치 후 변화된 주요 키워드의 수는 SA의 25개 토픽을 분석한 결과 내에서 더 확연하게 발생하였다. 이 중 SA6, SA20, SA21은 토픽 클러스터 내 LDA 결과의 50% 이상(10개 이상)의 키워드가 다른 토픽 클러스터로 재배치되었다. PA의 5개 토픽 클러스터 간에는 주요 키워드 재배치가 비교적 크게 발생하지 않았다.

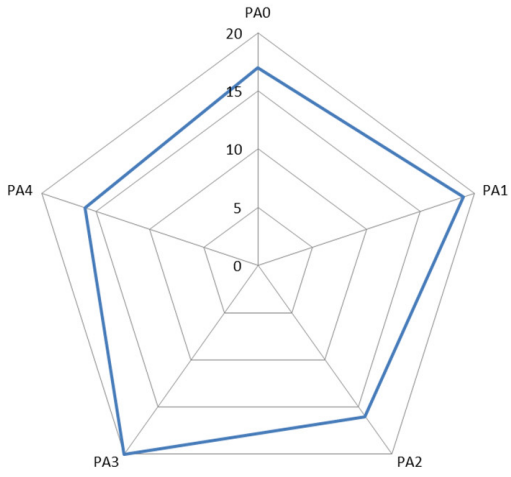
〈표 1〉은 RAMAK을 SA의 25개 토픽에 중복 적재된 주요 키워드를 재배치한 결과의 일부이다.

그리고 혼잡도 기반 5개 토픽 LDA에 RAMAK을 적용한 결과는 〈표 2〉와 같다. 적용 후 각 토픽의 상위 키워드 간의 중복 출현이 없으며, 토픽별로 상이한 주요 키워드가 적재되었다.



SA0	16	**SA5**	12	**SA10**	11	**SA15**	19	**SA20**	8
SA1	14	**SA6**	9	**SA11**	18	**SA16**	16	**SA21**	9
SA2	17	**SA7**	15	**SA12**	12	**SA17**	19	**SA22**	15
SA3	14	**SA8**	19	**SA13**	17	**SA18**	13	**SA23**	13
SA4	12	**SA9**	13	**SA14**	10	**SA19**	13	**SA24**	10

(a) SPLC 분석결과 기준 25개 토픽 내 주요키워드 재배치 결과



PA0	17
PA1	19
PA2	16
PA3	20
PA4	16

(b) 혼잡도 기준 5개 토픽 내 주요키워드 재배치 결과

〈그림 2〉 전체 문헌집단 내 주요 키워드 재배치 결과

〈표 1〉 SA의 25개 토픽 내 주요 키워드 재배치 결과 일부

	LDA 주요 키워드	LDA + RAMAK 주요 키워드
SA0	arrays, crystal, crystals enhancement, extraction, fabricated, fabrication, formation, GaN, growth, improvement, lithography, nanowires, output, patterned, photonic, power, sapphire, substrate, surface	arrays, enhancement, extraction, fabricated, fabrication, formation, improvement, lithography, nanowires, output, patterned, photonic, power, sapphire, substrate, surface
SA1	calculations, ceramics, chemical, chemistry, compounds, conversion, crystal, crystalstructure, diffraction, electronic, inorganic, nitridosilicates, optical, powder, silicon-nitride, solid-state, structural, structure, structures, x-ray	chemical, chemistry, crystal, crystalstructure, diffraction, electronic, inorganic, nitridosilicates, powder, silicon-nitride, structural, structure, structures, x-ray
SA2	bilirubin, color, conversion, efficacy, green, hyperbilirubinemia, index, jaundice, lamps, luminous, neonatal, newborn, optimization, phototherapy, purity, rendering, temperature, therapy, tuning, uniformity	bilirubin, color, efficacy, hyperbilirubinemia, index, jaundice, lamps, luminous, neonatal, newborn, optimization, phototherapy, purity, rendering, therapy, tuning, uniformity

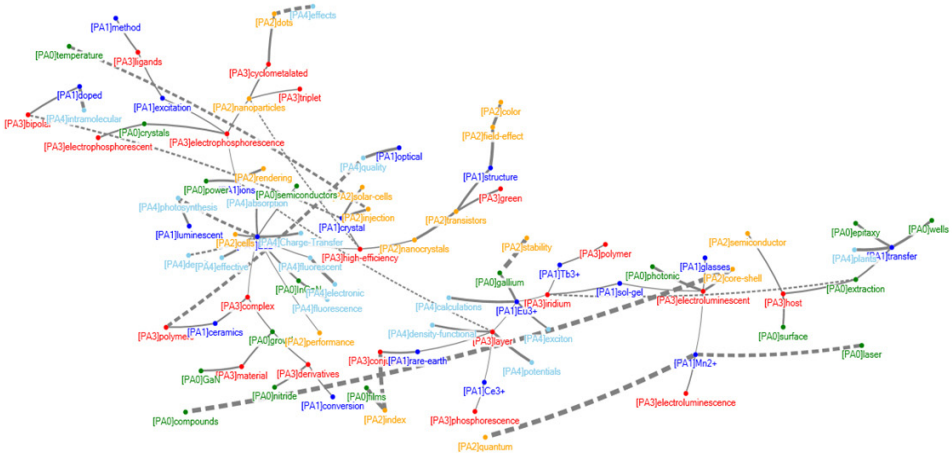
〈표 2〉 PA의 5개 토픽 내 주요 키워드 재배치 결과

	LDA 주요 키워드	LDA + RAMAK 주요 키워드
PA0	compounds, crystals, epitaxy, extraction, films, gallium, GaN, growth, InGaN, laser, layer, nitride, optical, photonic, power, quantum, semiconductors, surface, temperature, wells	compounds, crystals, epitaxy, extraction, films, gallium, GaN, growth, InGaN, laser, nitride, photonic, power, semiconductors, surface, temperature, wells
PA1	Ce3+, ceramics, conversion, crystal, doped, Eu2+, Eu3+, excitation, glasses, green, ions, luminescent, method, Mn2+, optical, rare-earth, sol-gel, structure, Tb3+, transfer	Ce3+, ceramics, conversion, crystal, doped, Eu2+, Eu3+, excitation, glasses, ions, luminescent, method, Mn2+, optical, rare-earth, sol-gel, structure, Tb3+, transfer
PA2	cells, color, conversion, core-shell, dots, field-effect, films, index, injection, layer, nanocrystals, nanoparticles, performance, polymer, quantum, rendering, semiconductor, solar-cells, stability, transistors	cells, color, core-shell, dots, field-effect, index, injection, nanocrystals, nanoparticles, performance, quantum, rendering, semiconductor, solar-cells, stability, transistors
PA3	bipolar, complex, conjugated, cyclometalated, derivatives, electroluminescence, electroluminescent, electrophosphorescence, electrophosphorescent, green, high-efficiency, host, iridium, layer, ligands, material, phosphorescence, polymer, polymers, triplet	bipolar, complex, conjugated, cyclometalated, derivatives, electroluminescence, electroluminescent, electrophosphorescence, electrophosphorescent, green, high-efficiency, host, iridium, layer, ligands, material, phosphorescence, polymer, polymers, triplet
PA4	absorption, calculations, charge-transfer, density, density-functional, derivatives, effective, effects, electronic, exciton, fluorescence, fluorescent, growth, intramolecular, photosynthesis, plants, potentials, quality, transfer, triplet	absorption, calculations, charge-transfer, density-functional, effective, electronic, exciton, fluorescence, fluorescent, intramolecular, photosynthesis, plants, potentials, quality

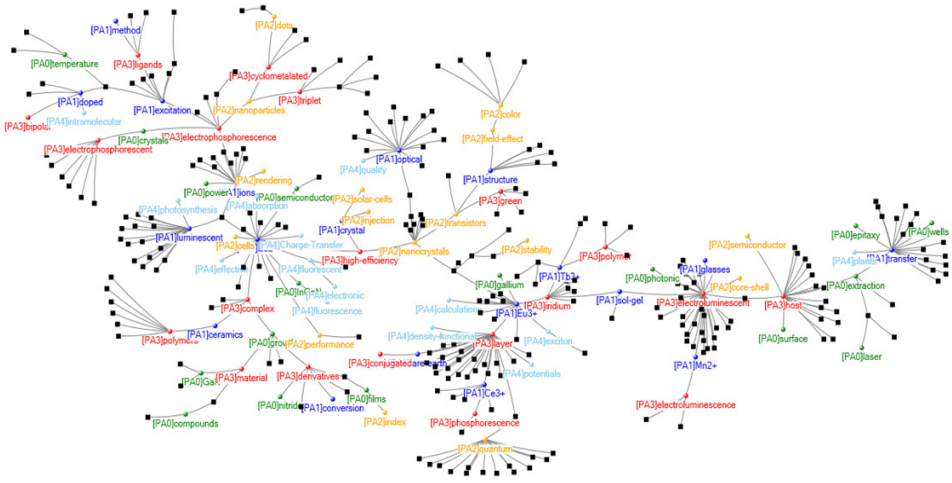
패스파인더 기법을 통해 확인된 두 네트워크 내의 주요 연결구조는 〈그림 3〉과 같다. 전자의 88개 키워드 중 86개 키워드가 SPLC 기준 25개 토픽 내 상위 키워드 동시출현 네트워크 내 출현하였으며, 포함되지 않은 키워드는 'density'와 'effect'이다. 그리고 87개의 주요 연결 중 14개 (16.1%)가 서로 다르게 파악되었는데, 이는 〈그림 3 - (a)〉에 점선으로 표시되었다. 혼잡도 기반 5개 토픽은 〈그림 3〉의 맵 상에서 서로 다른 노드 색상과 주요 키워드 레이블에 [PA 토픽번호]로 표시하였다.

14개의 서로 다른 단어 동시 출현의 주요 연결은 SPLC 분석 결과 기준 25개 토픽 내 상위 키워드의 동시 출현 네트워크 내에서는 직접 연결되지 않으나, 혼잡도 기준 5개 토픽 내 상위 키워드 동시 출현 네트워크에서는 직접 연결된 것이 주요 연결구조로 추출된 것을 말한다. 예를 들어, 2개의 주요 키워드 'quantum'과 'Mn2+'는 〈그림 3 - (b)〉의 패스파인더 네트워크에서는 직접적인 동시출현 유사도가 주요 연결의 하나로 파악되지 않고 다른 주요 키워드들과의 주요 연결이 더 많이 추출되었다. 그러나, 〈그림 3 - (a)〉에서는 이 둘 간의 동시출현 유사도가 패스파인더 네트워크를 통해 추출된 것을 알 수 있다.

두 네트워크에서 공통된 86개 키워드들간의



(a) 혼잡도 기준 5개 토픽 내 상위 키워드 동시출현 네트워크 (PA5)



(b) SPLC 분석결과 기준 25개 토픽 내 상위 키워드 동시출현 네트워크 (SA25)

〈그림 3〉 전체 문헌집단 내 상위 키워드 동시출현 네트워크

주요 연결관계를 비모수 상관 분석한 결과, 두 네트워크의 연결 구조는 통계적으로 유의미하게 유사한 것으로 나타났다(r= 0.412, p 〈 .01, N = 87). 하지만 상관계수가 0.6 미만으로 낮게 나타났으므로, 이 두 네트워크 간의 연결 구조는 유사하나 일부 차이가 있음을 알 수 있다.

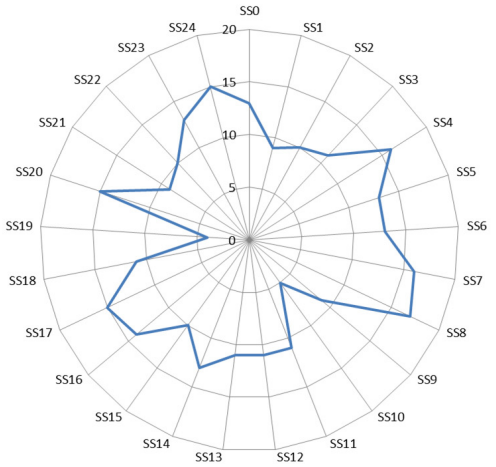
4.2 토픽 수에 따른 SPLCNN 문헌집합 내 LDA 기반 토픽 분석 결과

PS에서는 2개 토픽 내 주요 키워드 40개를 확인하였고, SS에서는 25개 토픽 내 주요 키워드 294개를 확인하였다. 그리고 두 개 이상의

토픽 클러스터에 중복 출현한 주요 키워드를 RAMAK을 적용하여 재배치한 결과는 〈그림 4〉와 같다. 재배치 후 변화된 주요 키워드의 수는 SS의 25개 토픽을 분석한 결과 내에서 비교적 더 확연하게 발생하였으며, 이는 SA의 25개 토픽 내의 변화보다도 더 활발한 것으로 나타났다. 예를 들어, SS10과 SS19에서는, 토픽 클러스터 내 LDA 결과의 75%이상 (15개 이상)의 키워드가 다른 토픽 클러스터로 재배치되었다. 반면에 PS의 2개 토픽 클러스터 간에는 주요 키워드 간 중복 출현이 없었으며, RAMAK을 적용한 재배치 후에도 동일한 키워드가 적재되었다.

〈표 3〉은 RAMAK을 SS의 25개 토픽에 중복 적재된 주요 키워드를 RAMAK을 이용하여 재배치한 결과의 일부이다.

〈표 4〉는 〈그림 5 - (a)〉 맵에 시각화된 주



SS0	13	**SS5**	13	**SS10**	5	**SS15**	10	**SS20**	15
SS1	9	**SS6**	13	**SS11**	11	**SS16**	14	**SS21**	9
SS2	10	**SS7**	16	**SS12**	11	**SS17**	15	**SS22**	10
SS3	11	**SS8**	17	**SS13**	11	**SS18**	11	**SS23**	13
SS4	16	**SS9**	9	**SS14**	13	**SS19**	4	**SS24**	15

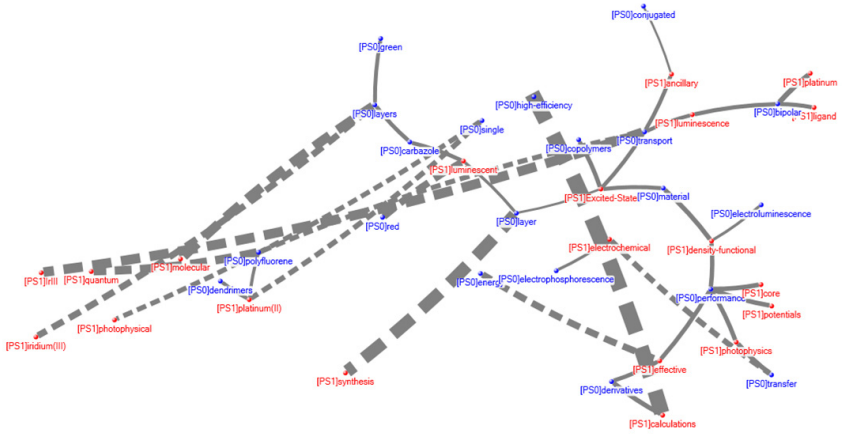
〈그림 4〉 SPLCNN 문헌집단 내 주요 키워드 재배치 결과

〈표 3〉 SS의 25개 토픽 내 주요 키워드 재배치 결과 일부

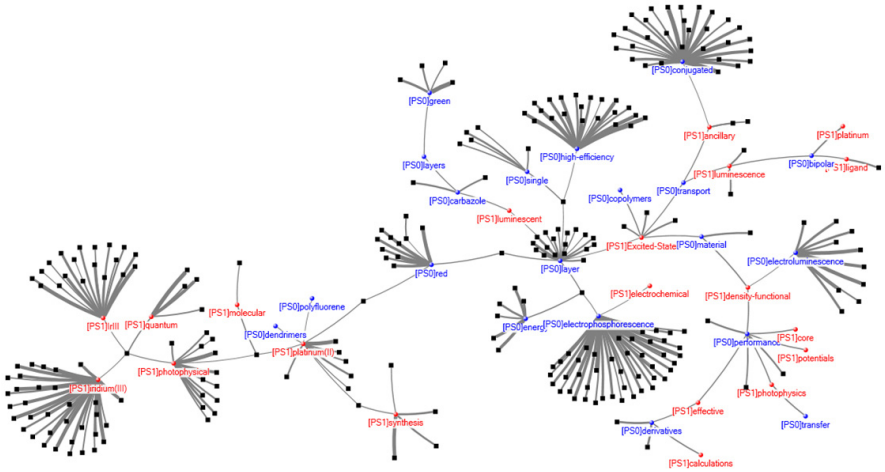
	LDA 주요 키워드	LDA + RAMAK 주요 키워드
SS0	blends, bright, color, confinement, conjugated, dopant, dopants, electroluminescence, electrophosphorescence, energy, exciton, green, layer, Poly(N-vinylcarbazole), polyfluorene, quenching, red, single, stable, transfer	blends, bright, confinement, dopants, energy, exciton, layer, performance, Poly(N-vinylcarbazole), quenching, single, stable, transfer
SS1	applications, carbene, chromophores, cyclometallated, design, electroluminescence, electrophosphors, excimer, luminescence, molecular, optoelectronic, photophysical, photophysics, platinum, platinum(II), PtII, quantum, red, saturated, synthesis	chromophores, cyclometallated, design, electrophosphors, excimer, luminescence, platinum, platinum(II), PtII
SS2	2-phenylpyridine, ancillary, color, crystal, derivative, electroluminescence, electrophosphorescence, green, heteroleptic, interligand, iridium(III), ligand, photoluminescence, red, skyblue, structure, substituted, synthesis, tetraphenylimidodiphosphinate, tuning	2-phenylpyridine, crystal, derivative, heteroleptic, interligand, photoluminescence, skyblue, structure, substituted, tetraphenylimidodiphosphinate

〈표 4〉 PS의 2개 토픽 내 주요 키워드 재배치 결과

	LDA 주요 키워드	LDA + RAMAK 주요 키워드
PS0	electrophosphorescence, high-efficiency, conjugated, electroluminescence, energy, layer, single, layers, carbazole, green, bipolar, performance, material, red, derivatives, transport, copolymers, dendrimers, transfer, polyfluorene	LDA 주요 키워드와 동일
PS1	IrIII, platinum(II), molecular, photophysical, ligand, iridium(III), luminescent, ancillary, density-functional, excited-state, core, photophysics, effective, potentials, synthesis, luminescence, quantum, electrochemical, platinum, calculations	LDA 주요 키워드와 동일



(a) 혼잡도 기준 2개 토픽 내 상위 키워드 동시출현 네트워크 (PS2)



(b) SPLC 분석결과 기준 25개 토픽 내 상위 키워드 동시출현 네트워크 (SA25)

〈그림 5〉 SPLCNN문헌집단 내 상위 키워드 동시출현 네트워크

요 키워드 및 이의 토픽 클러스터를 나타낸 것이다. PS의 2개 토픽 클러스터 내 주요 키워드들 간에는 두 개 이상의 클러스터에 중복 출현한 키워드가 없었으며, RAMAK을 적용한 후에도 주요 키워드는 재배치되지 않았다.

〈그림 5〉는 토픽 수에 따른 LDA 결과를 시각적으로 비교하기 위하여 (a)에는 SP 2개 토픽 내의 40개 주요 키워드들 간의 주요 연결을 표시하였고, (b)에는 SS 25개 토픽 내의 294개 주요 키워드들 간의 주요 연결을 표시한 것이다. 이 때 (a) 맵 내에 출현한 40개 키워드 중 (b) 맵에 동시에 출현한 40개 키워드를 동일한 위치에 표시함으로써 주요 연결 관계를 비교할 수 있도록 하였다. 비교 결과, (a) 맵 내의 39개의 주요 연결 중 12개 (30.8%)가 (b) 맵 내의 연결과 서로 다른 것으로 나타났으며, 이는 〈그림 5 - (a)〉에 점선으로 표시되었다.

두 네트워크에서 공통된 40개 키워드들 간의 주요 연결 관계간 유사도를 비모수 상관분석한 결과 두 개의 단어 동시출현 네트워크 간에는 통계적으로 유의미하게 유사한 연결 구조를 보이지 않는 것으로 나타났다(r= -0.299, p 〉 .05, N = 39).

4.3 분석 결과의 내적 품질 평가

〈표 5〉는 RAMAK을 적용한 LDA 기반 토픽 모델링 결과를 대상으로 내적 품질 평가를 실시한 결과를 비교한 것이다.

토픽 클러스터의 내적 품질 평가 결과, LDA를 적용한 문헌 집단의 크기에 상관없이 혼잡도에 따른 토픽수(PA5 와 PS2)로 분석한 것이 SPLC 결과 기반 25개의 토픽수로 분석한 것보다 토픽 간의 중복도가 낮은 것으로 나타났다. 이는 Mccallum 등(2009)이 특허 및 NYT, 20NG 데이터셋을 사용해서 토픽 수의 25개부터 100개까지 증가시키면서 정보변이계수(Variation of Information)의 변화량을 측정한 결과와 동일하다. 즉, PA5의 DBI 값(3.652)이 SA25의 DBI 값(3.755)보다 작았으며, PA5의 Index S의 값(-0.027)이 1에 더 가까워 DBI와 일관된 성능 평가 결과를 보여 주었다. 그리고 PS2가 SS25보다 DBI는 낮고 Index S는 높게 나타나 PS2의 토픽 클러스터 내적 품질이 높은 것을 알 수 있다.

각 분석 문헌 집단 간 성능을 비교해 보면, 동적 인용 네트워크 분석 방법인 SPLC 결과와 키워드의 분포 특성인 혼잡도를 모두 활용한 분석 결과(PS2)가 가장 좋은 성능을 보이는 것으로 나타났다. 즉, PS2의 DBI가 가장 낮고, Index S가 1에 가장 가깝게 나타났다(DBI = 3.270, Index S = 0.113). 반면에 SPLCNN 문헌을 사용하여 25개 토픽을 분석한 결과(SS25)

〈표 5〉 토픽 모델링 결과의 내적 품질 평가

	ALL		SPLCNN	
	SA25	PA5	SS25	PS2
주요 키워드 수	344	88	294	40
DBI	3.755	3.652	3.744	3.270
Index S	-0.076	-0.027	-0.120	0.113

의 Index S가 가장 낮은 값을 보였다(Index S = -0.076). 이는 분석 데이터셋의 주제가 'White LED'로 특정적임에 비해 설정된 토픽의 수가 상대적으로 많았고 이로 인해 각 토픽 내 출현한 주요 키워드 간의 중복도가 높아져 토픽 간 거리가 가까워졌기 때문으로 해석된다.

5. 결 론

확률모델을 사용하는 토픽모델링에서는 하나의 키워드가 모델링을 통해 파악된 토픽 클러스터에 얼마나 확률적으로 적재되는지를 알려주며, 이를 이용하여 각 토픽별로 일정 수의 주요 키워드를 제시하는 방법을 통해 토픽 모델링 결과를 제시하여 왔다. 이와 같은 결과 제시 방법은 특히 다른 데이터에 비하여 상대적으로 주제적 응집성이 높은 특정 연구 주제 분야 내 논문 데이터를 대상으로 한 토픽 분석 결과 내 중복 제시되는 키워드가 발생할 가능성이 더 높아진다고 볼 수 있다. 따라서 이 연구에서는 인용 네트워크 분석 및 자아 중심 네트워크 기법을 적용하여 토픽 모델링 결과 제시 방법을 정련하고자 하였다.

연구 결과, SPLC 네트워크 분석을 통해 생성한 문헌 데이터셋에 혼잡도에 따른 토픽수를 사용하여 LDA 분석을 수행한 후 중복 분류된 토픽 내 주요 키워드를 자아중심 네트워크 분석 기법을 적용하여 재배치한 경우가 제시되는 토픽별 주제적 응집성이 높고 각 토픽 간 변별력이 있는 것으로 나타났다. 이를 통해 볼 때, 데이터셋의 주제적 응집성이 높은 경우, 문헌 집단 내 출현 단어의 특성인 혼잡도를 고려하는 것이 문헌 집단 내 동적 인용 네트워크를 고려하는 것보다 더 적절함을 확인할 수 있었다.

그리고 주요 키워드의 동시 출현 네트워크 간 비교를 통해 설정된 토픽 수에 따라 파악되는 주요 키워드의 규모가 달라지면 패스파인더 네트워크를 통해 추출된 이들 간의 주요 연결 구조에도 통계적으로 유의미한 차이가 있음을 확인하였다. 따라서 동적 인용 네트워크 및 자아 중심 네트워크 분석을 적용하여 연구 동향을 분석함으로써 토픽 모델링에 의한 분석 결과를 보완하는 다면적인 연구 동향 분석이 가능할 것으로 보인다. 동시에 적절한 수의 토픽을 설정하는 방법을 정교화하고 내적 품질 평가와 함께 이용자 평가에 근거한 외적 품질 평가를 이용하여 토픽 모델링 결과를 제시하는 방법을 더 보완할 수 있을 것이다. 또한 다양한 논문 및 특허 데이터를 대상으로 한 분석 및 성능 평가를 통해 이 연구에서 제시한 접근 방법을 재검토할 필요가 있을 것으로 보인다.

참 고 문 헌

박자현, 송민 (2013). 토픽모델링을 활용한 국내 문헌정보학 연구동향 분석. 정보관리학회지, 30(1), 7-32. http://dx.doi.org/10.3743/KOSIM.2013.30.1.007

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

유소영 (2013). 문헌 단위 인용 네트워크 구조와 Topic Descriptor Profile을 활용한 연구경향 분석에 관한 연구. 2013 한국정보관리학회 추계 학술대회 논문집, 39-58.

이재윤, 김판준, 강대신, 김희정, 유소영, 이우형 (2011). 계량서지적 기법을 활용한 LED 핵심 주제영역의 연구 동향 분석. 정보관리연구, 42(3), 1-26. http://dx.doi.org/10.1633/JIM.2011.42.3.001

정우성, 양현재 (2013). 과학계량학 연구동향 및 과학기술 정책 분야 응용가능성. KISTEP ISSUE PAPER2013-06. 한국과학기술기획평가원.

한국과학기술정보연구원 (2013). 미래기술백서 2013. 한국과학기술정보연구원. Retrieved from http://mirian.kisti.re.kr/utility/tech_book/tech_book.jsp

한국과학기술정보연구원 (2014). 미래기술백서 2014. 한국과학기술정보연구원. Retrieved from http://mirian.kisti.re.kr/utility/tech_book/tech_book.jsp

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2, 224-227.

de Nooy, W., Mrvar, A., & Batagelj, V. (2011). Exploratory social network analysis with Pajek (Revised and expanded second edition). Cambridge: Cambridge University Press.

Ding, W., & Chen, C. (2014). Dynamic topic detection and tracking: A comparison of HDP, C-word, and cocitation methods. Journal of the Association for Information Science and Technology, 65(10), 2084-2097.

Garfield, E. (2001, September 19). From computational linguistics to algorithmic historiography. Lazerow lecture held in conjunction with panel on "Knowledge and language: Building large-scale knowledge bases for intelligent applications," presented at the University of Pittsburgh. Retrieved from http://garfield.library.upenn.edu/papers/pittsburgh92001.pdf

Garfield, E. (2001, November 27). From bibliographic coupling to co-citation analysis via algorithmic historio-bibliography: A citationist's tribute to Belver C. Griffith. Lazerow Lecture presented at Drexel University, Philadelphia, PA. Retrieved from http://garfield.library.upenn.edu/papers/drexelbevergrif?th92001.pdf

Garfield, E., Pudovkin A. I., & Istomin, V. S. (2002). Algorithmic citation-linked historiography-Mapping

the literature of science. Proceedings of the American Society for Information Science and Technology Annual Meeting, 39, 14-24.

Garfield, E., Pudovkin, A. I., & Istomin, V. S. (2003). Why do we need algorithmic historiography? Journal of the American Society for Information Science and Technology, 54(5), 400-412. http://dx.doi.org/10.1002/asi.10226

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. Proceedings of the National Academy of Sciences, 101(suppl 1), 5228-5235.

Hansen, D., Shneiderman, B., & Smith, M. A. (2010). Analyzing social media networks with NodeXL: Insights from a connected world. Morgan Kaufmann. Huang, W., Wu, Z., Liang, C., Mitra, P., & Giles, C. L. (2015). A Neural Probabilistic Model for Context Based Citation Recommendation.

Hummon, N. P., & Doreian, P. (1989). Connectivity in a citation network: The development of DNA theory. Social Networks, 11, 39-63.

Jiang, Z. (2015). Chronological scientific information recommendation via supervised dynamic topic modeling. In Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (pp. 453-458). ACM.

Mccallum, A., Mimno, D. M., & Wallach, H. M. (2009). Rethinking LDA: Why priors matter. In Advances in Neural Information Processing Systems (pp. 1973-1981).

Ramage, D., Hall, D., Nallapati, R., & Manning, C. D. (2009a, August). Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1 (pp. 248-256). Association for Computational Linguistics.

Ramage, D., Rosen, E., Chuang, J., Manning, C. D., & McFarland, D. A. (2009b, December). Topic modeling for the social sciences. In NIPS 2009 Workshop on Applications for Topic Models: Text and Beyond (Vol. 5).

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53-65.

Saka, A., & Igami, M. (2014). Science Map 2010&2012. Policy (NISTEP REPORT No. 159).

Song, Z. (2010). Research on text categorization based on LDA. Matster Degree Dissertation. Xi'an University of Techonology, Xi'an, China.

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical dirichlet processes. Journal of the American Statistical Association, 101(476), 1566-1581.

Walesiak M., & Dudek A. (2010). The cluster sim package for R. University of Wrac̷low,

Wrac̷low Retrieved from http://keii.ue.wroc.pl/clusterSim

Yu, S. Y. (2014). Exploratory study of developing a synchronization-based approach for multi-step discovery of knowledge structures. Journal of Information Science Theory and Practice, 2 (2) Korea Institute of Science and Technology Information. doi:10.1633/JISTaP.2014.2.2.2

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

Chung, Woo-Sung, & Yang, Hyun-Chae (2013). ISSUE PAPER 2013-06. KISTEP.

KISTI (2013). White Paper on Future Technologies 2013. KISTI.

KISTI (2014). White Paper on Future Technologies 2014. KISTI.

Lee, Jae-Yun, Kim, Pan-Jun, Kang, Dae-Shin, Kim, Hee-Jung, Yu, So-Young, & Lee, Woo-Hyoung (2011). A biliometric analysis on LED research. Journal of Information Management, 42(3), 1-26.

Park, Ja-Hyun, & Song, Min (2013). A study on the research trends in Library & Information Science in Korea using topic modeling. Journal of the Korean Society for Information Management, 30(1), 7-32. http://dx.doi.org/10.3743/KOSIM.2013.30.1.007

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.org/10.3743/KOSIM.2013.30.4.215

Yu, So-Young (2013). Applying TDP (Topic Descriptor Profile) with article-level citation flow for analyzing research trend, In Proceedings of the 2013 Korean Society for Information Management Conference in Autumn (pp. 39-58). Seoul: Korean Society for Information Management.