Journal of The Korea Society of Computer and Information
Vol. 21 No. 5, pp. 141-148, May 2016
www.ksci.re.kr
http://dx.doi.org/10.9708/jksci.2016.21.5.141

Trend Analysis of Data Mining Research Using Topic Network Analysis

Hyon Hee Kim*, Hey Young Rhee**

Abstract

In this paper, we propose a topic network analysis approach which integrates topic modeling and social network analysis. We collected 2,039 scientific papers from five top journals in the field of data mining published from 1996 to 2015, and analyzed them with the proposed approach. To identify topic trends, time-series analysis of topic network is performed based on 4 intervals. Our experimental results show centralization of the topic network has the highest score from 1996 to 2000, and decreases for next 5 years and increases again. For last 5 years, centralization of the degree centrality increases, while centralization of the betweenness centrality and closeness centrality decreases again. Also, clustering is identified as the most interrelated topic among other topics. Topics with the highest degree centrality evolves clustering, web applications, clustering and dimensionality reduction according to time. Our approach extracts the interrelationships of topics, which cannot be detected with conventional topic modeling approaches, and provides topical trends of data mining research fields.

▸Keyword: Topic network analysis, Research trend analysis, Topic modeling, Social network analysis

I. Introduction

최근 빅데이터에 대한 관심이 폭발적으로 증가함에 따라 빅데이터 분석의 핵심 기술인 데이터 마이닝 기법이 다시금 주목받고 있다. 통계학, 기계 학습, 데이터베이스 등 다양한 학문 분야의 학제적 연구 영역인 데이터 마이닝은 1990년대 중반 처음 “data mining”이라는 단어가 등장한 이후 꾸준히 연구되고 있는 분야이다. 특히 2010년 이후 기존의 데이터 마이닝 기법들이 대용량 데이터 처리 환경에 적용되거나 새롭게 등장한 데이터 형태에 따라 그 알고리즘이 변용되는 등 보다 다양한 응용 영역으로 확대되고 있다. 따라서 과거 20년간의 데이터 마이닝 연구 주제를 파악한다면 현재 데이터 마이닝을 연구하거나 실질적인 응용 분야에 적용하는 데 도움이 될 것이다.

텍스트 마이닝 기법은 문서를 자동으로 분류하는데 주로 사용되어 왔다 [1, 2]. 특히, 연구 동향 분석을 위해 연구 논문에 토픽 모델링[3]과 그 변형 알고리즘을 적용하여 관심이 많은 주제(hot topic)와 관심이 적은 주제(cold topic)를 찾아내는 연구가 주로 이루어졌다. [4]에서는 국내 문헌정보학 분야의 연구 동향을 분석하기 위해 문헌 정보학 관련 주요 학술지의 논문 초록에 Latent Dirichlet Allocation (LDA) 기반 토픽 모델링을 적용하여 주요 연구 주제 및 상승세와 하강세에 있는 연구 주제를 보여주었다. Blei[5]는 “Science” 저널에 실린 17,000편의 논문과 “Yale Law” 저널을 대상으로 LDA 모델을 적용하여 각 저널의 토픽을 제시하였으며, Mimno와 McCallum은 텍스트 문서뿐 만이 아니라 저자, 출판날짜, 참고문헌 등의 메타 데이터도 동시에 처리할 수 있는 Dirichlet-multinomial regression (DMR) 모델을 제안하고 컴퓨터 과학 분야의 논문에 적용하였다.

이처럼 토픽 모델링 기법은 연구 논문을 분석하여 연구 주제 및 시간에 따른 동향을 파악하는데 효율적이지만 찾아낸 주제 간의 관계를 고려한 연구는 거의 이루어지지 않고 있다. 특히 데이터 마이닝과 같은 학제 간의 연계가 활발히 이루어지는 연구 분야에서는 각각의 토픽이 독립적이기보다는 한 토픽 내에 다른 토픽이

• First Author: Hyon Hee Kim, Corresponding Author: Hyon Hee Kim
*Hyon Hee Kim(heekim@dongduk.ac.kr), Dept. of Statistics & Information Science, Dongduk Women's University
**Hey Young Rhee (jonju@dongduk.ac.kr), Dept. of Library & Information Science, Dongduk Women's University
• Received: 2016. 04. 11, Revised: 2016. 04. 21, Accepted: 2016. 05. 09.
• This work was supported by 2013 Dongduk Women's Univ. Research Grant.

공존할 수 있으며 토픽 간의 연결 관계가 중요한 정보가 될 수 있다. 따라서 연구 논문으로부터 토픽 간의 연결 관계 및 다른 토픽에 큰 영향을 미친 토픽 그리고 한 토픽에 공존하는 부분 토픽 등 찾아낸 토픽에 대해 다시 분석 하는 기법이 필수적으로 요구된다.

관심이 있는 엔티티 간의 관계를 분석하기 위해 많이 사용되는 기법은 사회 연결망 분석 (Social Network Analysis) [6]이다. 텍스트 문서의 경우, 문서에 등장한 키워드간의 관계를 찾기 위해 키워드 네트워크 분석이 사용되고 있다. [7]에서 연구 논문에 등장한 단어를 노드로 하고 한 문서에 동시에 등장한 단어들을 연결 관계로 지정하고, 다른 문서에 단어가 등장한 횟수를 가중치로 하여 키워드 네트워크를 구축 및 분석하여 시간에 따른 키워드 패턴을 발견하였다. 키워드 네트워크 분석은 각 노드가 한 개의 단어이기 때문에 토픽 모델링에서와 같이 여러 단어의 집합으로 구성된 토픽을 알기 어려워 거시적인 연구 동향을 파악하는데 제약점이 있다.

본 연구에서는 먼저 토픽 간의 관계를 분석하기 위해서 네트워크 분석 기법을 토픽 모델링의 결과로 생성된 토픽들에 적용한 토픽 네트워크 분석 방법을 제안하였다. 토픽 네트워크를 구축하기 위해서 토픽을 한 개의 노드로 보고 노드와 노드를 연결하는 간선은 서로 다른 두 토픽 간에 공통으로 등장한 단어의 수를 가중치로 사용하였다. 이처럼 구축된 토픽 네트워크를 분석하여 연결 중심성, 매개 중심성, 그리고 근접 중심성이 가장 높은 토픽들을 찾아내었다.

다음으로 시간에 따른 토픽의 경향을 파악하기 위해서 20년간 출판된 논문을 5년씩 분리하여 4개 구간으로 나눈 다음 각각 토픽 네트워크 분석을 하여 네트워크 구조의 시간에 따른 변화를 파악하였으며 특히 영향력이 큰 토픽들의 경향을 파악하였다.

연구 주제를 분석하기 위해서 데이터 마이닝 분야의 저명 저널 5개를 선정하고, Web of Science[8]에서 주제어를 “data mining"으로 지정하여 1996년부터 2015년까지 20년간 5개의 저널에 출판된 연구 논문 초록 2,039개를 수집하여 제안한 토픽 네트워크 분석 기법을 적용하였다. 특히, 시간에 따른 토픽의 변화를 보기 위하여 1996년에서 2000년 사이에 출판된 논문 86개, 2001년에서 2005년에 출판된 논문 314개, 2006년에서 2010년에 출판된 논문 736개, 그리고 2011년도에서 2015년도에 출판된 논문 991개로 분리하여 분석하였다.

토픽 네트워크의 전체적인 구조를 측정하기 위해서 연결 중앙성, 매개 중앙성, 그리고 근접 중앙성을 사용하였으며 분석 결과 데이터 마이닝 연구 초기인 2000년도까지 모든 중앙성 값이 최고값을 갖다가 이후 5년간 급하게 감소하는 형태를 보인다. 연결 중앙성의 경우는 이후 10년간 꾸준히 다시 증가하는 패턴을 보이며, 매개 중앙성과 근접 중앙성은 2010년도까지 증가하다가 최근 5년에 다시 감소하는 추세를 보인다. 이는 최근 5년간 빅데이터 마이닝에 관한 연구가 활발히 이루어지고 있으며 그 주제가 매우 다양하고 광범위하여 독립된 토픽을 연결해주는 토픽이나 모든 토픽과 가장 가까운 토픽이 등장하기 어려움을 반영한 것이라고 할 수 있다.

영향력이 큰 토픽들의 동향을 파악하기 위해서 연결 중심성, 매개 중심성, 그리고 근접 중심성이 가장 높은 토픽을 구간별로 분석하였다. 연결 중심성의 경우는 클러스터링, 웹 응용 프로그램, 클러스터링, 그리고 차원 축소의 주제가 구간에 따라 나타났다. 매개 중심성이 높은 토픽의 주제는 귀납 추론, 고객 관계 관리, 클러스터링, 그리고 순차 패턴 마이닝으로 변화되어 감을 알 수 있다. 마지막으로 근접 중심성이 높은 토픽의 주제는 귀납 추론, 병렬 질의 처리, 클러스터링, 그리고 순차 패턴 마이닝으로 변화되었다.

토픽의 동향에서 주목할 만한 점은 클러스터링에 대한 주제가 여러 구간에서 영향력이 높은 토픽으로 등장하였다는 점과 2001년도에서 2005년도 사이에 출판된 연구 논문들의 영향력 있는 토픽들의 주제는 데이터 마이닝 기법이라기보다는 응용 영역 및 데이터 처리에 관련된 것으로 나타났다는 것이다. 또한, 토픽 네트워크 구조에서 볼 수 있는 바와 같이 매개 중앙성과 근접 중앙성의 패턴이 유사하므로 2005년도 이후 출판된 연구 논문에서 매개 중심성과 근접 중심성이 높은 토픽은 같은 토픽으로 나타났다.

본 논문의 공헌은 다음과 같다. 첫째, 토픽 모델링과 네트워크 분석을 결합한 토픽 네트워크 분석 기법을 제안하여 다른 토픽에 영향을 많이 준 토픽과 다른 토픽을 연결하는 토픽, 그리고 다른 모든 토픽과 가장 근접한 토픽을 찾아내었다. 둘째, 과거 20년간 출판된 논문들을 5년씩 분리하여 4개 구간으로 나눈 다음 토픽 네트워크 분석을 하여 시간에 따른 네트워크 구조 변화 및 영향력 있는 토픽의 동향을 파악하였다.

본 논문은 다음과 같이 구성되어 있다. 2장에서 관련 연구를 살펴보고, 3장에서 본 연구에서 제안하는 토픽 네트워크 분석 기법에 대해 자세히 설명하고 실험 결과를 제시한다. 4장에서는 영향력 있는 토픽으로 선정된 토픽들의 주제가 시간에 따라 어떻게 변화하는지 연구 주제 동향을 살펴보고 마지막으로 5장에서 결론 및 향후 연구를 제시한다.

II. Related Work

본 연구에서 사용한 LDA 알고리즘[3]은 대표적인 확률론적 토픽 모델의 일종으로서 문서가 여러 개의 토픽으로 구성되어 있고 각 토픽은 단어의 분포로 이루어졌다는 가정에 따라 확률을 바탕으로 토픽을 생성하고 단어를 할당한다. LDA 알고리즘에 의한 토픽 모델링은 연구 동향을 분석[9]하거나 트위터에서 쟁점이 되고 있는 이슈 트래킹 시스템[10]에 성공적으로 적용되고 있다. 그러나 LDA 알고리즘은 토픽 간의 관계를 고려하지 않기 때문에 연구 토픽 간에 서로 영향을 주고받은 관계나 트위터 이슈 간의 관계를 파악할 수 없다는 제약점이 있다.

Blei는 LDA 알고리즘의 이와 같은 제약 사항을 극복하기 위해서 토픽 간의 상관관계를 고려한 Correlated Topic

Models(CTM) [11]을 개발하였다. CTM은 한 토픽 내에 다른 토픽이 함께 등장하는 경우를 고려하여 토픽을 생성하므로 보다 현실적인 모델이라고 할 수 있다. 또한 토픽간의 상관관계를 고려하였다는 점이 본 연구에서 다룬 토픽간의 관계를 파악하는 문제와 유사하다. 본 연구에서 제안하는 토픽 네트워크 분석은 LDA 알고리즘을 적용하여 먼저 독립적인 토픽들을 찾아낸 다음, 공통된 단어들을 중심으로 토픽 네트워크를 구축하여 연결 중심성, 매개 중심성, 그리고 근접 중심성이 높은 토픽들을 찾아냄으로써 다른 토픽들에 구체적으로 영향력을 끼친 연구 주제를 발견한 데에 그 독창성이 있다.

Mei[12]와 동료들은 연구 논문의 저자 관계를 기반으로 먼저 논문의 소셜 네트워크를 구축한 다음 관심 있는 노드에 토픽 모델링을 적용하여 네트워크 구조 상의 토픽들을 찾아내었다. 소셜 네트워크 분석과 토픽 모델링 기법을 통합하였다는 측면에서 본 연구와 가장 유사한 연구라고 할 수 있다. 그러나 Mei등의 연구는 먼저 생성된 소셜 네트워크를 기반으로 토픽을 추출하는 것에 그 핵심이 있다면, 본 연구에서는 형성된 토픽들의 소셜 네트워크 구조를 분석하는 것에 그 핵심이 있다. 토픽 모델링의 결과로 생성된 토픽들의 공유 키워드를 가중치로 하여 토픽 네트워크를 형성 및 분석한 연구는 거의 이루어지지 않고 있다.

토픽 모델링에 의해 생성된 토픽은 여러 개의 서브 토픽으로 구성될 수 있다. 특히 Mao[13]와 그의 동료들은 토픽의 계층 구조를 나타낼 수 있도록 hierarchical LDA와 latent topic들을 통합할 수 있는 프레임워크를 제안하였다. 또한 Wang[14]과 그의 동료들은 시간에 따른 토픽을 찾아낼 수 있는 non-Markov 모델을 제안하였다. 이러한 연구들과 본 연구의 가장 큰 차이점은 토픽 모델링을 통해 찾아낸 토픽보다는 이들 토픽이 형성할 수 있는 네트워크 구조에 관심을 두고 토픽 네트워크 분석을 한 결과로 영향력이 큰 토픽을 선정하여 시간에 따른 토픽의 경향을 분석한 것이라고 할 수 있다.

III. Topic Network Analysis

본 장에서는 먼저 제1절에서 토픽 네트워크 분석 프로세스를 살펴보고, 제2절에서 토픽 네트워크 구축에 대해 설명한다. 마지막으로 제3절에서 형성된 토픽 네트워크에 대해 분석을 실시한 결과를 제시한다.

1. Topic Network Analysis Process

그림 1은 토픽 네트워크 분석을 위한 프로세스를 나타내고 있다. 먼저 데이터 마이닝을 주제로 출판된 연구 논문들을 분석하기 위해서 Web of Science [8]에 등록된 저널 중에서 데이터 마이닝 분야 저널로 검색한 후, 데이터 마이닝 분야의 논문이 많이 출판되고 영향력 계수 (impact factor)가 상위인 저널 5개를 선정하여 논문 초록을 수집하였다. 선정된 저널은 IEEE Transactions on Knowledge and Data Engineering, Data Mining and Knowledge Discovery, Information Sciences, VLDB Journals, 그리고 Expert Systems with Applications이다. 1996년부터 2015년까지 주제어에 “data mining”이 포함된 연구 논문은 총 2,039개였으며, 연구 논문에 대한 초록만을 분석에 적용하였다.

다음으로, 수집된 논문 초록에 대해 LDA 알고리즘을 적용하기 위하여 통계적 분석 소프트웨어 환경인 R[15]과 topicmodels 패키지[16]를 사용하였다. 가장 적절한 토픽의 개수를 선정하기 위해서 10개, 20개, 30개, 40개의 토픽으로 나누어 실험한 결과 토픽의 의미가 더욱 명확하게 분리되는 개수가 30개였으므로 30개의 토픽으로 연구 논문 초록을 토픽 모델링 하였다.

토픽 모델링의 결과를 기반으로 토픽 네트워크를 구축하고 생성된 토픽 네트워크에 대해 [6]에서 네트워크의 중요 척도로 정의한 연결정도 중심성(degree centrality), 매개 중심성(betweenness centrality), 그리고 근접 중심성(closeness centrality)을 측정하였다. 과거 20년간 데이터 마이닝 연구 분야의 토픽 네트워크의 변화 과정을 살펴보기 위해서 1996~2000, 2001~2005, 2006~2010, 2011~2015년도로 5년씩 4개 구간으로 나누어 토픽 네트워크 분석을 하였다.

마지막으로 중심성이 높은 토픽들이 어떤 주제를 포함하는지 알아보기 위해서 온라인 백과사전인 위키피디아를 활용하였다. 위키피디아 데이터 마이닝 기사의 인포박스에서 분류된 데이터 마이닝 분류 기준을 활용하여 토픽을 구성하는 키워드들을 매핑한 후 이를 기반으로 특정 토픽의 연구 주제를 정의하였다.

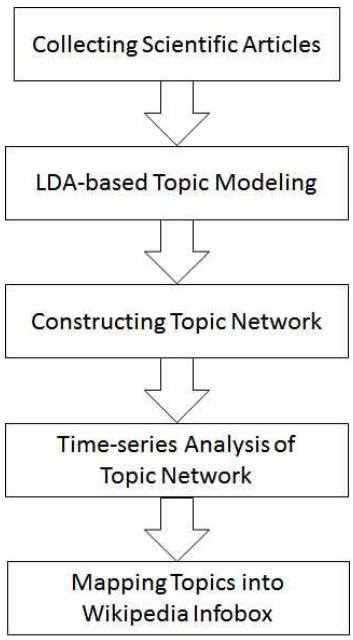


Fig. 1. Topic Network Analysis Process

2. Topic Network Construction

알고리즘 1은 토픽 네트워크 구축을 위한 알고리즘이다. 실

험에 사용한 LDA 알고리즘은 Blei[1]의 모델을 구현한 topicmodels 패키지의 LDA() 함수를 사용하였다.

논문 초록 집합인 A를 입력받아 문서-단어 행렬 DM[$f_{i,j}$]를 생성한다(1). 여기서 $f_{i,j}$는 i번째 문서에 j번째 단어가 출현한 빈도수이다. 단어의 빈도수만을 고려하면 data, user, information과 같은 데이터 마이닝 분야에서 일반적으로 사용되는 단어들이 상위에 랭크되기 마련이다. 따라서 일반적인 단어들은 제거하고 보다 중요한 단어를 선정하고자 TF-IDF 가중치[17]를 계산하여 임계치 이상의 단어들만을 선정하였다.

Algorithm 1. Algorithm for Topic Network Construction

Input: a set of abstracts A

Output: Topic Weights Matrix T[$t_{i,j}$]

1. **generate** document-term matrix DM[$f_{i,j}$]
2. **extract** important terms using TF-IDF weight
   DM[$f_{i,j}$] <- $f_{i,j}$ * log (N / $df_{i,j}$) >= k
3. **create** LDA model M with 30 topics
4. **generate** Topic Weights Matrix W[$w_{i,j}$]
   for each i
     for each j
       common <- intersect M[,i] with M[,j]
       if length(common) >= 1 **then**
         T[$t_{i,j}$] <- length(common)
       else T[$t_{i,j}$] <- 0
       endif
     endfor
   endfor

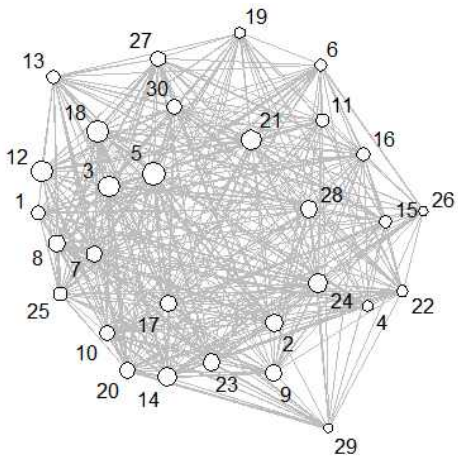


Fig. 2. Topic Network Graph (1996~2015)

본 연구에서는 단어들의 TF-IDF 평균값 및 최저값을 고려하여 임계치 k의 값을 0.8로 지정하였다(2). 생성된 문서-단어 행렬에 LDA 알고리즘을 적용하여 토픽 모델링을 실시하였다(3). 토픽 분포를 위한 alpha값은 k를 토픽의 수라고 할 때 50/k로 계산되며 초기값으로 고정하여 사용하였다. 여기서 k는 토픽의 수로 본 연구에서는 30으로 지정하였다. 또한 토픽을 위한 단어 분포를 위한 beta값은 토픽 분포를 고정함으로써 추정될 수 있도록 하였다. 표1은 LDA 알고리즘을 적용하여 토픽 모델링을 실시한 결과를 30개의 토픽에 대해 상위 5개 키워드를 제시한 것이다.

Table 1. 30 topics from Latent Dirichlet Allocation

Topic 1	Topic 2	Topic 3	Topic 4
text semantic sentiment sources source	trees software ranking numerical criteria	sequential sequence sequences distributed constraints	hybrid recommendat i-on location filtering recommender
Topic 5	Topic 6	Topic 7	Topic 8
web streams stream usage records	utility stock index constraint market	attribute rough reduction matrix approximati- on	event phase events changes change
Topic 9	Topic 10	Topic 11	Topic 12
spatial points outlier noise outliers	classifier positive instance negative programmi- ng	privacy credit factors svm risk	series objects distance local relations
Topic 13	Topic 14	Topic 15	Topic 16
relational adaptive weights relevance heterogene -ous	graph social parallel community graphs	product products association sales apriori	customer customers marketing company service
Topic 17	Topic 18	Topic 19	Topic 20
control image recognition access weight	frequent itemsets memory itemset closed	hierarchical expert traffic independe -nt induction	documents document internet frequency structured
Topic 21	Topic 22	Topic 23	Topic 24
detection temporal procedure anomaly unsupervis- ed	similarity regression functional active informative	construtions mobile stage multidimen- sional reasoning	fuzzy genetic scheme quantitative associative
Topic 25	Topic 26	Topic 27	Topic 28
clusters object ensemble categorical weighted	activities activity behaviors matching imbalanced	financial item transaction interactive window	kmeans medical random sample missing
Topic 29	Topic 30		
group structures gene expression map	query queries probability incremental sampling		

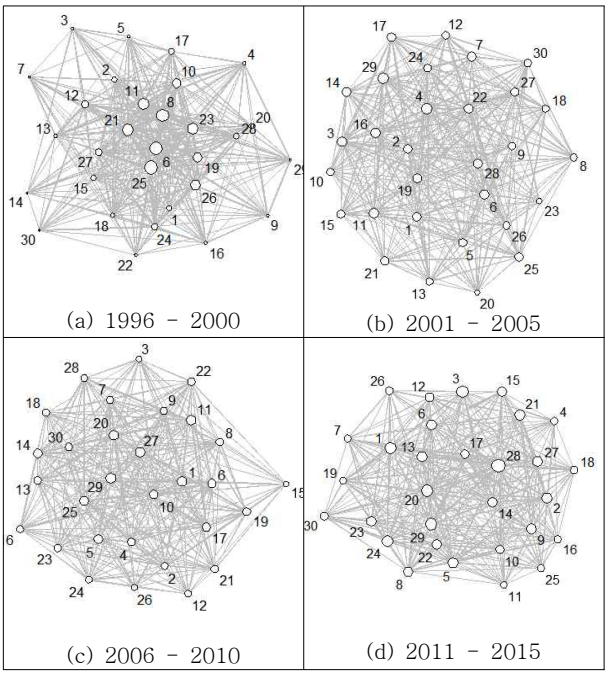
다음으로 생성된 토픽들을 기본으로 토픽 네트워크를 구축

하였다(4). 네트워크의 노드는 각 토픽이고, 노드와 노드를 연결하는 간선은 두 토픽 사이에 공통으로 포함된 단어의 개수를 가중치로 사용하였다. 토픽당 고려한 단어의 수는 중요도가 높은 상위 30개를 고려하였다. 가중치 값을 계산하기 위해서 i번째 토픽과 j번째 토픽의 교집합을 구한 다음, 교집합에 포함된 단어의 개수를 가중치로 하였다. 이와 같이 하면 30개 토픽들이 서로 같은 단어들을 공유하는 관계를 고려한 토픽 가중치 행렬 $T[t_{i,j}]$가 생성된다.

그림 2는 1996년도부터 2015년까지 20년간에 걸쳐 출판된 연구 논문의 초록 전체에 알고리즘 1을 적용하여 생성된 토픽 네트워크 그래프이다. 각 노드의 크기는 토픽의 연결 중심성의 배수로 표현하여 그래프에서 연결중심성이 높은 노드를 한눈에 알아 볼 수 있도록 하였다. 또한 그래프에서 간선의 굵기를 가중치 값으로 조절하였다.

3. Time-Series Analysis of Topic Network

생성된 토픽 네트워크를 분석하기 위해서 소셜 네트워크 분석에서 가장 기본적으로 사용되는 척도[18]인 연결성(degree), 매개성 (betweenness), 그리고 근접성 (closeness)에 대하여 중앙성(centralization)과 중심성(centrality)를 각각 측정하였다. 특히 과거 20년간 시간의 흐름에 따라 생성된 토픽 네트워크들의 변화를 알아보기 위하여 5년씩 구간을 나누어 토픽 네트워크를 재구축하였다.



3. Evolution of Topic Network Graph

연결 정도 중심성은 각 노드의 연결선 수를 측정하는 것으로 연결선의 수가 많을수록 중요한 노드로 본다. 연결선의 수가 많지 않을지라도 한 노드가 다른 두 노드를 연결하는 최단거리에 있다면 이 노드도 역시 중요한 역할을 한 것으로 볼 수 있다. 이 노드는 매개 중심성을 측정하여 알 수 있다. 마지막으로 모든 노드에 대한 인접 거리가 가장 짧은 노드 역시 중요한 노드로 볼 수 있는데 이는 인접 중심성을 측정하여 알 수 있다.

그림 3 (a)는 1996년도에서 2000년도의 연구 논문 초록을 분석한 토픽 네트워크로 전체 논문에서 86개의 논문이 이 구간에서 출판되었다. 같은 방식으로 (b)는 2001년도에서 2005년도에 출판된 314개 논문, (c)는 2006년도에서 2010년도에 출판된 736개의 논문, 그리고 (d)는 2011년도에서 2015년도에 출판된 991개 논문을 분류하여 처리하였다.

4개의 토픽 네트워크의 구조적 특성을 파악하기 위해서 각 네트워크의 중앙성을 파악하였다. 중앙성이란 전체 네트워크의 형태가 얼마나 중앙에 집중되어 있는지를 나타내는 개념으로 네트워크의 전체적인 특성을 파악하기 위하여 사용되고 있다. 중심성을 측정하는 척도에 대해 각각 그 중앙성을 구할 수 있으며 Freeman[18]에 의해 식 (1)과 같이 정의되었다.

$$C_A = \frac{\sum_{i=1}^{g}[C_A(n^*) - C_A(n_j)]}{\max \sum_{i=1}^{g}[C_A(n^*) - C_A(n_j)]} \quad (1)$$

$C_A(n_j)$를 한 노드의 중심성이라고 하고, $C_A(n^*)$를 네트워크에 속하는 모든 노드들의 중심성 중에서 최대값이라고 하자. $\sum_{i=1}^{g}[C_A(n^*) - C_A(n_j)]$는 최대 중심성값과 모든 노드의 중심성의 차들의 합이고,

$\max \sum_{i=1}^{g}[C_A(n^*) - C_A(n_j)]$은 논리적으로 가장 큰 차이를 합한 것이다. 최대 중심성값과 모든 노드의 중심성의 차들의 합을 논리적으로 가장 큰 차이들의 합으로 나눈 값이 중앙성이다.

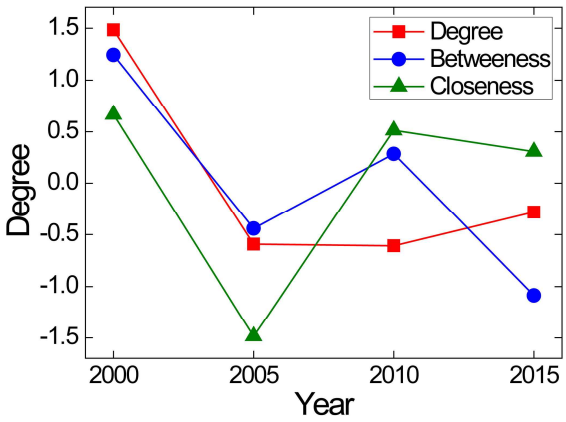


Fig. 4. Time-series analysis of centralization

그림 4는 시간에 따른 토픽 네트워크의 중앙성 분석 결과를 보여준다. 연결정도 중앙성, 매개 중앙성, 그리고 근접 중앙성의 값은 z-score 정규화 값으로 표현하였다. 먼저 연결정도 중앙성의 경우 데이터 마이닝 연구의 초기인 2000년에 최고치를 갖다가 이후 5년간 그 값이 감소하였으며 이후 10년간 다시 증가하는 추세를 보이고 있다. 매개 중앙성과 근접 중앙성은 같은 패턴을 보이는데 연결정도 중앙성과 마찬가지로 2000년도에 최고치를 갖다가 다시 감소하였으며 2010년도에 다시 증가하다가 최근 들어 감소하는 경향을 보이고 있다.

IV. Research Trend Analysis

본 장에서는 토픽 네트워크 분석 결과, 중요한 토픽으로 선정된 토픽들이 어떤 연구 주제를 나타내는지 구체적으로 살펴본다. 표 2는 구간별 연결 중심성이 가장 높은 토픽의 상위 키워드들을 나타내고 있다. 모든 표에서 1996년도에서 2000년도까지 A, 2001년도에서 2005년도까지 B, 2006년도에서 2010년도까지 C, 그리고 2011년도에서 2015년도까지를 D로 표기하였다.

Table 2. Top 10 keywords in topics with highest degree centrality

year	subject	Keywords
A	clustering	image, primary, neighborhood, eigenvectors, primitives, paths, optimal, inmemory, birch, clustering
B	Web Application	web, page, news, sites, success, enterprise, page, navigation, informative, content,
C	clustering	clustering, clusters, categorical, density, similarity, hierarchical, mixed, regions, hierarchy, points
D	dimension -ality reduction	rules, traffic, pca, component, principal, confidence, operation, forecast, constraint, evolutionary

표 2는 구간별 연결 중심성이 높은 토픽들의 상위 키워드 및 해당 주제를 나타낸다. 먼저 A 구간에서 연결 중심성이 높은 토픽의 주제는 클러스터링으로 neighborhood, birch, clustering등의 단어가 이에 속한다. B 구간에서 등장한 대표적인 단어들은 특정 데이터 마이닝 기법보다는 웹 응용 프로그램에서 사용되는 단어들이 다수 등장하였다. C 구간에서 연결 중심성이 높은 키워드 역시 클러스터링 기법에 관한 단어들이었으며 특히 계층적 클러스터링을 나타내는 hierarchical 이 등장한 것이 A구간의 클러스터링 기법과 차이라고 볼 수 있다. 마지막으로 D 구간에선 pca, principal, confidence등 dimensionality reduction에 속하는 키워드들이 다수 등장하였다.

Table 3. Top 10 keywords in topics with highest betweenness centrality

year	subject	Keywords
A	Inductive Reasoning	lip, inductive, logic, setting, assumption, handled, implementations, interpretations, propositional, load
B	CRM	fuzzy, products, marketing, promotion, transaction, managers, customer, personalized, shoppers, right
C	Clustering	measures, points, kmeans, relations, interestingness, location, relational, property, sensitivity
D	Sequential Mining	local, sequential, temporal, distributed, parallel, bound, aggregation, communication, operation, partitions

표 3은 매개 중심성이 가장 높은 토픽과 그에 해당하는 주제들을 보여준다. 매개 중심성이 높은 토픽이란 서로 독립적인 두 개의 토픽을 연결해 주는 토픽을 의미한다. 데이터 마이닝 연구 초기를 나타내는 A 구간에서는 inductive reasoning에 해당하는 단어들이 다수 등장하였고, B구간에서는 연결 중심성에서와 마찬가지로 특정 데이터 마이닝 기법이라기 보다는 응용 영역인 CRM에 해당하는 단어들이 주를 이루고 있다. C 구간에서는 연결 중심성이 높은 구간에서와 유사하게 clustering과 관련된 단어들이 많이 등장하였다. 마지막으로 D 구간에서는 시간을 고려한 sequential mining에 해당하는 단어들을 다수 볼 수 있다.

Table 4. Top 30 keywords in topics with highest closeness centrality

year	Subject	Keywords
A	Inductive Reasoning	lip, inductive, logic, setting, assumption, handled, implementations, load interpretations, propositional,
B	Parallel Query Processing	query, base, parallel, mechanism, uncertainty, execution, aggregation, distributed, partitioning, operations
C	Clustering	measures, points, kmeans, relations, interestingness, location, relational, property, sensitivity
D	Sequential Mining	local, sequential, temporal, distributed, parallel, bound, aggregation, communication, operation, partitions

마지막으로 근접 중심성이 높은 토픽들에 대해서 살펴보면 표 4와 같다. 근접 중심성이 높은 토픽은 다른 모든 토픽들과의

거리가 가장 짧은 토픽으로 모든 토픽과 가까이 있는 토픽을 의미한다. A, C, D 구간에서 근접성이 높은 토픽은 매개 중심성이 높은 토픽과 일치하였으며 이는 그림 4에서 살펴본 바와 같이 매개 중앙성과 근접 중앙성이 유사한 패턴을 갖는데서 유추할 수 있다. B 구간의 경우 병렬 및 분산 처리에서 사용되는 단어들이 주를 이루고 있다.

특히 C 구간 즉 2006년도에서 2010년도 사이에 출판된 연구 논문들에서 연결 중심성, 매개 중심성 그리고 근접 중심성이 높은 주제로 클러스터링이 나타났다는 것은 주목할 만하다. 클러스터링은 데이터 마이닝 연구 초기의 연결 중심성이 높은 주제에도 포함되었다.

가장 최근 5년간의 구간인 D 구간에서 연결 중심성이 높은 토픽의 주제는 차원 축소에 해당하는 것으로 나타났는데 최근의 데이터 형태가 다양하고 큰것에서 기인한 것으로 보이며 매개 중심성과 근접 중심성이 큰 토픽은 순차 패턴 마이닝에 해당하는 주제로 시공간 차원을 고려한 연구로 나타났다.

V. Conclusions and Future Work

본 연구에서는 데이터 마이닝 분야의 연구 주제들을 분석하기 위해서 2015년도까지 출판된 20년간의 데이터 마이닝을 주제로 한 연구 논문들을 5개의 관련 저널들로부터 수집하였다. 20년간 연구 주제의 동향을 파악하기 위해서 5년도씩 4개 구간으로 나누어 토픽 네트워크 분석을 실시하였다. 토픽 모델링을 실시하여 30개의 토픽과 토픽에 해당하는 관련 단어들을 선정하고, 공통으로 사용된 단어들의 개수를 연결 가중치로 하여 토픽 네트워크를 구축하였다. 구축된 토픽 네트워크에 소셜 네트워크 분석기법을 적용하여 연결 중심성, 매개 중심성 그리고 근접 중심성이 높은 토픽을 찾아내었다.

토픽 네트워크 구조 분석 결과 데이터 마이닝 연구 초기에는 연결 중앙성, 매개 중앙성, 그리고 근접 중앙성이 모두 최고치를 나타냈으며 이후 5년간 감소하다가 다시 서서히 증가하는 양태를 보였다. 특히 연결 중앙성의 경우 최근 들어 다시 크게 증가하였고, 매개 중앙성과 근접 중앙성은 다소 감소하였다. 이는 최근 5년간 빅데이터를 다루는 데이터 마이닝 연구가 많이 진행됨에 따라 연구 주제들이 다양하고 광범위해짐에 따른 결과로 해석된다.

구간별 연결 중심성, 매개 중심성, 그리고 근접 중심성이 최대인 토픽들을 분석한 결과 클러스터링 기법이 연결 중심성이 높은 토픽에서 다수 발견되었다. 또한 B 구간인 2001년도부터 2005년까지 토픽들은 데이터 마이닝 기법보다는 웹 응용 프로그램, 고객 관계 관리 시스템, 병렬 처리등 응용 영역이나 데이터 처리에 관련된 토픽이 중심성이 높은 토픽으로 선정되었다.

본 연구의 공헌은 사용된 토픽 모델링 알고리즘인 LDA 모델이 토픽간의 연결 관계를 고려하지 못한다는 한계점을 극복하기 위하여 토픽 네트워크를 구축하고 이를 분석하여 토픽들 간의 연결 관계와 영향력이 큰 토픽을 찾아낸 것이다. 또한 토픽 네트워크 분석을 시계열로 분석하여 영향력이 큰 토픽들이 시기에 따라 어떤 경향을 보이는지도 밝혀내었다.

현재 찾아낸 키워드를 보다 논리적으로 해석하기 위해 UniDM 온톨로지를 개발하고 토픽의 키워드를 매핑하여 주제를 해석하는 작업이 진행 중이다. 특히 온톨로지 구조가 계층적 구조임을 감안하여 본 연구에서 사용된 토픽 모델링 방식인 LDA외에 Hierarchical LDA를 적용하여 모델을 확장하는 방안을 연구 중이다. 또한 현재 위키피디아 기사만을 온톨로지의 지식 자원으로 활용하고 있으나 연구 논문이나 특허 문서 또한 온톨로지의 지식 자원으로 활용하고자 한다.

향후 연구는 제안하는 토픽 네트워크 분석 결과의 효용성을 파악하기 위하여 사용자가 본 모델을 실행해 볼 수 있는 웹 사이트를 구축하고자 한다. “data mining" 뿐만이 아니라 “machine learning", "Internet of Things"를 키워드로 연구 논문의 초록을 수집하여 테스트 케이스로 활용하고, 사용자가 토픽의 수와 키워드 수를 변화시켜 가면서 다양한 실험을 직접 실행해 볼 수 있도록 할 예정이다.

또한 토픽의 개수 및 토픽에 대한 단어의 분산을 다양하게 변화시켜 제안하는 토픽 네트워크 그래프가 척도 없는 그래프(scale-free graph)와 랜덤 그래프(random graph) 중 어떤 그래프의 특성을 갖는지 구체적으로 살펴볼 예정이다.

REFERENCES

[1] C. Kim and Y-S. Hong, “Classification Techniques for XML Document Using Text Mining”, Journal of the Korea Society of Computer and Information, Vol. 11, No. 2, May, pp. 15-23, 2006.

[2] J-P. Moon, W-S Lee, and J-H Chang, “A proper folder recommendation technique using frequent itemsets for efficient e-mail classification”, Journal of the Korea Society of Computer and Information, Vol. 16, No. 2, Feb. pp. 33-46, 2011.

[3] D. M. Blei, Y. N. Andrew, and M. I. Jordan, "Latent Dirichlet Allocation," Journal of Machine Learning Research Vol. 3, pp. 993-1022, 2003.

[4] J. Park and M. Song, "A Study on the Research Trends in Library & Information Science in Korea Ifor Information Management, Vol. 30, No. 1, pp. 7-32, March, 2013.

[5] D. M. Blei, "Probabilistic Topic Models," Communications of the ACM, Vol. 55, No. 4, pp. 77-84, April, 2012.

[6] S. Wasserman and K. Faust, "Social Network Analysis: Methods and Applications," Cambridge University Press, 1994.
[7] A. Duvvuru, S. Kamarthi, and S. Sultornsanee, "Undercovering Research Trends: Network Analysis of Keywords in Scholarly Articles," Proceedings of the 9th International Joint Conference on Computer Science and Software Engineering, pp. 265-270, 2012.
[8] Web of Science, "http://isiknowledge.com,"
[9] T. L. Griffiths and M. Steyvers., "Finding scientific topics," Proceedings of the National Academy of Sciences of the USA, Vol. 101 No. 1, pp. 5228-5235, April, 2004.
[10] J. Bae, N. Han, and M. Song., "Twitter Issue Tracking System by Topic Modeling Techniques," Journal of Intelligent Information Systems, Vol. 20, No. 2, pp. 109-122, June, 2014.
[11] D. M. Blei and J. D. Lafferty., "Correlated Topic Models," Proceedings of Neural Information Processing Systems, pp. 147-154, 2005.
[12] Q. Mei et al., "Topic Modeling with Network Regularization," Proceedings of International Conference on World Wide Web, pp. 101-110, 2008.
[13] X-L. Mao et al., "SSHLDA: A Semi-Supervised Hierarchical Topic Model," Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 800-809, 2012.
[14] X. Wang and A. McCallum, "Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends, " Proceedings of the 12th International Conference on Knowledge Discovery and Data Mining, pp. 424-433, 2006.
[15] R, The R Project for Statistical Computing, "https://www.r-project.org/,"
[16] B. Gruen and K. Hornik., "topicmodels: An R Package for Fitting Topic Models," Journal of Statistical Software, Vol. 40, No. 13, pp. 1-29, May, 2011.
[17] C. D. Manning, P. Raghavan, and H. Schuetze., "Introduction to Information Retrieval," Cambridge University Press, pp. 116-121, 2008.
[18] L. C. Freeman, "Centrality in Social Networks: Conceptual Clarification," Social Networks, Vol. 1, pp. 215-239, 1979.

Authors

Hyon Hee Kim received the B.S., M.S. and Ph.D. degrees in Computer Science and Engineering from Ewha Womans University, Korea, in 1996, 1998 and 2005, respectively

Dr. Kim joined the faculty of the Department of Statistics and Information Science at Dongduk Women's University, Seoul, Korea, in 2006. She is currently a Assistant Professor in the Department of Statistics and Information Science, Dongduk Women's University. She is interested in big data analysis, recommender systems, and ontologies.

Hey Young Rhee received the M.S. and Ph.D. degrees in Library and Information from Chung-Ang University, Korea, in 2000 and 2009, respectively

Dr. Rhee joined the faculty of the Department of Library and Information Science at Dongduk Women's University, Seoul, Korea, in 2014. She is currently a Assistant Professor in the Department of library and Information Science, Dongduk Women's University. She is interested in big data analysis.