ISSN 2093-5986(Print)
ISSN 2288-0666(Online)

The Korean Journal of Health Service Management
Vol.11 No.4 pp.213-224 December 2017
https://doi.org/10.12811/kshsm.2017.11.4.213

텍스트마이닝을 활용한 보건의료산업학회지의 토픽 모델링 및 토픽트렌드 분석

조경원[1], 배성권[1], 우영운[2]‡
[1]고신대학교 의료경영학과, [2]동의대학교 응용소프트웨어공학과

Analysis on Topic Trends and Topic Modeling of KSHSM Journal Papers using Text Mining

Kyoung-Won Cho[1], Sung-Kwon Bae[1], Young-Woon Woo[2]‡
[1]Department of Health Care Administration, Kosin University,
[2]Department of Applied Software Engineering, Dong-Eui University

<Abstract>

Objectives : The purpose of this study was to analyze representative topics and topic trends of papers in Korean Society and Health Service Management(KSHSM) Journal. **Methods** : We collected English abstracts and key words of 516 papers in KSHSM Journal from 2007 to 2017. We utilized Python web scraping programs for collecting the papers from Korea Citation Index web site, and RStudio software for topic analysis based on latent Dirichlet allocation algorithm. **Results** : 9 topics were decided as the best number of topics by perplexity analysis and the resultant 9 topics for all the papers were extracted using Gibbs sampling method. We could refine 9 topics to 5 topics by deep consideration of meanings of each topics and analysis of intertopic distance map. In topic trends analysis from 2007 to 2017, we could verify 'Health Management' and 'Hospital Service' were two representative topics, and 'Hospital Service' was prevalent topic by 2011, but the ratio of the two topics became to be similar from 2012. **Conclusions** : We discovered 5 topics were the best number of topics and the topic trends reflected the main issues of KSHSM Journal, such as name revision of the society in 2012.

Key Words : Text Mining, Topic Modeling, KSHSM Journal, LDA

* 이 논문은 2017년 고신대학교 교내연구비에 의하여 연구되었음.

‡ Corresponding author : Young-Woon Woo(ywwoo@deu.ac.kr) Department of Applied Software Engineering, Dong-Eui University

• Received : Nov 20, 2017 • Revised : Dec 19, 2017 • Accepted : Dec 21, 2017

Ⅰ. 서론

보건의료산업학회는 차세대 우리나라의 성장 동력으로 인식되고 있는 보건의료산업분야의 산-학-관 다학제간 연구의 활성화를 유도하여 국민건강 증진에 기여함을 물론 보건의료의 발전에 이바지 하고자 2009년에 설립되었다. 학회의 주요활동 중 하나인 학술지 및 보건의료산업 관련 학술자료 발간은 학회활동의 가장 근간이 되는 활동이며, 학술지를 통해 보건의료산업 분야의 주요한 학문적 주제들과 새로운 창의적 가치를 구현하고자 노력해 왔다. 2007년 “의료경영연구” 학술지로부터 시작된 보건의료산업학회지는 보건의료산업 분야의 대표적 학술지로서 의약학 분야 중 예방의학 부문에서 KCI 인용지수와 관련하여 모든 부문에서 높은 평가를 받고 있는 학회지이다. 또한, 한국연구재단 2015년 현재 KCI 인용지수관련 전 부문에서 1~3위 내의 수준을 유지하고 있으며, 최근 2년을 기준으로 하였을 때에는 국내외(WoS) 통합영향력 지수(IF)에서도 1.17로 1위, 2년간 영향력지수(IF)에 있어서도 1.17로 매우 우수한 평가를 받고 있는 실정이다[1].

현재 학술단체 분류에는 의약학-예방의학-의료관리-병원관리로 분류되어 있으며, 의약학 분야의 보건의료산업(의료서비스 산업, 의료기기기산업, 화장품산업, 의약품산업) 전반에 대한 다양하고 복합적인 다학제적 접근이 가능한 연구들을 포괄적으로 수용하고 있다.

보건의료산업학회지가 주로 게재하고 있는 논문의 주제는 전반적인 보건의료산업 관련영역을 다루고 있다. 우리나라 보건복지부의 주요연구기관인 한국보건산업진흥원 등에서 제시하는 보건의료산업의 개념은 인류의 건강권에 영향을 미칠 수 있는 다양한 요소들, 즉 의료서비스 산업, 의료기기 산업, 화장품산업, 의약품 산업 등을 포괄하여 접근하고 있다. 학술적 가치를 볼 때, 보건의료 분야의 주요 주제를 대상으로 가능한 산업적 영역과 가치로 연계시키는 것을 중심으로 제시하고 있다.

하지만, 2007년부터 현재까지 10년간 게재된 논문들의 주제와 그 비중에 대한 객관적인 검토를 해 볼 기회가 없었다. 따라서 본 연구에서는 그동안 게재된 논문들을 대상으로 주제를 분류하고 주요 주제들의 비중과 10년 동안의 주제들의 변화추이를 파악하는 것을 연구목적으로 한다. 이를 위해 KCI 등재지 논문에 대한 정보공유가 가장 활발하게 이루어지는 한국학술지인용색인을 자료원으로 빅데이터 분석에 주로 활용되는 텍스트마이닝 기법을 수행하였다. 텍스트마이닝은 텍스트의 구조를 파악하고 가장 영향력 있는 개념을 추출하는 것뿐 아니라 이를 시각화하는 데에도 유용하다[2]. 텍스트마이닝은 학술논문의 연구흐름 분석[3]과 뉴스 매체의 빅데이터를 활용한 담론 분석[4]이나 트위터 및 블로그와 같은 소셜 네트워크 분석[5][6], 온라인 리뷰를 통한 고객유형 분석[7][8] 등의 연구에서 활용되고 있다.

토픽 모델링은 비정형 대규모 문헌자료에서 반복적으로 제시되는 주제들을 찾기 위해 사용되는 빅데이터 분석 방법론 중의 하나이다. 일반적으로 글이 작성되는 과정에서 글쓴이가 특정 주제를 염두에 두고 글을 구성하기 때문에 기본적으로 주제어를 반복적으로 사용하게 되는 상황이 된다. 이런 상황을 가정하여 동시에 반복적으로 등장하는 다수의 주제어를 클러스터링 하여 주제그룹을 추정하며, 특정한 문서에서 서로 관련있는 단어들의 집합인 잠재적 토픽을 추출해준다[9].

Sleeman[10]은 특정 과학 분야의 발전을 이해하기 위해서 그 분야의 연구 저작물들에 대한 시간적 경향을 조사하는 것을 제시하였다. 기후 변화에 대한 연구 동향을 분석하기 위하여 30년간 지구과학자들에 의해 출간된 200,000편 이상의 논문들을 대상으로 토픽 모델링을 이용하여 저작물들을 시간대 별로 클러스터 매핑을 수행하는 방법을 활

용하였다.

Shiryaev[11]는 과학 기술 분야의 연구 동향을 파악하기 위하여 웹 페이지에서 수집된 관련 연구 논문들과 LDA(latent Dirichlet Analysis) 토픽 모델링 기법을 활용하였다. 분석 결과 제안된 기법에 의해 컴퓨터가 도출한 분석 결과가 전문가의 평가 결과와 일치함을 알 수 있어 기술 지식 분야의 경향을 파악하기에는 제안된 기법이 적합함을 제시하였다.

Choi[12]는 개인 정보 보호에 대한 학술 연구의 동향을 파악하기 위하여 Scopus 데이터베이스로부터 1972년부터 2015년까지의 저널 논문, 서적, 학술대회 논문 등 모두 2,356편의 문서 초록을 대상으로 LDA 기법을 활용하여 주제 분석을 실시하였다.

Amado[13]는 마케팅 분야에서의 빅데이터 활용에 대한 연구 동향을 파악하기 위하여 텍스트마이닝을 이용한 연구 자료 분석을 수행하였다. 이를 위하여 2010년부터 2015년 사이에 출판된 1,560개의 논문을 수집하여 분석에 사용하였다.

Sun[14]은 교통 분야에 대한 연구 동향을 파악하기 위하여 22개의 대표 저널로부터 1990년~2015년까지의 기사 17,163건을 수집하였다. 수집된 기사 데이터들로부터 LDA 분석을 통해 50개의 핵심 주제가 도출되었다. LDA에 의해 도출된 50개의 주제들은 전문가 평가에 의해서도 대표적인 연구 주제들이며 유의미하다는 것을 검증하였다.

국내에서는 Moon[15]은 비서학논총에 게재된 논문을 대상으로 분석하여 비서학의 지적구조를 파악하였고, Cho & Kim[16]은 산업공학논문지에 게재된 논문을 대상으로 텍스트마이닝을 적용하여 주제어 상관분석을 실시하였고, Park & Song[17]은 문헌정보학 분야의 연구동향을 규명하기 위해 논문초록을 대상으로 토픽 모델링 실험을 실시하였다.

최근 특정 산업의 연구 내용을 총괄적으로 이해하고 분석하는 연구들은 주로 특정 연구 분야에 대한 다년간의 연구 논문들을 빅데이터화하여 LDA 토픽 모델링 기법을 활용하여 접근하는 연구들이 주를 이루고 있다. 보건의료산업 분야에 있어서 새로운 학문적 가치를 구현하고 있는 보건의료산업학회지의 과거와 현재를 조명하고, 연구 동향과 학문적 변화를 분석·연구하는 것은 학회지 발간 10년을 맞는 시점에서 필요한 연구라고 여겨진다.

이에 본 연구에서는 토픽 모델링을 활용하여 보건의료산업학회지의 연구 주제를 분석하였으며, 10년간의 주제에 대한 변화 추이를 파악하고자 하였다. 이러한 시대적 변화에 따른 보건의료 분야의 연구 주제 추이를 파악하는 연구는 향후 보건의료산업학회지의 학문적 방향을 시대변화에 적응할 수 있는 전략을 수립하는 데에 기초자료가 될 것이다.

구체적인 연구목적은 다음과 같다.

첫째, 보건의료산업학회지의 10년간 논문들의 주제를 파악하고 주제들 간의 관련성과 비중을 분석한다.

둘째, 파악된 주제들의 연도별 비중변화 추이를 파악한다.

셋째, IDM(Intertopic Distance Map)을 통해 보건의료산업학회지의 의미 있는 대표 연구주제들을 추출한다.

Ⅱ. 연구방법

1. 자료 수집

보건의료산업학회 논문지의 연구 주제 변화를 분석하기 위하여 학술지가 시작된 2007년부터 2017년 상반기까지 약 10년 동안 논문지에 게재된 논문들의 발간년월, 국문제목, 영문초록, 영문키워드를 논문별로 수집하였다. 이 정보들을 자동으로

수집하기 위하여 Python 웹 스크랩핑 프로그램을 이용하였으며[18], 한국연구재단의 KCI 통합검색 사이트에서 10년 동안 게재된 논문 516편을 자동으로 추출하여 CSV 형태로 저장하였으며 수집된 논문들의 통계 정보는 <Table 1>과 같다.

<Table 1> Statistics of data sets

Year	Volume Number	Freq. (# of papers)	%
2007	1(1)	11	2.13
2008	2(1)	9	1.74
	2(2)	8	1.55
2009	3(1)	8	1.55
2010	4(1)	12	2.33
	4(2)	10	1.94
2011	5(1)	13	2.52
	5(2)	17	3.29
	5(3)	17	3.29
	5(4)	14	2.71
2012	6(1)	21	4.07
	6(2)	21	4.07
	6(3)	19	3.68
	6(4)	25	4.84
2013	7(1)	15	2.91
	7(2)	17	3.29
	7(3)	23	4.46
	7(4)	23	4.46
2014	8(1)	15	2.91
	8(2)	21	4.07
	8(3)	19	3.68
	8(4)	23	4.46
2015	9(1)	14	2.71
	9(2)	9	1.74
	9(3)	22	4.26
	9(4)	15	2.91
2016	10(1)	13	2.52
	10(2)	18	3.49
	10(3)	20	3.88
	10(4)	19	3.68
2017	11(1)	14	2.71
	11(2)	11	2.13
		516	100.00

2. 자료 전처리

자료 분석을 위해서는 Rstudio와 R 언어를 사용하였다. 자료 분석을 위해 가장 먼저 수집된 데이터의 전처리 과정이 필요하다.

이 논문에서는 데이터 전처리를 위하여 가장 먼저 영문초록, 영문키워드의 두 가지 데이터를 한 단위로 처리할 수 있도록 한 문장으로 병합하였다. 그런 후 R 언어의 NLP Library에서 제공되는 영단어 Stemming 함수를 활용하여 단어들의 원형을 추출하여 논문별로 저장하였다. Stemming 추출을 위해서 입력문장의 구두점들과 숫자들을 모두 제거한 후, 소문자로 통일하여 불용어를 제거하였다. 또한, 추출된 단어들 중 빈도수가 비교적 높으나 연구 주제를 파악하는데 관련이 없다고 판단되는 단어(study, purpose, method, data, analysis, result, research, general 등)들을 모두 제거한 후 남은 단어들만을 논문별로 다시 저장하여 전처리 과정을 완료하였다. R 프로그램에서 영단어 Stermming 처리를 위한 관련 함수 적용 과정은 다음과 같다.

```
sp.corpus = Corpus(VectorSource(txt))
  docs <- tm_map(sp.corpus,
removePunctuation)
  docs <- tm_map(docs, removeNumbers)
  docs <- tm_map(docs, tolower)
  docs <- tm_map(docs, removeWords,
  stopwords("english"))
  docs <- tm_map(docs, PlainTextDocument)
  docs <- tm_map(docs, stemDocument,
  "english")
```

3. 자료 분석

토픽 모델링 분석을 위하여 전처리가 완료된 논문 데이터들을 모두 하나의 처리 단위로 하여

LDA 분석 기법을 적용하였다[19]. 토픽 모델링 기법에는 LSA(Latent Semantic Analysis), pLSA(probabilistic Latent Semantic Analysis), LDA(Latent Dirichlet Allocation) 등이 있으나 LDA 기법이 여러 개의 토픽이 내재되어 있는 많은 문서들을 분석하는데 사용 가능하며, 동일한 의미를 지닌 서로 다른 단어, 문맥에 따른 다른 의미의 단어들을 분리, 통합하여 효과적으로 다룰 수 있다는 특징 있다. 또한 Dirichlet 분포를 이용하기 때문에 추출된 토픽들 간의 독립성이 두드러져 여러 자료에 공통으로 나타나는 단어보다 토픽을 결정짓는 특징적 단어들이 나타난다는 장점이 있다[20]. R 프로그램의 LDA 기능에서는 VEM(Variational Expectation-Maximization) 기법과 collapsed Gibbs sampling 기법을 제공하는데, 본 연구에서는 collapsed Gibbs sampling 기법을 이용하였다.

토픽 모델링에서 토픽수를 결정하는 것은 중요한 이슈이다. 원 자료에 뚜렷한 단위가 없는 소설 등에서는 문서 단위를 무엇으로 할 것인지가 중요한 결정 사항 중 하나이지만, 논문 등에서는 대체로 논문 한편을 그대로 한 문서로 사용하므로 토픽의 해석가능성에 따라 토픽의 수를 결정하는 것에 큰 문제가 없다[21]. 이 논문에서는 적절한 토픽의 수를 결정하기 위하여 토픽의 수를 5개에서 30개까지 변화시켜 가면서 LDA 분석 알고리즘의 VEM 기법을 이용하여 R 언어에서 제공되는 perplexity 함수 결과값과 토픽의 해석 가능성, 의미 유용성 등을 고려하여 9개로 결정하였다.

토픽수가 결정된 후에는 LDA 분석 기법에서 collapsed Gibbs sampling 알고리즘을 이용하여 9개의 토픽 별로 빈도수가 높은 상위 15개의 단어들을 최종적으로 추출하였다. 또한, 연도 별로 토픽의 비중의 추이를 분석하기 위하여 2007부터 2017년까지 토픽별 비중을 산출하였다.

추출된 9개의 토픽들의 비중과 유사도를 파악하기 위하여 IDM을 생성하였다. 이를 통해 유사도가 높은 토픽들은 의미 유용성을 고려하여 하나의 토픽으로 통합하였다.

Ⅲ. 연구결과

1. 주제 분석

토픽 모델링의 분석 결과는 <Table 2>와 같다. 분석 방법에서 제시한 바와 같이 토픽 모델링에서 토픽수를 결정하는 것은 중요한 이슈이다. Perplexity 분석결과를 통해 산출된 9개로 토픽수를 설정하고 토픽 모델링을 실시하였다. 첫 번째 토픽은 "의료", "서비스", "요인", "만족", "치료", "환자", "건강", "관리", "사용", "시스템", "고객" ","품질","성과"의 단어를 포함하고 있다. 이들은 보건의료 산업의 주된 주제인 병원 서비스와 관련된 단어들이다. 두 번째 토픽은 "건강", "구강", "치과", "요인", "삶", "우울", "사용", "나이", "행동", "학생", "프로그램", "노인", "사회 ","교육 ","학교"의 단어를 포함하고 있다. 이들은 보건교육 및 행태와 관련된 단어들로 보건교육의 필드들을 포함하고 있으며, 학교 보건이나 지역사회 건강증진에서 다루어지는 건강행위 및 행태와 관련된 단어들이 포함되어 있음을 알 수 있다. 첫 번째 토픽에서의 "요인", "건강", "사용"등의 단어가 두 번째 토픽에서도 동일하게 다수 언급되었다.

세 번째 토픽은 "직장", "조직", "만족", "간호사", "이직률", "스트레스", "일", "문화", "갈등", "조직", "효과", "설문", "위생", "감정 ","리더십"으로 보건조직관리 분야에서 다루어지는 단어들로 구성되어 있으며, 조직관리 연구에서 간호사를 대상으로 한 연구가 다수 이루어졌음을 알 수 있었다.

<Table 2> Extracted 9 topics and their top 15 words, significance ratio

Cluster	Topic	Words	Ratio
1	Hospital service	"hospital","medical","service","factor","satisfaction","care","patient","health","manage","use","system","customer","insure","quality","performance"	0.531
2	Health care education and behavior	"health","oral","dental","factor","life","depress","use","age","behavior","student","program","elder","social","education","school"	0.347
3	Health care organization management	"job","organizational","satisfaction","nurse","turnover","stress","work","culture","conflict","organization","effect","questionnaire","hygiene","emotion","leadership"	0.050
4	Education for professions of healthcare	"nurse","student","clinic","program","education","college","practical","correlation","stress","profession","curriculum","positive","cope","compete","self-concept"	0.026
5	Hospital finance management	"hospital","profit","product","ratio","revenue","performance","manage","expense","efficient","index","cost","personnel","financial","bed","increase"	0.018
6	Elderly healthcare	"function","cognition","disable","live","elder","daily","ADL","therapy","upper","stroke","rehabilitate","active","improve","physical","exercise"	0.011
7	Medical industry	"industrial","induce","effect","employ","healthcare","product","employee","pay","tool","major","use","factor","economic","compare","self"	0.008
8	Hospice	"attitude","death","hospice","volunteer","conscious","nurse","toward","busan","case","experience","organization","safety","nutrition","exist","whether"	0.006
9	Infection control	"infection","control","experiment","drug","april","small","spss","hygiene","internet","offer","collect","goal","mental","respect","association"	0.002

네 번째 토픽도 세 번째 토픽과 유사한 개념의 단어들이 언급되었는데, "간호사", "학생", "클리닉", "프로그램", "교육", "대학", "실용적인", "상관관계", "스트레스", "직업", "커리큘럼", "긍정적" ","경쟁하다 ","자기 개념"을 포함하고 있어 본 연구에서는 보건의료 인력의 영역으로 구분하였다.

다섯 번째 토픽은 “병원”, “수익”, “제품”, “비율”, “이익”, “성과”, “관리”, “비용”, “효율”, “지표”, “인사”, “재정”, “병상”, “증가”를 포함하고 있다. 여기서 추출된 단어들은 병원경영성과지표와 관련된 개념으로 병원재무관리에서 다루어지는 단어들을 다수 포함하고 있다.

여섯 번째 토픽은 "기능", "인식", "비활성화", "삶", "노인", "매일", "ADL", "치료", "위", "스트로크", "재활" ,"신체활동"을 포함한다. 주로 노인의 삶의 특징을 나타내는 단어들이 주로 추출되었고, 노인보건의 영역으로 판단할 수 있다.

일곱 번째 토픽은 "산업", "유도", "효과", "고용", "의료", "제품", "직원", "지불", "도구", "주요", "사용", "요인", "경제 ", "비교" 등을 포함한다. 이는 첫 번째 토픽에서 나온 단어들과 다수 일치하며 의료산업전반에 관한 연구들이 소수 이루어진 것으로

파악된다.

여덟 번째 토픽은 "태도", "죽음", "호스피스", "자원 봉사", "의식", "간호사", "부산", "사례", "경험", "조직", "안전", "영양", "존재" 등을 포함한다. 이 토픽은 극히 작은 군집으로 형성되었으며, 호스피스관련 연구에서 집중적으로 나타나는 단어들로 구성되어 있다.

아홉 번째 토픽은 "감염", "통제", "실험", "약물" 등을 포함하며, 이 또한 여덟 번째 토픽과 함께 극히 작은 군집으로 형성되어 있다. 이는 감염관리관련 연구에서 감염, 통제, 약물 등의 단어가 한 문서에서 집중적으로 나타나 군집을 이룬 것으로 판단된다.

2. IDM 분석

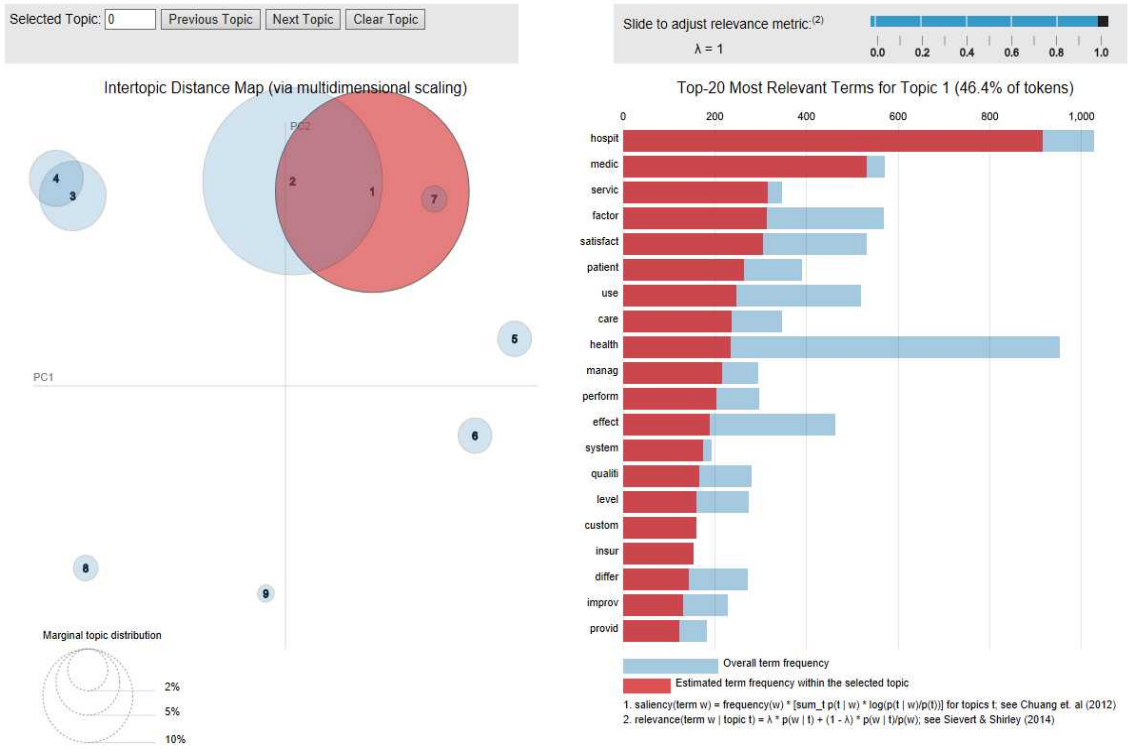
분석 결과 추출된 9개의 토픽에 대한 IDM은 <Figure 1>과 같다. 그림에서 알 수 있듯이 하나의 토픽으로 볼 수 있는 중복된 토픽들이 여러 개로 나뉘어 추출되었다. 중복된 토픽들과 그 토픽들에 대한 실제 추출된 단어들을 분석해 본 결과 토픽 1, 2가 일부 중복되는 것으로 보이나, 키워드들의 의미를 분석한 결과, 다른 주제영역으로 명확히 구별되는 것으로 판단하였다. 그러나 토픽 7은 토픽 1의 일부로 나타나 토픽 1과 7은 하나의 주제인 '의료 서비스'로 정의하였다.

토픽 3, 4는 IDM 상 유사도도 높으며 심층적인 키워드 의미 분석 결과 동일한 주제인 '병원조직 및 인적자원관리'로 정의하였다. 토픽 2는 '보건관리', 토픽 5는 '병원재무관리', 토픽 6은 '노인보건'로 나타났다. 토픽 8과 9는 극소수의 논문에서 집중적으로 반복되어 나타난 키워드에 의해 각기 하나의 주제로 도출되었으나 그 비중이 0.01에도 미치지 못하여 연구 분야로서 하나의 영역으로 간주하기에는 미흡하여 영역으로 구분하지 않았다. 따라서 2007년부터 2017년까지 보건의료산업학회에 투고된 논문들의 연구 주제는 크게 5가지로 구분할 수 있었다<Table 3>.

3. 년도별 토픽 트렌드 분석

토픽 트렌드 분석은 시간의 흐름에 따라 연구자들이 어떤 주제로 보건의료산업학회에 논문을 게재하였는지를 분석하는 것을 의미한다. 이를 통해 게재된 논문들의 대표 주제들의 변화 추이와 지속성을 파악할 수 있다.

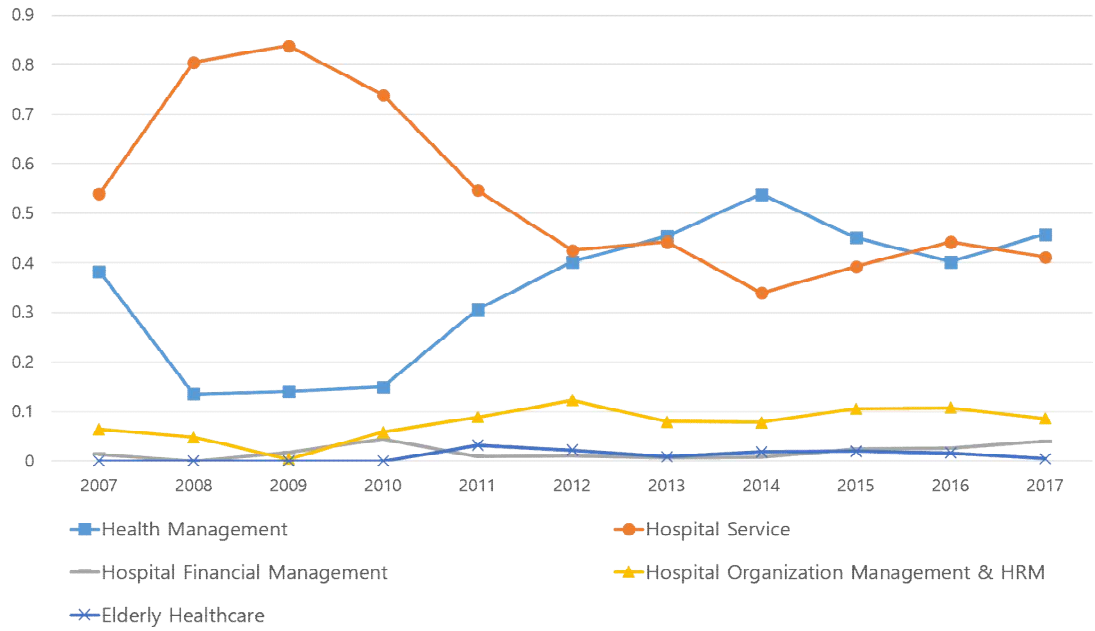
<Figure 2>는 IDM 분석 결과 얻어진 5개의 토픽에 대한 년도별 분포 비중을 그래프로 나타낸 것으로, 시계열 간격은 1년 단위로 분석하였다. 2007년부터 2017년 상반기까지의 분석 결과를 살펴보면, 전체적으로는 'Health management'와 'Hospital service'가 약 80%에서 90%의 비중으로 지배적인 것을 알 수 있었고, 나머지 3개의 주제는 비중이 매우 낮기는 하나 독립적인 주제로서 지속성이 유지되고 있음을 알 수 있었다. 또한 학회가 설립된 초기에는 'Hospital service'의 비중이 거의 대부분을 차지할 정도로 높았고 상대적으로 'Health management' 주제는 다른 비중이 낮은 주제들과 유사한 수준으로 나타났다. 그러나 2012년도를 기점으로 이 두 가지 주제의 비중이 거의 같아졌으며 그 이후에는 유사한 수준으로 유지되고 있음을 파악할 수 있었다.



<Figure 1> Intertopic Distance Map(IDM) created by 9 topics

<Table 3> Revised Topics by IDM Analysis

Topic	Words	Ratio
Medical service	"hospital","medical","service","factor","satisfaction","care","patient","health","manage","use","system","customer","insure","quality","performance","industrial","induce","effect","employ","healthcare","product","employee","pay","tool","major","use","factor","economic","compare","self"	0.538
Health management	"health","oral","dental","factor","life","depress","use","age","behavior","student","program","elder","social","education","school"	0.347
HRM for healthcare	"job","organization","satisfaction","nurse","turnover","stress","work","culture","conflict","organization","effect","questionnaire","hygiene","emotion","leadership""nurse","student","clinic","program","education","college","practical","correlation","stress","profession","curriculum","positive","cope","compete","self-concept"	0.077
Hospital finance management	"hospital","profit","product","ratio","revenue","performance","manage","expense","efficient","index","cost","personnel","financial","bed","increase"	0.018
Elderly healthcare	"function","cognition","disable","live","elder","daily","ADL","therapy","upper","stroke","rehabilitate","active","improve","physical","exercise"	0.011



<Figure 2> Topic trends from 2007 to 2017

Ⅳ. 고찰

보건의료산업학회지는 우리나라 보건의료산업과 관련된 주요한 이슈[22]와 연구 내용을 담고 있는 학회지로서 기존의 특정 보건 분야(병원경영, 보건교육 등)에 한정하여 제공되던 내용보다는 보다 보건의료산업 전반의 연구주제를 포함한 학제간 융합적 연구주제를 수용하고 보건의료산업 전반적인 변화를 주도하고자 노력해왔다.

현재까지 학술지에 게재된 논문들의 30~40%는 의료경영 분야의 주요 주제들을 다루고 있는 것으로 파악해왔다. 또한, 간호학 및 치위생학, 방사선학, 물리치료학 등 다양한 보건산업 영역의 다양한 주제도 전체 논문의 40% 이상 포함되어 있는 것으로 파악해 왔다. 또한, 각 주제에 있어서 산업적 가치에 중점을 두고 심층 연구된 논문들을 지속적으로 게재하고 있었다. 논문 편수에 있어서는 현재 본 학술지는 매호마다 평균 20편 내외의 논문을 게재하는 것을 지향하였다. 학술지 출판의 초기에 비해 지속적으로 투고편수와 게재편수가 증가하고 있는 실정으로 보건의료산업의 다양한 시험적 연구와 접근을 적극적으로 수용해 왔다.

본 연구에서는 토픽 모델링의 주요기법인 LDA와 IDM 분석을 통해 보건의료산업학회의 연구 주제를 실증적으로 분류해 보고자 하였다.

크게 두 가지 결과를 도출하였는데, 첫째로 perplexity 산출을 통해 9개 토픽으로 분석하였으나, IDM의 결과에 따라 일부 통합하여 5개의 토픽으로 최종 분류하였다. 토픽 즉, 연구 주제들로는 의료서비스, 보건관리(보건교육 및 행태), 병원조직 및 인적자원관리, 병원재무관리, 노인보건 영역으로 구분되었다. 이는 학술지 편집에서 의료경영 분야와 간호학, 치위생, 방사선, 물리치료 등 다양한 보건산업영역으로 크게 구분해 온 것과는 다소 차이를 보였다. 텍스트마이닝 기법은 연구자가 자료수집과 결과를 해석하는 데에만 개입을 하고, 분석

처리 과정에서 연구자의 어떠한 가치 판단도 개입되지 않고 완전히 자동화된 방법이다. 따라서 본 연구의 결과가 일반적으로 학문적으로 보건의료산업에 대해 연구하는 전문가들이 인지하는 연구 분야의 범위와 다소 다르게 나타날 수 있다. 그러나 전체적인 분류로 보았을 때는 병원관리, 보건관리, 의료서비스산업, 고령친화산업으로 분류해오던 것과는 어느 정도 일치를 보였다.

둘째로, 보건의료산업을 의료서비스 산업, 의료기기산업, 화장품산업, 의약품산업으로 광범위하게 보았을 때, 보건의료산업학회지의 논문들은 의료서비스 산업에 집중되어 있다. 물론, 작은 군집으로 나왔던 감염관리관련 연구에서 감염, 통제, 약물 등의 단어가 한 문서에서 집중적으로 나타나 독립적인 군집을 이룬 것을 보았을 때 의약품산업의 영역까지 일부 포함하는 것으로 파악되었다. 의약품산업과 같이 화장품산업과 의료기기 산업도 소수의 논문이 게재되었으나, 독립적인 군집을 형성하기에는 그 수가 상대적으로 미미하였다. 따라서 보건의료산업학회지의 보건의료산업의 전반에 대한 다양하고 복합적인 다학제적 접근이 가능한 연구들을 포괄적으로 수용하고자 하는 취지를 살리기 위해서는 토픽으로 분류되지 않았지만, 현재 소수의 논문 게재되고 있는 산업분야의 관심과 투고를 유도할 필요가 있다고 판단한다. 이를 위해 보건의료산업 영역의 범학문적 분야들에 대해 더욱 더 가까이 접근할 있도록, 학회 내의 운영 체제에 대한 개선 방안과 연구 분야의 재정립이 필요할 것으로 판단된다.

본 연구는 보건의료산업학회지의 주제 분류를 위해 보건의료산업 전문가들의 주관적인 개입 없이 텍스트마이닝을 통해 객관적인 결과값을 분석할 수 있었다는 것이 특징이다. 또한, 토픽 모델링이 개별 논문에 대한 주제 탐색이라기보다는 검증하고자 하는 보건의료산업의 학문분야에 대한 전체적인 이해와 통찰을 할 수 있다는 점에서 매우 유용하였다.

셋째로, 토픽 트렌드 분석 결과 전반적으로 'Health management'와 'Hospital service' 주제에 대한 논문들의 게재가 지배적임을 확인할 수 있었다. 다른 3가지 주제들은 상대적으로 비중이 매우 낮기는 하나 정체성을 가지고 차별화된 주제로 자리 잡고 있음을 판단할 수 있다. 학회 설립 초기에는 학회 명칭이 의료경영학회였기 때문에 한동안 'Hospital service' 주제가 주류를 이루었으나, 학회 명칭이 보건의료산업학회로 변경된 이후 2012년부터는 학회에서 지향하는 보건의료산업학회지의 다학제적 접근에 대한 의도가 반영되어 'Health management'와 'Hospital service'의 두 주제가 유사한 비중으로 지속성을 유지하고 있는 것으로 파악한다.

Ⅴ. 결론

본 연구에서는 보건의료분야의 학술지인 보건의료산업학회지를 대상으로 지난 10년간 출판된 논문들의 영문초록과 영문키워드에서 도출된 토픽들의 상관관계를 파악하였다. 또한 LDA를 활용하여 유사한 연구 분야를 나타내는 논문들을 군집화하였다. 그리고 collapsed Gibbs sampling 알고리즘과 IDM 분석을 이용하여 각 주제어들 사이의 관계를 수치화하고 시각화함으로써 정량적 분석을 위한 기초 연구 자료를 도출하였다.

이와 같이 다양한 텍스트마이닝 기법을 통해 전체 논문의 초록과 키워드의 데이터를 분석함으로써 논문이 발간된 이후부터 현재까지의 보건의료산업분야 연구의 현황과 추이를 살펴볼 수 있었다. 본 연구에서는 특정 학술지를 이용하여 분석하였지만, 본 논문에서 제시된 분석기법 및 처리 과정은 다른 분야에도 적용 가능하며, 보건의료 관련기사 및 SNS 데이터를 대상으로 텍스트마이닝과 토픽 모델링의 적용을 통해 보건의료 동향을 관찰하

는 데에도 유용하게 활용될 수 있을 것이다.

보건의료산업학회지는 설립 후부터 지금까지 보건의료산업과 관련된 다양한 주제의 연구들이 투고되어 왔으며, 이를 통해 학회설립의 본래 목적을 적극적으로 수용하며 운영되어 온 것으로 파악된다. 또한 학회의 연구 활동이 활발히 이루어져 의약학 학술 분야를 주도하고 있는 점도 높이 평가할 수 있는 부분이다[1]. 토픽 모델링을 적용하여 과거 10여년의 기간 동안 연구 성과와 의의를 분석하였을 때, 보건의료산업학회지는 의료서비스 분야를 중심으로 하는 보건의료산업 영역에서 주요한 연구 성과를 도출하고 제시하는 데 있어서 큰 역할을 해 왔다고 판단된다. 그러나 이러한 성과와 성취에도 불구하고 학회지가 보건의료산업 전 분야를 모두 다루는 것에는 한계가 있었다. 향후 학회 창립의 정신을 재정립하고 다양한 보건의료산업 분야와의 연계를 위해서는 학회지 운영에 있어서 체계 개선 등의 다양한 노력들이 필요할 것이다. 추후 본 연구에서 도출된 결과의 타당성을 입증하는 연구가 수행될 필요가 있다고 판단되며, 향후 연구에서는 통합 주제와 그에 따른 추가된 세부 주제어들이 보건의료산업학회의 비전과 목적에 적합한 지에 대한 객관적인 검증이 수행되는 것이 필요하며, 학회 임원을 포함한 전문가 패널을 통해 보다 심도 있는 검증이 이루어지는 것이 필요할 것이다.

REFERENCES

1. National Research Foundation of Korea(2017), https://www.nrf.re.kr/
2. D. Paranyushkin(2011), Visualization of text's polysingularity using network analysis, Prototype Letters, Vol.2(3);256-278.
3. Y. Cho, P. Fu, C. Wu(2017), Popular Research Topics in Marketing Journals, 1995-2014, Journal of Interactive Marketing, Vol.40;52-72.
4. I. Flaounas, S. Sudhahar, T. Lansdall-Welfare, E. Hensiger, N. Cristianini(2012), Big Data Analysis of News and Social Media Content, www.see-a-pattern.org
5. D. Scanfeld, V. Scanfeld, E.L. Larson(2010), Dissemination of health information through social networks: Twitter and antibiotics, American Journal of Infection Control, Vol.38(3);182-188.
6. M. Michelson, S.A. Macskassy(2010), Discovering users' topics of interest on twitter: a first look, ACM, pp.73-80.
7. R. Chen, W. Xu(2017), The determinants of online customer ratings: a combined domain ontology and topic text analytics approach, Electronic Commerce Research, Vol.17(1);31-50.
8. Z. Qiao, X. Zhang, M. Zhou, G.A. Wang, W. Fan(2017), A Domain Oriented LDA Model for Mining Product Defects from Online Customer Reviews, Proceedings of the 50th Hawaii International Conference on System Sciences, pp.1821-1830.
9. D.M. Blei, A.Y. Ng, M.I. Jordan(2003), Latent dirichlet allocation, Journal of machine Learning research, Vol.3(Jan);993-1022.
10. J. Sleeman, M. Halem, T. Finin, M. Cane(2016), Advanced Large Scale Cross Domain Temporal Topic Modeling Algorithms to Infer the Influence of Recent Research on IPCC Assessment Reports, Synthesis, Vol.99(81);89.
11. A.P. Shiryaev, A.V. Dorofeev, A.R. Fedorov, L.G. Gagarina, V.V. Zaycev(2017), LDA models for finding trends in technical knowledge domain, IEEE, pp.551-554.
12. H.S. Choi, W.S. Lee, S.Y. Sohn(2017), Analyzing research trends in personal information privacy using topic modeling, Computers & Security,

Vol.67;244-253.
13. A. Amado, P. Cortez, P. Rita, S. Moro(2017), Research trends on Big Data in Marketing: A text mining and topic modeling based literature analysis, European Research on Management and Business Economics, Vol.24;1-7.
14. L. Sun, Y. Yin(2017), Discovering themes and trends in transportation research using topic modeling, Transportation Research Part C: Emerging Technologies, Vol.77;49-66.
15. J.Y. Moon(2009), Understanding the Intellectual Structure of Secretarial Studies with Text Mining, Journal of secretarial studies, Vol.18(1);83-98.
16. S.G. Cho, S.B. Kim, Finding Meaningful Pattern of Key Words in IIE Transactions Using Text Mining, Journal of the Korean Institute of Industrial Engineers, Vol.38(1);67-73.
17. J. H Park, M. Song(2013), A Study on the Research Trends in Library & Information Science in Korea using Topic Modeling, Journal of the Korean society for information management, Vol.30(1);7-32.
18. R. Mitchell(2015), Web scraping with Python: collecting data from the modern web, O'Reilly Media, Inc, pp.1-256
19. J. Silge, D. Robinson(2017), Text mining with R: A tidy approach, O'Reilly Media, Inc, pp.1-194
20. J. Kim, S. Baek(2016), Analysis of Issues on the College and University Structural Reform Evaluation Using Text Big Data Analytics, Asian journal of education, Vol.17(3);409-436.
21. B. Grun, K. Hornik(2011), topicmodels: An R package for fitting topic models, https://cran.r-project.org/web/packages/topicmodels/vignettes/topicmodels.pdf
22. J. Ahn, J. Suh(2017), Economic Effects of South Korea's Smart Healthcare Industry, Korea Journal of health service management, Vol.11(2);55-64.