Journal of The Korea Society of Computer and Information
Vol. 23 No. 12, pp. 153-161, December 2018
www.ksci.re.kr
https://doi.org/10.9708/jksci.2018.23.12.153

Trend Analysis of Thyroid Cancer Research in Korea with Text Mining Techniques

Tae-Gyeong Lee*, Seong-Min Heo*, Seung-Hyeok Shin*, Ji-Yeon Yang *

Abstract

In this paper, we propose a text-centered approach to identify the research trend of thyroid cancer in Korea. We incorporate statistical analysis, text mining and machine learning techniques with our clinical insights to find connective associations between terminologies and to discover informative clusters of literatures. The incidence of thyroid cancer in Korea increased rapidly in the 2000s, which fueled the debate regarding overdiagnosis, but recently the number of patients undergoing surgery has decreased significantly due to conscious reform efforts from various circles. We analyzed the abstracts and keywords of related research papers from DBpia. It was found that most were case reports in the 1980s, and some papers in the 1990s discussed the early detection of thyroid cancer by mass screening. While many papers focused on different diagnostic techniques and the detection of small cancers in the 2000s, many emphasized more on the quality of life of patients in the 2010s. There was an apparent change in the topics of thyroid cancer research over past decades. The results of this study would serve as a reference guide for current and future research directions.

▸Keyword: hierarchical clustering, social network, text mining, thyroid cancer, word cloud

I. Introduction

국내 갑상선암의 발생은 1990년대 동안 서서히 증가하다가 이후 2000년대에 들어서며 급격히 증가하였다. 2011년 갑상선암 진단 비율은 1993년에 비해 15배나 증가하였으며, 2009년부터 2014년까지 국내 암 발생순위 1위를 기록하였다[1]. 2012년에는 10만명 당 52.8명으로 세계에서 제일 높은 발생률을 나타냈는데(ASR, Age-Standardized Rates), 이는 두 번째로 높은 뉴칼레도니아의 10만명 당 22.2명(ASR) 보다 2배 이상 높은 수치이다[2]. 이러한 현상은 한때 세계적인 이목을 집중시킨 바 있다[3-5]. 연구진들은 지난 20~30년간 갑상선암 발생률이 급증한 주요 요인으로 의료기술 발전으로 인한 과잉진단(overdiagnosis)을 꼽았으며 한국이 그 대표적 사례라고 밝히고 있다. 반면 발생률이 급증한 것에 비해 갑상선암으로 인한 사망률은 크게 변화가 없는 것으로 보고된다[3].

이러한 논란에 휩싸인 이후, 국내 갑상선암 수술 환자 수는 급격히 줄어 2011~2013년 동안 연간 4만~5만 건에 달하던 수술 환자 수는 2014년~2016년 연간 2만~3만 건으로 급격히 감소한 것으로 나타났다[6]. 하지만 여전히 2015년 현재 암 발생환자 214,701명 중 갑상선 암 발생환자의 수는 25,029명에 달한다. 전체적으로 위암, 대장암 다음으로 가장 높은 발생률이며, 여성만 보았을 때에는 암발생률 1위를 기록하고 있다[1].

이에 따라 지난 수십 년간 갑상선암에 대한 사람들의 관심이 높아지고, 관련 논문도 증가함을 볼 수 있다. 본 연구에서는 이러한 관련 논문들의 초록과 키워드를 분석하여 갑상선암 연구의 시대별 동향을 파악하고 활발한 연구 활동을 위한 기초 자료를 제공하고자 한다. 반면 초록과 키워드 자료의 비정형성으로 인해 전통적인 자료 분석방법에 추가적으로 텍스트

• First Author: Tae-Gyeong Lee, Seong-Min Heo (Lee and Heo contributed equally to the work),
Corresponding Author: Ji-Yeon Yang
*Tae-Gyeong Lee (leetk97@kumoh.ac.kr), Dept. of Applied Mathematics, Kumoh National Institute of Technology.
*Seong-Min Heo (cjsm03@kumoh.ac.kr), Dept. of Applied Mathematics, Kumoh National Institute of Technology.
*Seung-Hyeok Shin (shinbaad@kumoh.ac.kr), Dept. of Applied Mathematics, Kumoh National Institute of Technology.
*Ji-Yeon Yang (jyang@kumoh.ac.kr), Dept. of Applied Mathematics, Kumoh National Institute of Technology.
• Received: 2018. 10. 18, Revised: 2018. 11. 09, Accepted: 2018. 11. 10.
• This research was supported by Kumoh National Institute of Technology(2018-104-066).

마이닝(text mining) 방법론을 활용하여야 한다. 텍스트 마이닝은 텍스트로 이루어진 데이터에서 형태소 분석과 자연 언어처리(natural language processing) 기술을 이용해서 텍스트 안에서 패턴 또는 관계를 추출하여 지식을 발견하는 일련의 과정을 총칭한다[7].

다양한 분야에서 이러한 텍스트 마이닝을 기반으로 대용량의 텍스트 데이터를 탐색, 분석해 오고 있다. [8]은 기계 학습 분야의 특허를 수집하여 핵심 키워드를 추출하고 이를 토대로 키워드 네트워크를 구축한 후, 클릭분석을 실시하여 기계 학습 분야의 특허 경향을 파악하고 있다. [9]는 인터넷 뉴스 기사에 대한 빈도 분석 및 연관관계 분석을 통해 기후변화 및 식품관련 정보가 어느 정도의 연관성을 가지고 얼마나 자주 나타나고 있는지를 검토하고 있다. [10]에서는 SNS에서 수집한 충북 관광에 관한 글들을 이용하여 이슈분석, 연관분석, 감성분석을 수행하였으며, 이를 통해 관광진흥정책의 수립에 필요한 기초자료를 제시하고 있다. [11]에서는 국내 남자 프로배구 경기의 문자중계 데이터를 이용한 소셜네트워크 분석을 통해 경기력과 관련된 핵심 키워드를 추출하고 이를 토대로 배구 구단의 경기력을 분석하고 있다.

의생물학에서도 텍스트 마이닝 기법을 활용하여 추출된 정보를 각종 기초 자료와 결합함으로써 한층 연계된 의료 정보를 구성하려는 시도가 많이 이루어지고 있다. 예를 들어, [12]에서는 웹에 분산되어있는 의료정보를 활용하여 텍스트 마이닝을 통해 파악한 질환별 판별요인을 개인 맞춤형 정보 형태로 제공하는 서비스를 소개한다. [13]에서는 TF-IDF 기반의 가중치를 고려하여 진료 소견데이터로부터 학습된 질병들의 증상을 이용해 질병명을 추론하는 시스템을 제안하고 있다. 또한 다량의 기존 연구문헌을 통해 특정 질병에서 DNA, RNA, 단백질등과 같은 생체분자들 간의 관계와 약물과 질병 간의 관계 정보를 자동으로 추출하려는 노력 역시 활발히 이루어지고 있다. [14]는 생물학 문헌들을 활용하여 문헌별로 지역 유전자 네트워크를 구성하고, 병합하는 과정을 통해 하나의 전역 유전자 네트워크를 구축하는 방법을 제안하고 있으며, 이를 통해 질병과 관련한 유전자를 추론할 수 있다. [15]에서는 생의학 분야의 논문을 활용하여 유전자-단백질-질환의 연관 관계를 식별할 수 있는 시스템을 제안하고 있다. [16]에서는 문헌 자료의 초록으로부터 약물과 질병의 관계 및 약물과 유전자의 관계를 식별한 후, 질병 기반 약물 유사도와 유전자 기반 약물 유사도를 측정하고 있다. [17]에서는 췌장암 관련 문헌을 이용하여 유전자-단백질 상호작용 네트워크를 구성하고, 관련 연구에서 주요하게 언급되는 유전자-단백질의 유발관계 사슬을 파악하고 있다.

본 연구에서는 지난 수십 년간 사람들의 관심이 증가한 갑상선암에 초점을 맞추어, 텍스트 마이닝 기법을 사용하여 해당 분야의 연구동향을 파악하고 있다. 이를 위해 관련 국내 논문들의 초록과 키워드를 수집·분석하고, 시대별로 연구 주제의 관심사가 어떻게 변화했는지, 연구 주제들이 어떠한 상관관계가 있는지를 살펴보고자 한다. 이를 통해 갑상선암에 대한 의료진들의 임상적인 시각 변화가 실제 연구 방향의 변화로 이어졌는지 살펴볼 수 있을 것이다. 또한 향후 갑상선암 연구의 방향 모색에 필요한 기초자료로 활용될 수 있으리라 판단된다. 본 연구의 구성은 다음과 같다. 제2장에서는 자료의 수집과 분석 방법에 대해 기술하고, 제3장에서는 자료 분석 결과를 제시한다. 마지막 4장에서는 본 연구의 결론으로 연구의 의의, 그리고 향후 필요한 연구과제에 대해 논의하고 있다.

II. Data Collection & Methods

2.1 Data Collection

본 연구 분석에 사용된 자료는 디비피아(http://www.dbpia.co.kr/)에서 "갑상선암"을 키워드로 검색하여 수집된 논문이다. 2018년 5월 24일자에 의약학 239개, 공학 10개, 사회과학 7개, 자연과학 3개, 농수해양학 1개 총 260개의 논문이 검색되었고 그 중 학술대회, 보고서, 한글초록 등 관련 없는 158개의 논문을 제거하여 102개의 논문에서 제목, 영문초록, 키워드, 학회지, 학회지 종류, 연도 등을 추출하였다.

데이터 전처리 과정으로 구두점, 공백, 기호, 불용어를 제거하였으며, 추가적으로 오타 수정, 소문자 변환을 시행하였다. 하지만 약어나 고유 명사의 경우에는 대문자로 처리하였으며, 엔그램(n-grams)의 경우에 단어 간 "_"로 연결하였다. 반면 출현 빈도가 매우 높지만 분석 결과에는 영향을 미치지 않는 단어들(thyroid, cancer, carcinoma 등)은 사전에 제거하였다. 그 결과, 초록에서 319개의 단어와 키워드에서 181개의 단어를 얻었다.

2.2 Methods

우선 학회지의 연구 분야별 논문 수의 분포 및 초록의 길이를 검토하였다. 또한 연구 주제의 추세를 살펴보기 위해, 논문 초록에 포함된 키워드를 대상으로 시대별 워드 클라우드(word cloud)를 구성하였다. 워드 클라우드는 문서에 사용된 단어의 빈도를 표시하는 시각화 유형이다. 그림에 크게 나타나는 단어가 상대적으로 사용 빈도가 높은 단어이다. 본 연구에서는 시대별 워드 클라우드를 시각화하여 비교함으로써 연구 주제의 시대적 변화를 쉽게 파악할 수 있다.

반면 단어 간 관계를 살펴보기 위해 소셜네트워크 분석(social network analysis)을 실시하였다. 단어 간의 유사성은 두 단어가 문서 또는 텍스트 단위에서 동시에 출현(co-occurrence)하는 정도를 통해 측정된다[18]. 본 연구에서는 키워드를 이용하여 연관성을 보이는 중요한 단어들의 관계를 파악하고 있다. 특히 각 키워드를 나타내는 노드(점, node)와 키워드 간의 동시출현 관계를 의미하는 링크(선, link)로 표현되는 소시오그램(sociogram)을 제공함으로써, 각 키워드가 네트워크 내의 다른 키워드들과 얼마나 근접하게 연결되어 있는지 그리고 중심이 되

는 영향력을 가진 키워드가 무엇인지 파악할 수 있다. 또한 네트워크의 중범위 수준에서 키워드들이 어떻게 군집되어 있는가를 분석하고 해당 군집의 주제어를 제시하였다.

또한 자율 군집(unsupervised clustering)의 한 방법인 계층적 군집분석(hierarchical clustering)을 이용하여 단어들이 어떻게 군집을 이루는지 검토하였다. 계층적 군집분석은 각 개체의 유사성을 측정하여 가장 유사한 개체를 묶어 나가는 과정을 반복하여 원하는 개수의 군집을 형성하는 방법이다. 개체 간 유사성(또는 거리)의 정의 및 군집 간 연결하는 방식은 다양하다[19, 20]. 그 중 본 분석에서는 키워드-문서 행렬을 구성한 후, 키워드 간의 비유사성을 코사인(cosine) 거리로 측정하였으며 군집내의 오차제곱합에 기초한 와드연결법(ward linkage)을 이용하여 군집을 형성했다. 분석결과를 나무그림(dendrogram)으로 시각화하여 연관이 높은 키워드를 파악하였다. 그런 후, 이를 바탕으로 각 군집의 특성을 파악하고 연구주제를 추론하였다.

끝으로 피어슨 상관계수(Pearson correlation coefficient)를 이용한 상관관계 분석을 통해 주요 용어와 연관이 높은 단어를 살펴보았으며, 시대별로 많이 등장하는 통계 기법을 검토하였다.

III. Results

3.1 General Characteristics

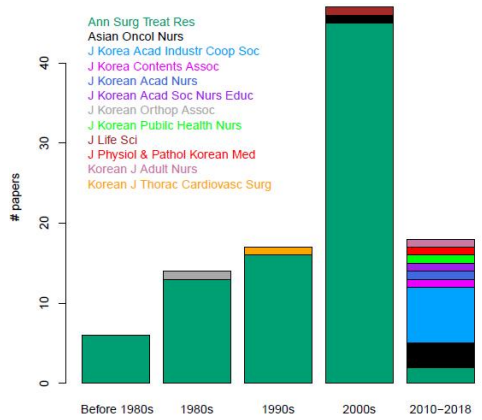

Fig. 1. The number of publications in domestic journals over time

디비피아에서 검색되는 갑상선암 관련 국내논문은 2018년 5월 24일 기준으로 1965년부터 2018년까지 총 102편이었다. 이 중, 그림 1에서 보듯이, 1980년 이전에는 6편, 80년대에는 14편, 90년대에는 17편이다가 2000년대에는 47편까지 증가하였다. 이후 현재까지 18편의 논문이 추가적으로 게재되었다. 2000년대까지 대부분의 논문은 대한외과학회지(Annals of Surgical Treatment and Research)에 게재되었으며 2010년대에 들어서 다양한 학회지에 게재됨을 볼 수 있는데, 이는 갑상선암 연구와 관련해서 최근 연구진과 주제의 다변화를 짐작할 수 있게 한다. 이러한 점은 표 1에서도 확인할 수 있다. 표 1은 연구 분야별 발표 논문 수의 분포를 보여주는데, 2000년대까지 1편을 제외한 모든 논문이 의약학에 속하는 학회지에 발표되다가, 2010년대에 들어서 44.4% 가량의 논문이 한국산학기술학회논문지(Journal of Korea Academia-Industrial Cooperation Society), 한국콘텐츠학회논문지(The Journal of the Korea Contents Association) 등 공학 학회지에 발표되었음을 확인할 수 있다. 한편 표 1의 오른쪽 부분에서 영문 초록의 길이를 확인할 수 있는데, 이는 숫자, 기호, 불용어를 제거한 영문 초록의 단어의 개수를 의미한다. 의약학 논문을 대상으로 봤을 때, 초록의 길이는 점점 증가하는 추세를 보이다가 2010년대에 다시 감소세로 돌아섰다. 만일 게재 논문의 개수 및 초록의 길이와 같은 양적 특성이 시대적 연구 트렌드를 반영한다면, 갑성선암 연구에 대한 관심도는 2000년대에 정점을 찍고 차츰 감소하는 추세를 보이고 있다.

3.2 Trend of Research Topics

본 절에서는 논문 초록에 포함된 키워드를 분석함으로써 갑상선암 연구의 현황 및 시대별 추이를 살펴보고 있다. 1980년대 이전의 논문들(6편)과 일부 1980년대 논문들(10편)은 키워드를 포함하고 있지 않아 본 절의 분석에서 제외한다.

1980년대에는 4편의 논문이 분석에 포함되었으며, 주요 단어는 parathyroid(부갑상선), incidence(발생), I-131(방사성 옥소 치료방법), thyroidectomy(절제술), lobectomy(엽절제술), near_total(근전절제술) 등으로 나타났다(그림 2 참조). 이 당시에는 갑상선암의 증례, 후향적 보고(retrospective report)가 대부분으로 갑상선암 치료 및 수술에 대한 기술이 주를 이루었다.

반면 1990년대의 논문에는 thyroidectomy(절제술), dissection(청소술), surgery(수술), total(전절제), treatment(치료), papillary(유두상암), lymph-node(림프절), recurrence(재발), metastasis(전이) 등의 단어가 많이 등장하는데, 가장 많이 발생하는 아형인 papillary(유두상암)에 대해서 림프절 전이 및 재발 여부에 따른 외과적 수술범위와 방법에 대한 논의가 많이 이루어진 것으로 보인다. 동시에 complication(합병증)의 단어는 수술과 치료 이후 나타나는 합병증이 검토되고 있음을 알 수 있다. 또한 일부 논문에서는 Graves' disease(Graves병), Hashimoto's(하시모토 갑상선염) 같은 갑상선 질환을 같이 검토하고 있다. mass(집단), screening(검진), early(조기) 등의 단어는 건강 검진을 통한 갑상선암의 조기 진단이 논의되고 있음을 짐작하게 한다.

2000년대에는 많은 연구들이 다양한 주제로 진행되었는데, 그 중 대부분은 좋은 예후를 보이는 differentiated(분화암), papillary(유두상암)을 대상으로 한다. 당시에도 lymph-node(림프절), distant(원격), metastasis(전이), recurrence(재발),

Table 1. The number of publications and the average length of abstracts by discipline of the journals

	Number of publications			Average length of abstracts		
	Engineering	Medicine & pharmacy	Natural science	Engineering	Medicine & pharmacy	Natural science
Before the 1980s	0	6	0	NA	98.2	NA
1980s	0	14	0	NA	113.0	NA
1990s	0	17	0	NA	136.7	NA
2000s	0	46	1	NA	143.6	118.0
2010-2018	8	10	0	103.5	114.6	NA

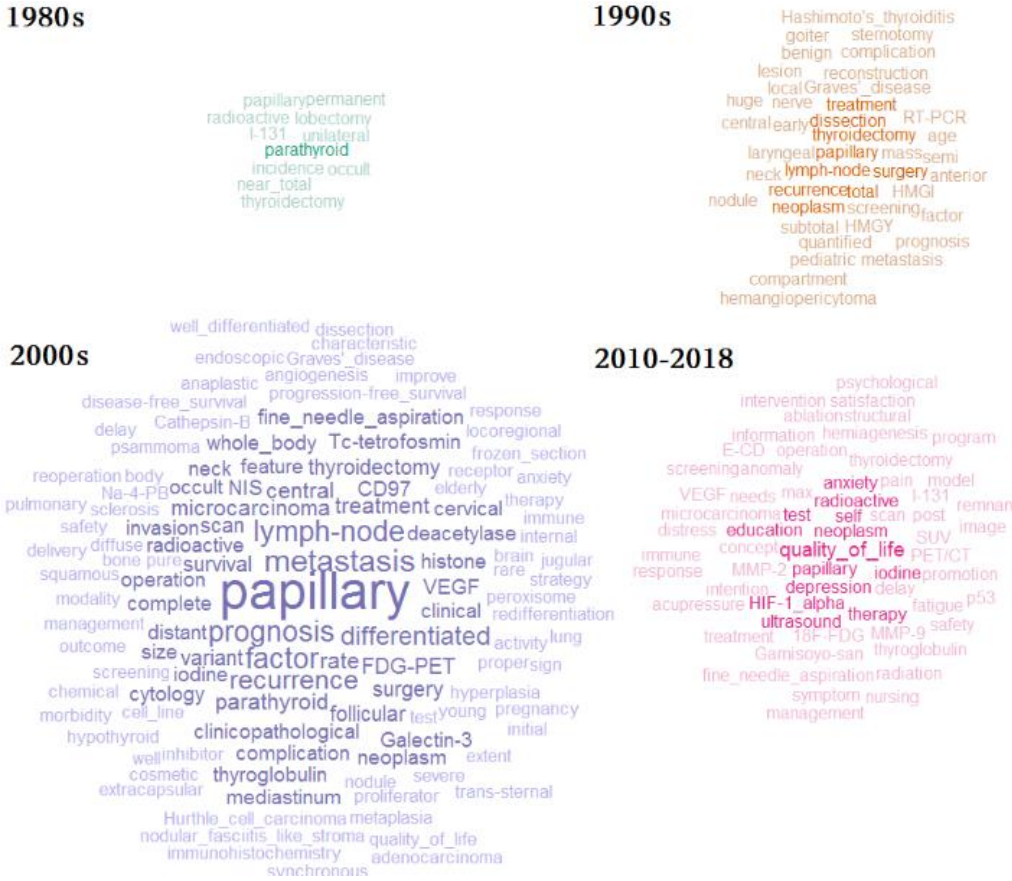

Fig. 2. Word clouds of keywords by decade

invasion(침윤), thyroidectomy(절제술), complete(잔존갑상선절제술) 등의 단어가 많이 나타났는데, 이는 절제 범위 및 수술 방법을 결정하는데 국소침윤, 림프절 전이, 진행속도 등이 영향을 주고 있음을 추론할 수 있다. 2000년대에 들어와서는 whole_body(전신), scan(스캔), FDG-PET(양전자단층촬영방법), Tc-tetrofosmin(핵의학 영상검사방법), thyroglobulin(갑상선글로불린 수치 검사), fine_needle_aspiration(세침흡인검사), cytology(세포검사) 등의 단어가 많이 등장하는데, 암의 진단 및 재발의 진단을 위한 여러 가지 검사방법에 관한 연구들이 이 당시 많이 이루어졌음을 알 수 있다. 더불어 occult(잠재성암), microcarcinoma(미소암), size(크기) 등의 단어들을 통해 이 당시에 정밀한 장비들의 도입으로 미세하고 치명적이지 않은 갑상선 이상 발견이 증가하였음을 추측할 수 있다. 또한 prognosis(예후), factor(요인), survival(생존), rate(비율), clinicopathological(임상병리학적), feature(특징) 등의 단어들은 생존분석(survival analysis)이 많이 이루어졌음을 알 수 있다. 뿐만 아니라 VEGF(혈관내피성장인자), histone(히스톤), deacetylase (탈아세틸화효소), CD97 (CD97 mRNA), Galectin-3 (Galectin-3 단백질), NIS (sodium iodide symporter mRNA) 등 생물학적 객체를 나타내는 단어가 많이 나타난 점으로 보아 마이크로어레이(microarray) 데이터 분석 역시 활발히 이루어졌음을 추측할 수 있다. 기타 아형의 변이(variant) 역시 활발히 연구되어졌다.

2010년대에는 quality_of_life(생활의 질)이라는 단어를 중심으로 환자의 복지에 관한 연구가 많이 이루어졌다. 관련된 단어로 anxiety(불안), depression(우울)도 많이 사용되었다. 또한 therapy(치료), iodine(요오드), radioactive(방사성의), education(교육), self(자가), test(검사) 등의 단어가 자주 쓰인 점을 미루어보아, 수술과 방사성요오드 치료에 대한 정보 및 교육 제공 그리고 자가진단에 대한 내용이 많이 검토되어졌음을 알 수 있다. 그리고 일부 논문에서 HIF-1_alpha(HIF-1 α) 발현의 임상적 연구가 이루어졌음을 알 수 있다.

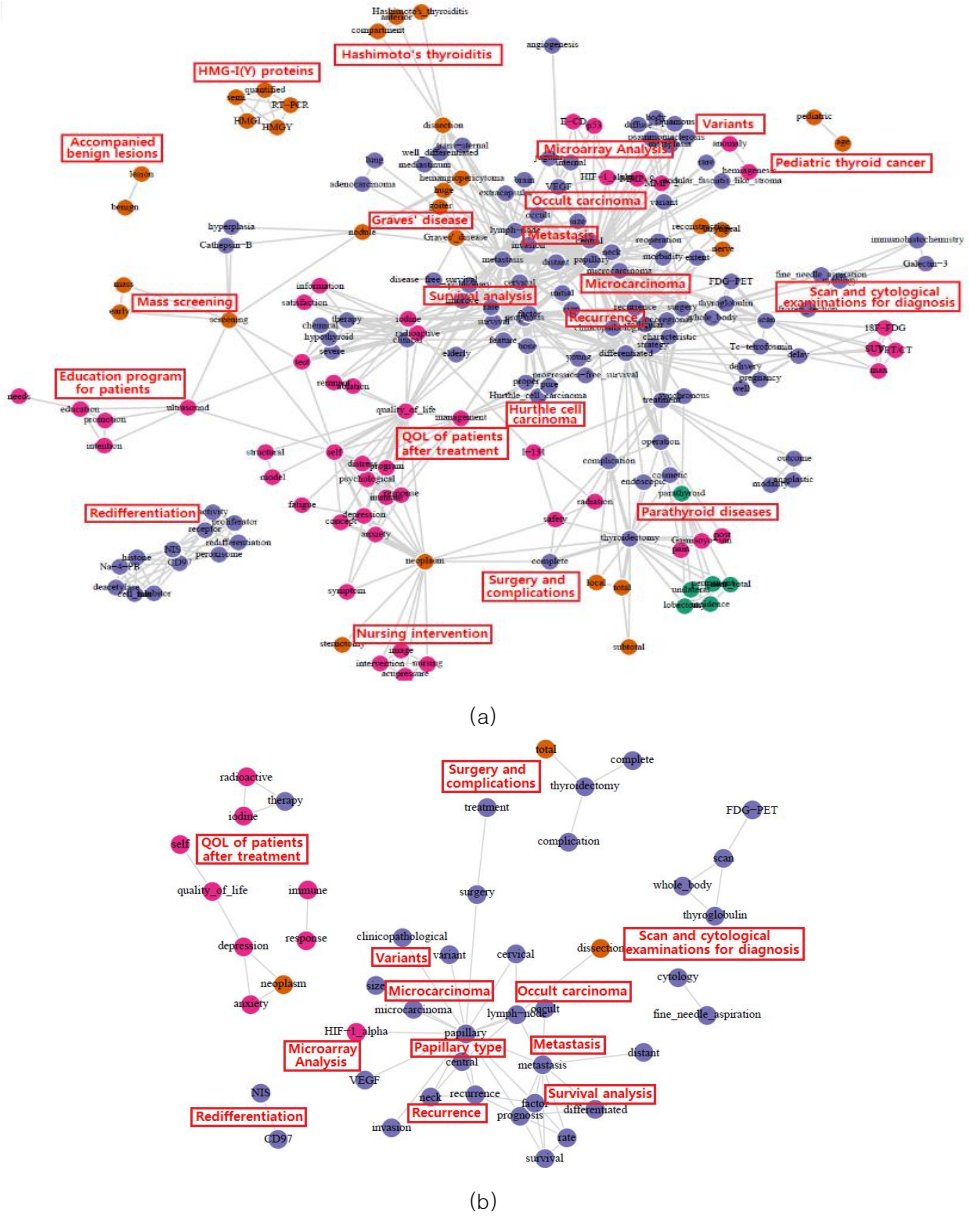

Fig. 3. (a) Complete and (b) simple social networks of keywords with cluster annotations

3.3 Associations between Terminologies

본 절에서는 소셜네트워크와 계층적 군집분석을 이용하여 대상 논문 초록에 포함된 키워드 간 유사성을 측정하고 연구 주제 네트워크를 구축하며, 상호 연결성이 높은 단어들을 군집화하고자 한다. 이를 통해 주제별 연구의 상호관련성과 주제별 핵심어를 파악할 수 있을 것이다. 논문의 키워드를 대상으로 한 소셜네트워크 분석결과가 그림 3에 제시되었다. 패널 (a)의 네트워크는 모든 키워드를 대상으로 한 반면, 패널 (b)의 네트워크는 차수가 0인 정점(vertex)과 가중치가 1 이하인 엣지(edge)들을 제거한 46개의 용어만을 대상으로 한 간단한 네트워크이다. 각각의 용어에 대해서 시대별 등장 빈도를 확인한 후, 가장 많이 나타난 시대를 반영하고자 정점의 색을 달리 했다. 10년 단위 시대별 색은 그림 2의 워드클라우드의 색과 동일하다(초록: 1980년대, 주황: 1990년대, 파랑: 2000년대, 분홍: 2010-2018). 전 절에서 발견한 결과와 비슷한 패턴을 확인할 수 있는데, 2000년대에는 많은 연구들이 다양한 주제로 진행된 반면 2010년대에는 환자의 복지, 교육 프로그램, 간호 중재에 대한 연구가 활발하였음을 알 수 있다.

전반적으로 네트워크는 papillary(유두상암)을 중심으로 다양한 연구 주제들이 고르게 연결되어 있는 형태이며, 이와 별도로 환자의 삶의 질(QOL of patients), 소아 갑상선암(pediatric thyroid cancer), 재분화(redifferentiation) 등의 주제어 그룹이 나타났다. 특히 패널 (b)의 간단한 네트워크를 살펴보면, 환자의 삶의 질 관련 주제가 비교적 큰 독립적인 군집을 형성하고 있다. 이러한 패턴은 그림 4의 나무그림(dendrogram)에서도 확인할 수 있다. 나무그림은 단순 네트워크에 포함된 46개의 용어를 대상으로 코사인 유사도(cosine similarity)와 워드링크법(ward-link method)을 사용한 결과

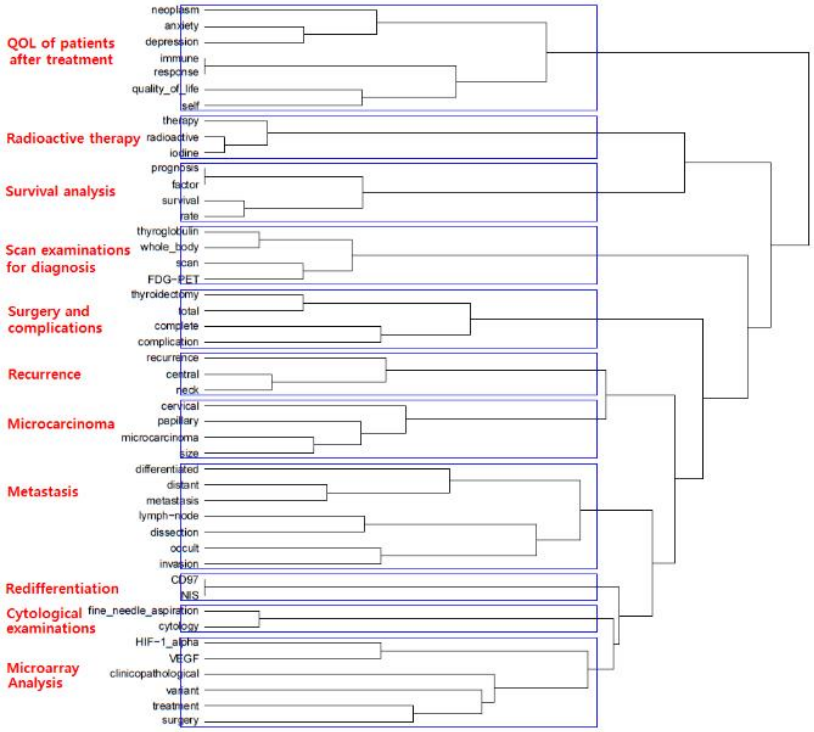

Fig. 4. Dendrogram from hierarchical clustering of keywords with cluster annotations

Table 2. The top correlated words with specific main terminologies

	Text type	Top five correlated words
thyroidectomy	Keywords	total, subtotal, complete, complication, pain
	Abstracts	complete, total, transient, morbidity, permanent
radioactive	Keywords	iodine, therapy, information, satisfaction, severe
	Abstracts	iodine, therapy, attention, coping, content
metastasis	Keywords	distant, brain, lymph-node, bone, mediastinal
	Abstracts	distant, lymph-node, cervical, organ, risk
scan	Keywords	whole_body, FDG-PET, thyroglobulin, 18F-FDG, PET/CT
	Abstracts	whole_body, specificity, Tc-tetrofosmin, discontinuation, T4
factor	Keywords	prognosis, rate, survival, progression-free_survival, elderly
	Abstracts	prognosis, univariate, multivariate, independent, entity
quality_of_life	Keywords	depression, self, concept, fatigue, structural
	Abstracts	stepwise, data, depression, intervention, regression

이다. 이를 보면 환자의 삶의 질이 별도의 군집을 이루고 있으며, 나머지 주제들은 하나의 큰 군집에 속하는데 이 중에는 수술 및 합병증, 재발, 전이, 스캔과 세포 검사, 미소암, 재분화, 생존분석, 마이크로어레이 분석 관련 주제들이 포함된다. 갑상선암의 연구주제는 2010년대 이후를 기점으로 큰 변화가 있었음을 알 수 있는데, 이전에는 임상병리학적 특성에만 주된 초점을 두었다면 이후에는 환자의 복지까지 고려하고 있다.

표 2에서는, 각각 초록과 키워드를 대상으로 할 때, 일부 주요 용어와 가장 상관관계가 높은 용어들을 보여주고 있다. 그 중 일부를 살펴보면, thyroidectomy(절제술)은 수술 부위, 수술방법, 일시적, 영구적, 합병증, 고통과 같은 단어와 높은 상관관계를 보이고 있다. metastasis(전이)는 경부 림프절, 종격동을 비롯한 뼈, 뇌, 조직 등 원격전이 단어와 많이 쓰이고 있다. 반면 요인(factor)은 나이, 생물학적 객체와 관련이 있음을 알 수 있고, 예후, 무진행생존율, 일변량, 다변량 생존분석과 많이 사용되고 있다.

3.4 Statistical Methods Used Over Time

본 절에서는 초록에 나오는 단어들을 바탕으로 많이 사용된 통계 기법을 살펴보고 있다. 물론 논문에서 사용된 모든 통계적 기법이 초록에 언급되지 않을 수 있기 때문에, 정확한 결과라고 보기는 어렵다. 게다가 표 3에 보듯이, 구체적인 통계적 방법의 언급 없이 statistical(통계의), p_value(p값)라는 단어만 등장하는 경우가 많아 사용된 통계적 처리 방법을 전부 파악하기에는 무리가 있다. 따라서 전반적인 시대별 추세만 살펴보는 데 의의를 두고자 한다. 주요 통계적 기법을 나타내는

Table 3. Frequency of appearances of statistical related terms in abstracts

	Before the 1980s	1980s	1990s	2000s	2010-2018
p_value	0	0	1	13	6
statistical	0	0	2	4	4
Chi-square	0	0	2	1	1
Fisher's_exact_test	0	0	0	1	0
t-test	0	0	0	0	2
t_value	0	0	0	1	0
pre-post_test	0	0	0	1	2
Duncan's test	0	0	0	0	1
Sheffe_test	0	0	0	0	1
one-way	0	0	0	0	2
ANOVA	0	0	0	0	2
Pearson	0	0	0	0	3
Cox	0	0	0	1	0
log-rank	0	0	0	4	0
multivariate	0	0	0	5	0
univariate	0	0	0	5	0
regression	0	0	0	1	4
stepwise	0	0	0	0	3
Wilson_and_Cleary_model	0	0	0	0	1
structural_model	0	0	0	0	1
confidence_interval	0	0	0	0	1
# papers	6	14	17	47	18

용어들을 선정하여 이 용어들이 얼마나 자주 등장하는지 시대별로 살펴본 결과가 표 3에 제시되어져 있다. 두드러진 특징은 1980년대 이전과 1980년대의 논문에서는 선정된 통계 용어가 한 번도 초록에 등장하지 않는다는 점이다. 비로소 1990년대에 일부 논문에서 카이제곱 검정(Chi-square test)을 사용하였음을 확인할 수 있다. 2000년대에 들어서는 생존분석이 상대적으로 많이 시행되었으며, 기타 카이제곱 검정, 피셔의 정확 검정(Fisher's exact test), T-검정 등이 사용되었다. 2010년대에는 오히려 생존분석 보다 모형 선택, 회귀분석, 구조방정식, 분산분석(ANOVA) 모형이 많이 사용되었다.

시대가 지날수록 다양한 통계학적 분석 방법이 사용되고 있으며, 이러한 현상은 앞으로 더 두드러지리라 여겨진다. 데이터 양이 방대해지고 분석내용이 복잡해짐에 따라 이를 정확히 분석하고 해석할 수 있는 다양한 통계학적 분석 방법이 요구되어지며, 연구의 질의 향상을 위해 그 중요성이 강조되고 있다.

IV. Discussion & Conclusion

2010년대 급증한 국내 갑상선암의 발생률의 원인에 대해 [3]의 기존 연구들은 초음파 검사 장비를 비롯한 컴퓨터단층촬영(CT), 자기공명영상촬영(MRI) 등 정밀한 장비들의 도입으로 미세한 결절에서도 암 세포를 발견할 수 있게 되었고, 개인 건강검진이 활성화되면서 치명적이지 않은 갑상선 이상을 발견하는데 영향을 주었다고 분석하고 있다. 2016년 11월, 대한갑상선학회에서는 갑상선암 진료 권고안 개정안을 발표하였는데, 주요 골자는 과잉치료를 피하기 위해 적극적으로 검사나 치료를 하지 말라는 것이다. 침범소견이나 전이소견이 있지 않는 한, 재발위험성에 대한 적극적 대비보다는 삶의 질을 고려한 치료를 권하고 있다. 또한 수술범위를 줄이고 수술 후 방사성요오드치료도 적극적으로 권하지 않고 있다[21]. 이러한 갑상선암에 대한 임상적 시각 및 치료의 변화는 연구 주제에 직접적인 영향을 미친 것으로 나타났다. 갑상선암 관련 국내 논문을 대상으로 텍스트 마이닝을 적용한 결과, 증례·사례 보고 위주였던 내용은 1990년대에 들어서 건강 검진을 통한 갑상선암의 조기 진단이라는 주제가 자주 검토됨이 확인되었다. 2000년대에는 많은 연구들이 다양한 주제로 진행되었는데, 특히 주목할 만한 점은 여러 장비들을 활용한 검사방법과 미세한 암의 발견에 대한 논의가 증가하였다. 반면에 2010년대에는 수술과 치료 후 환자의 복지에 관련된 주제가 많이 다루어졌다.

의생물학에서 텍스트 마이닝 기법을 적용한 연구는 대부분 진료데이터를 이용해 질병 정보를 파악하거나 패스웨이(pathway)를 구축하는 데 초점을 맞추고 있다. 본 연구에서는 관련 문헌 정보를 이용해 주요 연구 분야와 시계열 트렌드를 밝히고 연구 분야 간의 상관관계를 시각적으로 가시화하였다. 이러한 시도는 암 특히 갑상선암 관련 분야에서는 국내 처음이며, 향후 다른 암, 질병에 확장하여 분석함으로써 연구 활성화 방향의 기초자료로 활용될 수 있을 것이다.

REFERENCES

[1] K. Jung, Y. Won, H. Kong, and E. Lee, "Cancer statistics in Korea: incidence, mortality, survival, and prevalence in 2015," Cancer Research and Treatment: Official Journal of Korean Cancer Association, Vol. 50, No. 2, pp. 303-316, Mar. 2018.

[2] International Agency for Research on Cancer, "GLOBOCAN 2012: Estimated Cancer Incidence, Mortality and Prevalence Worldwide in 2012 v1.0", http://globocan.iarc.fr

[3] H. Ahn, H. Kim, and H. Welch, "Korea's thyroid cancer epidemic-screening & overdiagnosis," The New England Journal of Medicine, Vol. 371, No. 19, pp. 1765-1767, Sep. 2014.

[4] L. Davies, "Overdiagnosis of thyroid cancer," BMJ: British Medical Journal (Online), Vol. 355, Nov. 2016.

[5] S. Jegerlehner, J. L. Bulliard, D. Aujesky, N. Rodondi, S. Germann, I. Konzelmann, A. C. Chiolero, and NICER Working Group, "Over- diagnosis and overtreatment of thyroid cancer: a population-based temporal trend study," Vol. 12, No. 6, Jun. 2017.

[6] National Health Insurance Service, "Main Surgery Statistical Yearbook for 2016," National Health Insurance Service, 2017. http://www.nhis.or.kr/bbs7/boards/B0079/22737

[7] M. A. Hearst, "Untangling text data mining," Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, pp. 3-10, Jun. 1999.

[8] H. Kim, D. Kim, and J. Jo, "Patent data analysis using clique analysis in a keyword network," Journal of the Korean Data and Information Science Society, Vol. 27, No. 5, pp. 1273-1284, Sep. 2016.

[9] Y. Hyun, J. Kim, J. Jeong, S. Yun, and M. Lee, "Text mining on internet-news regarding climate change and food," Journal of the Korean Data and Information Science Society, Vol. 26, No. 2, pp. 419-427, Mar. 2015.

[10] W. S. Cho, A. Cho, K. Kwon, K. and H. Yoo, "Implementation of smart Chungbuk tourism based on SNS data analysis," Journal of the Korean Data and Information Science Society, Vol. 26, No. 2, pp. 409-418, Mar. 2015.

[11] B. Kang, M. Huh, and S. Choi, "Performance analysis of volleyball games using the social network and text mining techniques," Journal of the Korean Data and Information Science Society, Vol. 26, No. 3, pp. 619-630, May 2015.

[12] J. Lee and M. Lee, "Big data-based information recommendation system," Journal of the Korea Institute of Information and Communication Engineering, Vol. 22, No. 3, pp. 443~450, Mar. 2018.

[13] H. Park, M. Lee, S. Hwang, and S. Oh, "TF-IDF based association rule analysis system for medical data," KIPS Transactions on Software and Data Engineering, Vol. 5, No. 3, pp. 145-154, Mar. 2016.

[14] J. Kim, H. Kim, Y. Yeo, M. Shin, and S. Park, "Inferring disease-related genes using title and body in biomedical text," KIISE Transactions on Computing Practices, Vol. 23, No. 1, pp. 28-36, Jan. 2017.

[15] S. Choi, S. Yoo, and H. Cho, "A study on the semiautomatic construction of domain-specific relation extraction datasets from biomedical abstracts - mainly focusing on a genic interaction dataset in Alzheimer's disease domain," Journal of Korean Library and Information Science Society, Vol. 47, No. 4, pp. 289-307, Dec. 2016.

[16] G. Jang, Y. Hwang, M. Oh, T. Lee, and Y. Yoon, "Novel Drug Similarity Measuring Method based on Text Mining for Predicting Similar Drugs," The Journal of Korean Institute of Information Technology, Vol. 14, No. 7, pp. 127-137, Jul. 2016.

[17] H. Ahn, M. Song, and G. E. Heo, "Inferring undiscovered public knowledge by using text mining analysis and main path analysis: the case of the gene-protein brings about chains of pancreatic cancer," Journal of the Korean BIBLIA Society for library and Information Science, Vol. 26, No. 1, pp. 217-231, Jan. 2015.

[18] M. J. Lee and J. W. Kim, "Design and Implementation of the Menu Navigation using Social Network Analysis among the Menus of Management Information System," Journal of the Korea Society of Computer and Information, Vol. 19, No. 9, pp. 151-160, Sep. 2014.

[19] S. J. Oh and M. K. Won, "Using Text Mining Techniques for Intrusion Detection Problem in Computer Network," Journal of the Korea Society of Computer and Information, Vol. 10, No. 5, pp. 27-32, Nov. 2005.

[20] S. J. Oh and C. W. Park, "Development of Automatic Rule Extraction Method in Data Mining : An Approach based on Hierarchical Clustering Algorithm and Rough Set Theory," Journal of the Korea Society of Computer and Information, Vol. 14, No. 6, pp. 135-142, Jun. 2009.

[21] Korean Thyroid Association. "Revised Korean thyroid association management guidelines for patients with thyroid nodules and thyroid cancer," Journal of the Korean Society of Radiology, Vol. 64, No. 4, pp. 389-416, Dec. 2010.

Authors

Tae-Gyeong Lee is an undergraduate student of Applied Mathematics from Kumoh National Institute of Technology, Korea. She is currently an undergraduate researcher in Applied Statistics Laboratory, Kumoh National Institute of Technology. She is interested in machine learning and text mining.

Seong-Min Heo is an undergraduate student of Applied Mathematics from Kumoh National Institute of Technology, Korea. He is currently an undergraduate researcher in Applied Statistics Laboratory, Kumoh National Institute of Technology. He is interested in big data analytics and text mining.

Seung-Hyeok Shin received the B.S. degrees in Applied Mathematics and received the M.S. and Ph.D degrees in Department of Computer Engineering from Kumoh National Institute of Technology, Korea, in 1998, 2000 and 2013, respectively. He joined the faculty in 2016 and is currently an Assistant Professor of the Department of Applied Mathematics at Kumoh National Institute of Technology, Gumi, Korea. He is interested in Mathematical Algorithms, Micro Systems and Data Visualization.

Ji-Yeon Yang received MS and PhD degrees in Statistics from University of Illinois at Urbana-Champaign, in 2006 and 2010, respectively. She joined the faculty in 2014 and is currently an Assistant Professor of the Department of Applied Mathematics at Kumoh National Institute of Technology, Gumi, Korea. She is interested in big data analytics, Bayesian analysis and biomarker development.