대한한의학원전학회지
2018년 11월 31권 4호

pISSN 1229-8328
J Korean Med Classics 2018:31(4) 027-046
https://doi.org/10.14369/jkmc.2018.31.4.027

JKMC
Journal of Korean Medical Classics

텍스트마이닝(Text mining)을 활용한 한의학 원전 연구의 가능성 모색

-『黃帝內經』에 대한 적용례를 중심으로 -

[1]가천대학교 한의과대학 생리학교실 대학원생· [2]가천대학교 한의과대학 생리학교실 교수
[3]한의학고전연구소 박사후연구원 · [4]가천대학교 한의과대학 원전의사학교실 교수
배효진[1] · 김창업[2] · 이충열[2*] · 신상원[3] · 김종현[4]**

Investigation of the Possibility of Research on Medical Classics Applying Text Mining

- Focusing on the Huangdi's Internal Classic -

Bae Hyo-jin[1] · Kim Chang-eop[2] · Lee Choong-yeol[2*], Shin Sang-won[3] · Kim Jong-hyun[4]**

[1]Graduate Student at Department of Physiology, College of Korean Medicine, Gachon University
[2]Professor at Department of Physiology, College of Korean Medicine, Gachon University
[3]Post-doctoral Researcher at Institute of Oriental Medical Classics
[4]Professor at Dept. of Medical Classics and History, College of Korean Medicine, Gachon University.

Objectives : In this paper, we investigated the applicability of text mining to Korean Medical Classics and suggest that researchers of Medical Classics utilize this methodology.

Methods : We applied text mining to the Huangdi's internal classic, a seminal text of Korean Medicine, and visualized networks which represent connectivity of terms and documents based on vector similarity. Then we compared this outcome to the prior knowledge generated through conventional qualitative analysis and examined whether our methodology could accurately reflect the keyword of documents, clusters of terms, and relationships between documents.

Results : In the term network, we confirmed that Qi played a key role in the term network and that the theory development based on relativity between Yin and Yang was reflected. In the document network, Suwen and Lingshu are quite distinct from each other due to their differences in description form and topic. Also, Suwen showed high similarity between adjacent chapters.

Conclusions : This study revealed that text mining method could yield a significant discovery which corresponds to prior knowledge about Huangdi's internal classic. Text mining can be used in a variety of research fields covering medical classics, literatures, and medical records. In addition, visualization tools can also be utilized for educational purposes.

Keywords : Text mining, Medical classics(原典), Huangdi's Internal Classic(黃帝內經), Network analysis

* Corresponding Author : Lee Choong-yeol .
Department of Physiology, College of Korean Medicine, Gachon University
Tel: +82-31-750-5419 E-mail : cylee@gachon.ac.kr
** Corresponding Author : Kim Jong-hyun.
Dept. of Medical Classics and History, College of Korean Medicine, Gachon University.
Tel: +82-31-750-5422 E-mail : ultracoke@gachon.ac.kr
Received(October 31, 2018), Revised(November 19, 2018), Accepted(November 19, 2018)

I. 서론

텍스트 마이닝(text mining)은 텍스트 데이터의 분석과 처리를 위한 기술을 포괄하는 용어로, 정형화되지 않은 문헌 자료를 수리적 분석이 가능한 구조화된 데이터로 변환해 유의미한 정보를 추출하는 기법이다. 주로는 기업들의 마케팅 전략 수립에 활용되어왔으나[1] 최근 들어서는 문헌 자료를 대상으로 한 학술분야의 활용도 역시 높아지는 추세이다.[2] 서양의학에서는 의료기록에서 증상, 질환, 발병 요인의 연결 패턴을 추출하고 기계학습(machine learning)을 접목함으로써 임상 의사결정 지원 시스템(clinical decision support system)을 개발하는 등의 연구가 활발히 진행되고 있다.[3]

텍스트 마이닝의 장점은 자동화된 분석기술을 이용하여 기존에 사람이 다룰 수 없었던 대량의 문헌을 분석할 수 있다는 것으로, 한정된 수의 연구자들이 방대한 양의 문헌을 다루어야 하는 원전학 분야의 어려움을 고려하면 활용 가치가 높을 것으로 기대할 수 있다. 하지만 주로 정성적 방법론을 통해 연구를 수행해왔던 원전 연구자들에게 텍스트 마이닝과 같은 데이터 처리 기법은 아직 생소한 것이어서 어떻게 활용할 수 있을지 충분하게 검토되지 못한 상태다. 따라서 본 연구는 텍스트 마이닝이 한의학 원전 연구에 적용될 수 있는지 가능성을 검토하고, 앞으로의 활용방안을 모색하고자 하였다. 연구 결과는 향후 원전 연구자들이 선택할 수 있는 연구 방법론의 다양성 제고에 기여할 수 있을 것이라 기대된다.

지금까지 한의학 원전 분야에서 데이터 처리 기법을 적용한 연구는 대체로 시험 단계에 있다고 볼 수 있으며, 한의학 용어와 같이 비교적 접근이 용이한 영역을 중심으로 이루어져 왔다. 대한한의학원전학회지를 중심으로 살펴볼 때 관련 연구들은 주로 이병욱에 의하여 선도되어 왔는데, 여러 의서에 나타나는 용어[4][5][6][7], 본초[8], 방제[9]에 대한 연구나 이러한 정보의 추출 및 검색을 위한 방법론 연구가 주를 이루고 있다.[10]

최근 들어서는 한의학 원전을 대상으로 텍스트 마이닝을 적용한 연구들도 보고되기 시작하였다. 오준호[11]는 "용어 사용 빈도가 해당 텍스트의 내용을 반영한다."는 가설을 검증하기 위해 『東醫寶鑑』, 『醫學入門』, 『景岳全書』에서 특정 용어의 출현 빈도를 계측하여 문헌 간 유사성을 비교하는 한편, 유사성의 편차가 발생한 원인을 정성적으로 분석하여 계량적 분석 결과가 기존의 정성적 연구 결과와 부합

1) 텍스트 마이닝의 하위 범주인 오피니언 마이닝(opinion mining)은 대상에 대한 주관적 의견을 문서에서 추출하는 분야로, 감성분석으로 더 잘 알려져 있다. 기업들은 소비자의 요구를 파악하고 마케팅 전략을 수립하기 위해 오피니언 마이닝 기법을 소비자층의 SNS 분석, 제품 리뷰 분석 등에 적극적으로 활용하고 있다.

2) 예를 들어, 생물학 문헌들을 대상으로 텍스트 마이닝을 적용해 질병과 관련된 유전자를 추론하는 방법론을 제안하거나(김정우 外 4人. 생물학 문헌 데이터의 제목과 본문을 이용한 질병 관련 유전자 추론 방법. 정보과학회 컴퓨팅의 실제 논문지. 2017. 23(1). pp.28-36.), 논문들로부터 단백질간의 상호작용을 추출하는 텍스트 마이닝 기법 등이 연구되었다.(이현철 외 4人. 단백질 상호작용 추출을 위한 텍스트 마이닝 기법. 컴퓨터정보통신연구. 2004. 12(1). pp.61-66.)

3) 의료기록 분석을 통해 질병과 증상의 관계를 추론하는 시스템 개발은 의료정보학의 대표적인 텍스트 마이닝 적용 분야이다. 의료기록에 나타난 증상들을 의학 분야 온톨로지(ontology)를 활용해 유사도를 평가한 후 예측 가능한 질병 후보들을 제시함으로써 의사결정의 참고자료로 활용할 수 있다.(박호식 外 3人. 데이터 공학: 의료 정보 추출을 위한 TF-IDF 기반의 연관규칙 분석 시스템. 정보처리학회논문지. 소프트웨어 및 데이터 공학. 2016. 5(3). pp.145-154.)

4) 박찬영, 이병욱, 김기욱. 『鍼灸甲乙經』의 用語體系에 관한 연구. 대한한의학원전학회지. 2013. 26(3).

5) 송인우, 이병욱. 『동의보감』에 기재된 인체 용어 관계를 이용한 검색효율성 향상 방법. 대한한의학원전학회지. 2012. 25(4).

6) 김명희, 이병욱, 김은하. 비위론에 기재된 술어의 분류에 관한 연구. 대한한의학원전학회지. 2010. 23(1).

7) 김민건, 이병욱, 김은하. 小兒藥證直訣과 脾胃論에 기재된 용어 비교에 관한 연구. 대한한의학원전학회지. 2010. 23(1).

8) 백진웅, 이병욱. 『方藥合編』 收錄 處方 內의 藥物 조합 頻度 연구. 대한한의학원전학회지. 2011.24(4).

9) 오월환 외 3人. 『太平惠民和劑局方』과 『素問宣明論方』과 『蘭室秘藏』의 방제구성 비교. 대한한의학원전학회지. 2014. 27(4).

10) 김기욱, 김태열, 이병욱. 본초 목록을 이용한 방제의 본초 구성 자동 추출 방법. 대한한의학원전학회지. 2014. 27(3).

11) 오준호. 의학 사상의 유사성은 계량 분석 될 수 있는가. 대한한의학원전학회지. 2018. 31(2).

함을 확인했다. 이태형 등[12]은 『東醫寶鑑·五臟門』의 針灸法 관련 내용을 취합하고, 내용 중에 나타난 病因과 經穴의 관계를 밝히기 위한 연구를 수행하였다. 이 연구는 해당 기법을 적용하기에 앞서 문헌의 관점을 파악하기 위한 고찰을 거치고 있어, 정량적 방법과 정성적 방법의 접목을 시도했다는 의의가 있다. 언급한 두 편의 논문은 텍스트 마이닝을 활용한 원전 연구의 가능성을 일부 확인한 연구결과라 할 수 있다. 그러나 이들 연구의 진행과정을 살펴보면, 본격적인 분석에 앞서 특정 부분을 발췌 또는 각색하는 과정을 거쳤기 때문에[13] 한의학 문헌의 원자료(raw data), 즉 가공하지 않은 원문 자료에 자동화된 분석방법을 적용해도 의미 있는 결과를 얻을 수 있을지를 온전히 검증했다고는 보기 어렵다. 연구자가 사전에 많이 개입하면 할수록 자동화된 분석을 적용하는 장점이 줄어들며, 연구 방법론으로서 가치도 낮아지기 때문이다.

텍스트 마이닝을 통한 원전 연구의 가능성을 모색하기 위해 본 연구는 『黃帝內經』 원문 전체를 예시로 텍스트 마이닝 분석을 시행하였으며, 그 과정 중에 드러난 결과들을 기초로 가능성과 한계점을 고찰하였다. 『黃帝內經』을 예시로 선택한 까닭은 『黃帝內經』이 한의학의 가장 오래된 원전이고, 이 책의 문장과 논리가 후대 문헌들에서도 꾸준히 발견되므로 분석과정에서 드러나는 장단점을 한의학 문헌 전반에 확장할 수 있는 대표성을 띠고 있기 때문이다.

본 논문은 『黃帝內經』을 예시로 텍스트 마이닝 방법을 원전 연구에 활용할 수 있을지를 모색해 보는 것 외에 원전 연구자들에게 아직 생소한 텍스트 마이닝 분석 방법에 대한 기본적인 정보를 제공하려는 목적도 있다. 따라서 논문의 전반부는 텍스트 마이닝의 방법으로 문헌 데이터를 처리하는 과정과 기본 원리를 설명하는 데에 할애하였다. 후반부에서는 앞선 과정의 결과를 전체 용어(字) 간, 문서(篇) 간 관계로 나누어 시각화하고, 기존에 알려진 지식들과 비교하는 방식으로 분석하였다. 이후에는 『黃帝內經』의 분석 과정에서 나타난 결과들을 토대로 해당 방법론의 활용 가능성과 한계점 및 보완점을 고찰하는 방향으로 논의를 진행하였다.

12) 이태형 외. 텍스트마이닝을 이용한 동의보감 질병인식방식과 내경편 경혈 분석. 대한경락경혈학회지. 2013. 30(4).

13) 오준호의 연구는 본 연구와 유사한 목적으로 수행되었으나 五臟, 六腑, 八綱 등 직접 추출한 핵심어만을 분석 대상으로 한정하였다. 이태형 등의 연구 또한 본격적인 분석에 앞서 『東醫寶鑑』에서 주치 병증에 관한 서술만을 취합하여 분석하였다.

II. 본론

1. 연구 방법

텍스트 마이닝은 비정형의 문자 자료를 컴퓨터가 인식할 수 있는 데이터 형식으로 구조화하고 데이터 마이닝 분석기법을 적용함으로써, 방대한 텍스트로부터 정보를 요약하거나 이면에 숨겨진 패턴을 드러나게 한다.[14] 본격적인 분석에 앞서 연구자는 연구 대상으로 삼은 문헌에서 불필요한 정보를 제거하고, 이어서 의미단위를 기준으로 텍스트를 분할하는 과정을 거친다. 분할된 텍스트는 연구자가 확인하고자 하는 속성이 반영된 값(數値)으로 변환되며 자동 분석에 용이한 형태를 갖추게 된다. 이후 다양한 정량적 분석기법을 이용하여 텍스트를 분석하고, 분석결과를 효율적으로 요약, 제시하기 위해 데이터 시각화 기법들을 동원한다. 마지막으로 연구자는 추출된 결과를 해석하고 결론을 도출한다.

본 연구의 경우 『黃帝內經』에 쓰인 용어들 중 불필요한 부호들을 제거하는 데이터 전처리과정(1.1 Data Preprocessing)을 거친 후, 162篇과 글자 단위로 원문을 분할(1.2.1 텍스트분할)하고, 분석 대상의 전체 집합인 사전(1.2.2 Dictionary)을 구성하였다. 이어서 篇마다 글자들이 등장하는 빈도 패턴을 수치화한 다음 글자와 글자, 편과 편 사이의 유사도를 계산하였다(1.3 Word Vectorization). 이를 바탕으로 용어 네트워크와 문서 네트워크를 구성하고 시각화(1.4 Visualizing)하였다.

14) Vishal Gupta and G.S. Lehal. A Survey of Text Mining Techniques and Applications. Journal of Emerging Technologies in Web Intelligence. 2009. 1(1). pp.60-76.

본 연구의 데이터분석은 범용 프로그래밍 언어인 Python(ver3.6.5)을 이용해 수행되었다.[15]

1.1 데이터 전처리 (Data Preprocessing) - 연구 목적에 맞는 데이터 범위 선정

본격적인 분석에 앞서 데이터 전처리 과정이 필요한 까닭은 연구대상인 문헌 자료의 기록 양식이 일정치 않기 때문이며, 때로는 연구 목적에 따라 데이터를 한정할 필요가 있기 때문이다.

이 단계를 세밀히 할수록 연구 목적에 부합하는 결과물을 얻기에 용이하지만 그만큼의 시간과 노력이 소요되므로 대규모 데이터에 적용하기 어렵다는 단점이 있다. 따라서 연구자는 어떤 수준에서 데이터를 추출할 것인지 결정해야 하며, 그에 따라 발생할 결과의 변화를 충분히 인지하여 해석 시 오류를 범하지 않도록 주의해야 한다. 또한 연구 목적에 적합한 데이터 선정을 위해서는 해당 분야 혹은 문헌에 대한 전반적인 이해가 요구된다.

본 논문의 데이터 전처리 과정은 비교적 단순하게 진행하였다. 원문 텍스트 파일에는 『素問』, 『靈樞』의 모든 편이 포함되었으며, 순수 텍스트만 남기기 위해 漢字 외에 인용부호, 쉼표, 마침표, 특수문자, 공백문자 등을 제거하였다. 그 결과 총 162篇의 156982字를 분석 대상으로 설정하였다.

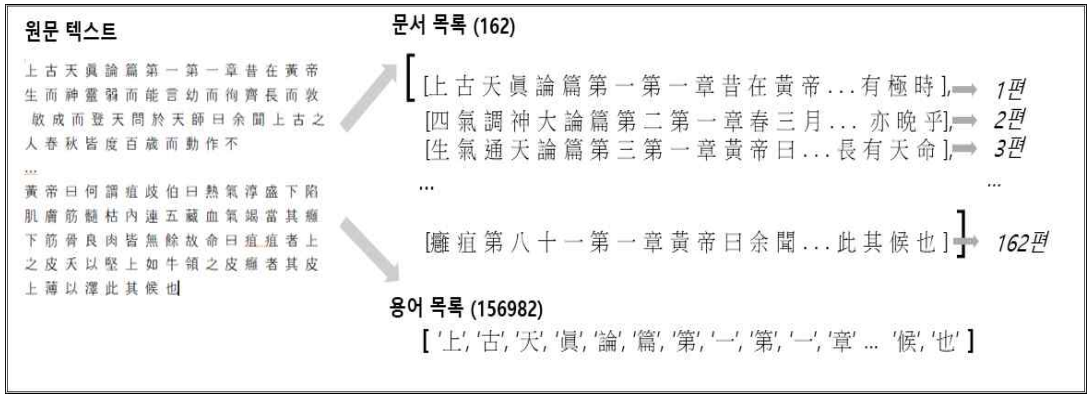
1.2. 데이터 정형화 - 데이터 구조 정의하기

1.2.1 텍스트 분할 - '문서'와 '용어'의 단위 지정 및 원문 자르기

텍스트 마이닝은 비정형의 문헌자료에서 유의미한 정보를 발견하는 과정이다. 따라서 '비정형의 데이터를 어떻게 정형화할 것인지'를 정의하는 작업이 선행되어야 한다. 전체 텍스트에서 문서(document)의 단위와 용어(term)의 단위를 규정해야 하는데, 문서의 단위는 장, 편, 문단, 문장 등 목적에 따라 지정하고, 용어의 단위 역시 단어, 음절, 형태소 등 분석하고자 하는 의미단위에 맞게 지정한다.

본 연구에서 최소 분석 단위의 문서는 篇으로, 최소 분석 단위의 용어는 글자(一字)로 설정했으며, 줄글 형태의 원문을 편 단위로 분할한 '문서 목록'과, 글자 단위로 분할한 '용어 목록'을 각각 작성하였다.(Fig. 1)

Fig. 1. Construction of document list and term list through text division



15) 데이터처리 및 분석 전반 과정을 위해 numpy, scipy, pandas 라이브러리를, TF-IDF, 코사인유사도 계산을 위해 sklearn 라이브러리를 이용하였으며 데이터 시각화를 위해 seaborn, 네트워크 분석을 위해 networkx 라이브러리를 사용하였다

1.2.2 사전(Dictionary)의 구성 – 분석에 사용할 용어집 작성

사전은 분석에 사용할 용어들의 집합이다. 컴퓨터는 문서를 읽어내려 가면서 사전에 포함된 용어가 등장할 경우, 해당 용어의 빈도 값에 1씩 더해가는 식으로 누계한다.

관찰 대상을 포괄하는 사전이 이미 구축되어 있는 경우라면 컴퓨터로 하여금 기존의 사전과 대비하면서 용어를 인식하도록 할 수 있지만, 그렇지 못한 경우에는 문헌에 사용된 용어들을 직접 추출해서 사전을 구성해야 한다.[16] 사전을 직접 구성하고자 할 때 중요한 점은 '용어'의 단위를 지정하는 방법이다. 보통은 단어(word)를 의미 단위로 파악하지만, 공백문자(space)를 기준으로 단어를 쉽게 추출할 수 있는 영어와 달리 띄어쓰기를 하지 않는 한문 텍스트의 경우 자동으로 단어를 인식하여 분절하기 어려운 점이 있다.

본 연구의 경우 『黃帝內經』에 사용된 한의학적 용어를 포괄하는 사전이 구축되어 있지 않은 상황에서 진행되었다. 또한 연구자가 일일이 단어를 추출해 사전을 구성하는 것은 대량의 문헌에 대한 텍스트 마이닝의 활용 가능성을 검토하려는 본 연구의 취지에 부합하지 않는다. 그러므로 단어 단위의 사전을 구성하지 않았으며, 표의문자라는 漢字의 특성을 반영해 글자(一字) 단위의 사전을 구성했다.

본 연구에서는 총 3가지 유형의 사전을 구성했는데, 이는 분석대상을 한정하는 것이 결과에 어떠한 영향을 미치는지 관찰하기 위해서이다. 첫 번째 사전은 앞서 글자 단위로 분절한 용어 목록 중 중복을 제외한 2422글자 전체를 사전으로 사용한 것이다.(이하 사전1로 지칭한다). 두 번째 사전은 문법적으로 사용된 글자 등 현재의 분석 초점에서 벗어나있다고 판단되는 글자들을 수동으로 제외하고 남은 2344개의 글자로 구성하였다.(이하 사전2로 지칭한다.) 마지막으로 소수의 편에만 지엽적으로 등장한 글자를 분석 대상에서 제외하기 위해 최소 3편 이상에 등장한 1461개 글자들로만 사전을 구성하였다.(이하 사전3으로 지칭한다.)

16) 그러나 이처럼 연구자가 단어를 원문에서 하나씩 추출해서 사전을 구성하는 방식은 분석 데이터 규모가 커짐에 따라 현실적으로 적용하기 어려워지며 연구자의 주관에 의한 오류의 발생도 배제할 수 없다는 단점이 있다.

1.3. 용어의 벡터화(Word Vectorization) – 텍스트를 컴퓨터가 이해할 수 있는 데이터로 바꿔주기

용어의 벡터화(word vectorization)는 컴퓨터에 용어를 입력할 때 해당 용어가 가지고 있는 속성을 컴퓨터가 인식하여 처리할 수 있는 형태로 변환하기 위한 방법이다.[17][18] 컴퓨터로 정보를 처리하기 위해서는 그 정보를 컴퓨터가 다룰 수 있는 코드로 변환하여야 하며, 다양한 경우를 표현하기 위해서 개별 정보를 숫자들의 배열로 구성된 '벡터'(vector)의 형태로 코딩하는 것이 효율적이다. 간단한 예시로 '강아지'=[1,0,0], '고양이'=[0,1,0], '돌'=[0,0,1]과 같이 고유의 벡터 코드를 지정하는 것을 생각할 수 있다. 그러나 이러한 방식으로 용어를 규정할 때 컴퓨터는 벡터를 각 용어에 대응시킬 뿐 해당 용어의 속성은 이해할 수 없다. 따라서 용어 벡터화는 용어의 속성을 표현할 수 있는 벡터(word representation vector)를 만드는 것을 목표로 한다.

1.3.1 빈도기반 용어벡터화 (Frequency based Word Vectorization) – 용어의 속성을 빈도로 표현하기

본 논문에서는 용어의 문서 내 분포(distribution)를 통해 용어 벡터의 속성을 규정하는 방법을 사용하였다. 이 방법은 "출현 빈도가 유사한 패턴을 나타내는 용어들은 유사한 속성을 가질 것이다."라는 가정에서 출발한다.[19] 예를 들어, 『黃帝內經』에서

17) 단어의 문서 내 분포에 기반 하여 벡터공간상에 단어벡터를 표현하고 그 유사성을 정량화하는 방식은 언어학에서 논의되어지던 전통적인 워드임베딩(word embedding) 개념에 포함되지만, 최근 컴퓨터언어학 등 에서 일반적으로 논의되어지는 워드임베딩은 다양한 차원축소기법을 적용한 용어의 벡터화 혹은 수치화 방식을 포괄하는 개념이다.

18) Salton G, Wong A, Yang CS. A vector space model for automatic indexing. Commun ACM. 1975. 18(11). pp.613-620.

19) Zellig S. Harris. Distributional Structure. WORD.

'淸', '濁', '官', '於'라는 4가지 글자의 등장 패턴을 篇 단위로 분석하면 '淸'과 '濁'은 다양한 편에 등장하되 주로 함께 등장할 것이므로 그 분포가 유사할 것이다. 이에 비해 '官'은 「靈蘭秘典論」에서 장부의 직능을 표현할 때 여러 번 언급되지만 나머지 편에 거의 등장하지 않으므로 '淸', '濁'과는 다른 양상을 보일 것이라 예상할 수 있다. '於'의 경우 어조사로써 다빈도의 불규칙한 패턴을 보일 것이다.

1.3.2 TF-IDF 계산 : 특정 문서에 대한 특정 용어의 중요도를 출현 빈도로 정량화하기

빈도(frequency) 기반으로 용어를 벡터로 표현하기 위해선 용어마다 각각의 문서에서의 출현빈도를 계산하고 이 값들을 벡터로 표현해야 한다. 이 과정에서 '각 문서에 대한 특이적 관련성'을 고려하기 위해서는, 절대적인 출현빈도가 아닌 특정 문서에 대한 특이적 출현빈도를 반영해야 한다. 본 연구에서는 『黃帝內經』의 각 篇을 문서로, 개별 글자를 용어로 정의하였고, 특정 용어(a)의 특정 문서(A)에 대한 출현 빈도는 TF-IDF (term frequency-inverse document frequency)를 이용하여 계산하였다.

$$tf-idf(t,d) = tf(t,d)*idf(t)$$

$$idf\,(t) = \log\,[\frac{n}{df(t)}]$$

t : 용어, d : 문서, n : 전체 문서 수
$tf(t,d)$: 용어t가 문서d에 등장한 횟수
$df(t)$: 용어t가 등장한 문서수

TF-IDF는 TF(용어빈도)와 IDF(역문서빈도)의 곱으로 정의되며, '특정 용어가 특정 문서 내에서 얼마나 중요한가'를 나타내는 수치이다.[20] TF는 문서에 등장한 용어의 단순 빈도로서 특정 문서에서 특정 용어의 등장 횟수를 계수한 값이다.

IDF는 전체 문서 수를 주어진 용어가 출현한 문서들의 수, 즉 문서빈도(DF)로 나눈값에 로그를 취한 값으로 정의한다.[21] IDF 값은 용어의 문서빈도가 높을수록 낮아지기 때문에 TF에 IDF를 곱해줌으로써 편과 글자 사이의 특이적 연관성을 반영할 수 있다. 즉 TF는 높지만 모든 편에 자주 등장하여 특이성이 낮은 용어들이 지나치게 부각되지 않도록 하는 것이다. 예를 들어 영어의 경우 'the' 와 같은 관사는 모든 용어에 높은 빈도로 등장하여 높은 TF 값을 가지지만 동시에 높은 DF값으로 인해 IDF는 아주 낮아지므로 전체 TF-IDF 값은 낮아진다. 종합해볼 때 "a라는 용어가 A라는 문서에서 TF-IDF 값이 높다."는 것은 a용어가 A문서에만 자주 등장한다는 것을 의미하므로, "용어 a는 문서 A의 키워드이다."라고 말할 수 있다.

1.3.3 문서-용어 행렬 (Document-Term Matrix)의 구성 : TF-IDF 값을 행렬 형태로 표현하기

특정 용어에 대해 각각의 문서들을 대상으로 계산된 TF-IDF 값들을 모아 벡터로 표현하면 컴퓨터가 용어의 속성을 '전체 문서에 대한 TF-IDF 값들의 분포'로 인식할 수 있게 하는 용어 벡터(term vector)가 완성된다. (예시 : 氣=[$\overset{1편}{0.1}$, $\overset{2편}{0.4}$ $\overset{3편}{0.4}$, ... , $\overset{162편}{0.2}$]). 반대로 특정 篇에 대하여 각 용어를 대상으로 계산된 TF-IDF 값들을 모아서 벡터로 표현한다면 문서의 속성을 '해당 문서 내 전체 용어들의 TF-IDF 값 분포'로 인식할 수 있게 하는 문서 벡터(document vector)를 구성할 수 있다.[22] (예시 : 1편(상고천진론) = [$\overset{之}{0.5}$, $\overset{脈}{0.2}$, $\overset{伯}{0.3}$, ... , $\overset{也}{0.1}$])

1954. 10. pp.2-3, 146-162.

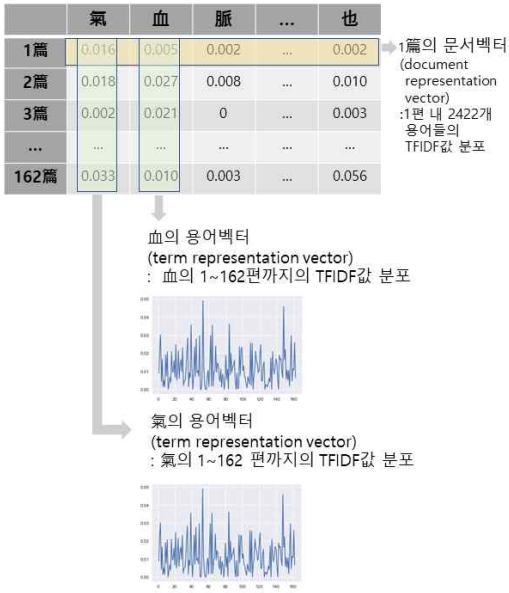
20) Wu HC, et al.. Interpreting TF-IDF term weights as making relevance decisions. ACM Trans Inf Syst. 2008. 26(3). pp.1-37.

21) 본 연구에서 TF-IDF값 계산은 python의 내장 라이브러리인 sklearn.feature_extraction.text.의 TfidfVectorizer를 사용하였으며 parameter는 default 값으로 설정하였다. (norm='l2', smooth idf = True)

22) P. D. Turney, P. Pantel. From Frequency to Meaning: Vector Space Models of Semantics. Journal of Artificial Intelligence Research. 2010. 37. pp.141-188.

용어 벡터와 더불어 문서 벡터는 이어지는 분석에서 기본적인 분석단위로 작용하게 되므로, 이상의 벡터 정보를 효율적으로 처리하기 위해 TF-IDF 값들을 행렬 형태로 구성하였다.(Fig. 2) 행렬의 행(row)에는 162개의 篇이, 열(column)에는 2422개의 용어(글자)가 배치되었으며, 행렬의 각 원소는 해당 행의 篇과 해당 열의 용어 사이에 계산된 TF-IDF 값으로 구성되었다. 이렇게 만들어진 문서-용어 행렬에서 특정 행을 취하면 해당 篇에 대한 문서벡터가 되고, 특정 열을 취하면 해당 용어에 대한 용어벡터가 된다.

Fig. 2. Document-Term Matrix
(row : document(篇), column : term(字))



1.3.4. 벡터 유사도(Vector Similarity) 계산 - 벡터의 유사도를 수치로 표현하기

계산된 용어벡터와 문서벡터는 빈도 정보를 기반으로 용어나 문서의 속성을 표현하고 있으므로 벡터 간 연산을 통해 용어 간, 문서 간 관계에 대한 정량적인 분석이 가능하다. 본 연구에서는 용어와 문서의 연결을 표현하는 네트워크를 구성하기 위해 벡터 간 유사도를 계산하였다. 벡터 유사도를 정의하는 방식은 여러 가지가 있지만 가장 보편적으로 사용되는 '코사인 유사도(cosine similarity)'를 이용하였다.[23]

$$COS(A,B) = \frac{A \bullet B}{\| A \| \ \| B \|}$$

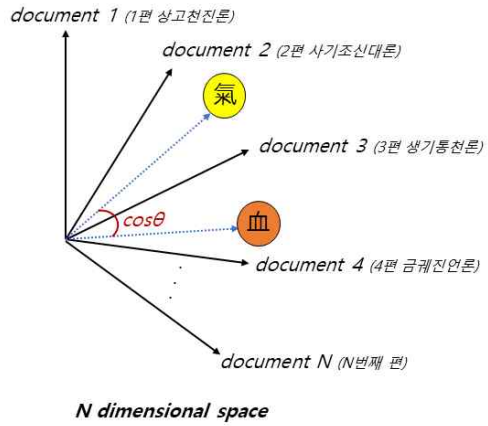
A, B = 용어벡터, 혹은 문서벡터

$A \bullet B$ = 벡터 A와 벡터 B의 내적

$\| A \|$: 벡터 A의 $norm$ 거리

n개의 원소로 구성된 두 벡터간의 코사인 유사도는 n차원 공간상의 두 벡터가 만드는 사잇각에 대한 코사인 값으로 계산된다.(Fig. 3). 이 값은 -1에서 1 사이의 값을 가지며 두 벡터가 동일한 경우 1, 두 벡터가 직교할 경우 0, 반대 방향을 가리킬 경우 -1을 갖는다.

Fig. 3. Term vectors represented in N dimensional Space



1.4 데이터 시각화 (Data visualization)

1.4.1 네트워크(Network) - 점과 선으로 이루어진 그래프

데이터 시각화는 앞서 계산한 용어 간, 문서 간 유사도 수치를 직관적으로 확인 가능한 다양한 방식으로 표현하는 단계이다. 본 연구에서는 전체 구성

23) Ye J. Cosine similarity measures for intuitionistic fuzzy sets and their applications. Mathematical and Computer Modelling. 2011. 53(1). pp.91-97.

요소들을 노드(node, 점)로, 요소 간 연결을 엣지(edge, 선)로 나타내는 형태의 네트워크로 표현하였다. 이를 통해 전체 요소 간 연결양상을 관측하고 네트워크의 전반적 특성을 분석할 수 있다.[24]

용어와 용어, 문서와 문서 사이의 유사성을 기반으로 용어네트워크(term network)와 문서네트워크(document network)를 구성하였으며, 시각화 작업은 Cytoscape 프로그램을 사용하였다.[25] 네트워크에서 노드는 용어 혹은 문서에 해당하고, 각각 노드들이 역치 값(cutoff) 이상의 유사도를 가질 경우 노드 사이를 엣지로 연결한다.

하나 이상의 노드와 역치 값 이상의 유사도를 갖지 않는 노드들은 네트워크상에 표시되지 않도록 필터링할 수 있는데, 이 경우 역치 값을 낮게 지정할수록 많은 노드가 표현되어 네트워크의 전체 구조(network topology)를 보기에는 유용하나 세부 정보를 해석하기 어려운 단점이 있다. 반대로 역치 값을 올릴수록 강하게 연결된 노드만 표시되기 때문에 주요 관계들을 명확히 부각하기는 좋으나 그만큼의 정보손실(information loss)이 발생한다. 따라서 역치 값의 조정에 따른 네트워크의 변화를 관찰하여 적절한 값을 찾는 과정이 필요하다.

1.4.2 히트맵 (Heatmap) - 네트워크상에서 연결된 용어들을 하나하나 비교해보기

히트맵은 행렬 형태의 수치들을 색깔로 표시하여 효율적으로 정보를 드러내는 시각화기법이다. 네트워크가 용어들이나 문서들 사이의 전반적인 관계를 보여준다면, 히트맵에서는 소수 요소들 간의 관계를 각각의 벡터수준에서 확인할 수 있다. 수치의 높낮이를 밝기로 표현함으로써 벡터 값들의 분포를 쉽게 파악할 수 있고, 몇 번째 값이 서로 유사하고 다른지 세밀하게 분석할 수 있다는 장점이 있다.

이는 네트워크상에서 독립된 군집 형태를 나타내는 일부 요소들 사이의 관계를 자세히 살펴보려 할 때 유용하다. 본 논문의 경우 용어 네트워크상에 독립된 군집을 형성한 '肝', '心', '脾', '肺', '腎'을 히트맵으로 작성하여 TF-IDF 값이 실제 어떤 편에서 유사했고, 어떤 편에서 차이가 난 것인지 확인해 보았다.(Fig. 5)

2. 연구 결과 및 분석

지금까지『黃帝內經』원문에 대한 텍스트 마이닝 기법의 적용 과정을 살펴보았다. 이어서는 텍스트 마이닝의 최종 결과물이라 할 수 있는 시각화 자료를 분석함으로써 기존 연구들을 통해 우리가 이미 알고 있는『黃帝內經』의 내용적 특징이 어느 정도까지 정확하게 표현되는지 확인해보고, 발견되는 장단점을 한의학 원전에 대한 텍스트 마이닝 기법의 활용 가능성을 가늠하는 근거로 삼고자 한다.

2.1. 용어 네트워크 (Term Network)

용어 네트워크는『黃帝內經』에 쓰인 용어(字)들의 관계를 시각화한 것이다. 점과 선이 연결된 그래프 형태로 문헌을 표현함으로써 네트워크의 허브(hub) 역할을 하는 용어의 종류, 개수, 그로부터 분파된 형태, 독립된 군집 등, 용어 간의 관계구조를 보다 직관적으로 파악할 수 있다.

용어 네트워크의 경우 각 용어를 노드(점)로 지정하고, 용어 벡터 간 코사인 유사도 값을 엣지 가중치(edge weight)로 부여하였다. 즉 연결선의 굵기는 해당 용어들의 유사도에 비례한다. 노드의 크기와 투명도는 해당 부분에 연결된 엣지의 개수(node degree)가 많을수록 크고 진하게 표시되도록 설정하였다.

Fig. 4-a는 앞서 설명한 사전 중 사전1을 분석한 것으로,『黃帝內經』에 등장하는 모든 용어를 사전으로 등록하여 분석한 결과이다. Fig. 4-b는 문법적으로 사용된 용어들을 제외한 사전2를 분석한 결과이다.

24) Christofides N. Graph theory: An algorithmic approach (Computer science and applied mathematics). Orlando, USA. Academic Press, Inc. 1975.

25) Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Research. 2003. 13(11). pp.2498-2504.

Fig. 4-a. Term Network
(* Dictionary_1, Edge_cutoff=0.7)

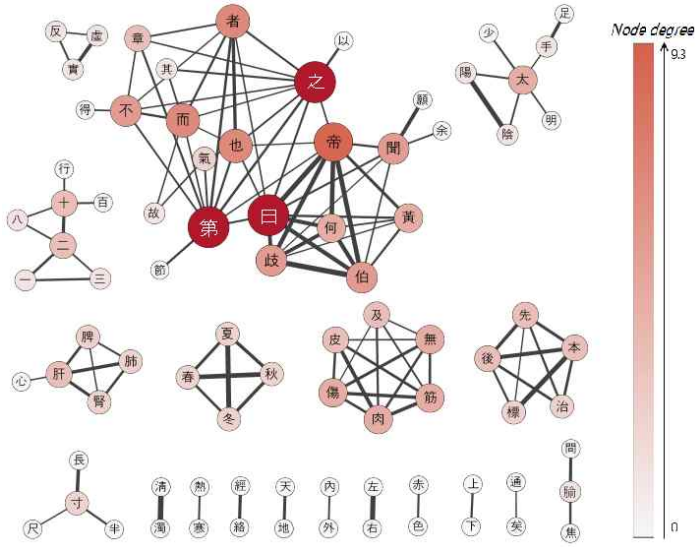
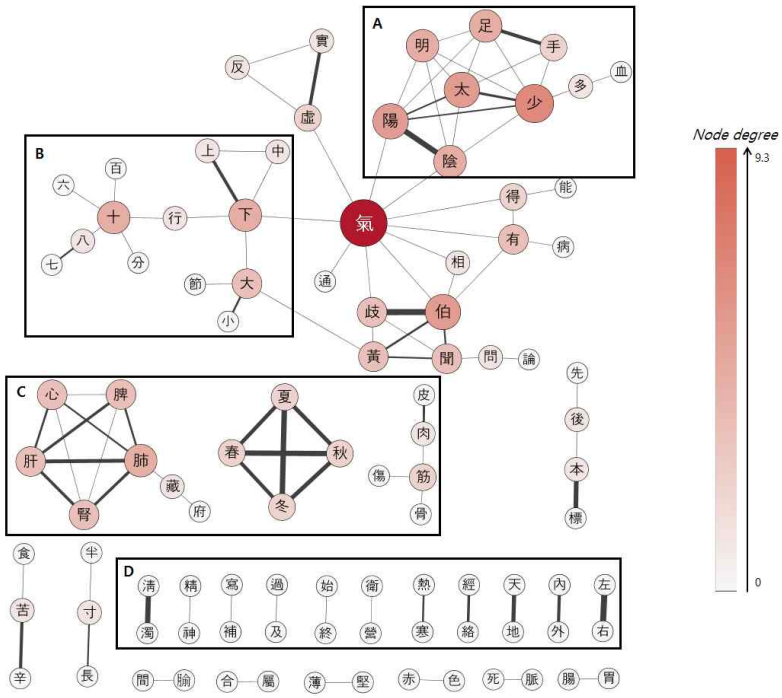


Fig. 4-b. Term Network
(*Dictionary_2, Edge_cutoff = 0.6)



2.1.1 氣

Fig. 4-a의 경우 '之', '曰', '第', '而', '也', '者', '何', '其' 등이 가장 많은 연결을 보이며, 따라서 『黃帝內經』에 사용된 중심용어라 말할 수 있다. 이를 통해 『黃帝內經』이 주로 黃帝와 岐伯의 문답형식으로 구성되었다는 점은 유추할 수 있으나, 문법적인 용도로 사용된 용어들이 대부분이어서 내용상의 상관관계를 파악하기 어렵다. 이에 비해 Fig. 4-b에서는 氣가 네트워크의 중심점 역할을 한다는 것을 직관적으로 파악 할 수 있다. 모든 노드들의 중심성(centrality)척도를 계산해 보더라도 '氣'가 연결 중심성(degree centrality)과 매개 중심성(betweenness centrality)에서 모두 높은 값을 가짐을 확인할 수 있었다.[26] 즉 '氣'가 이 네트워크의 허브 역할을 하고 있으며, 『黃帝內經』은 氣를 중심으로 의학 이론을 설명한 문헌이라 할 수 있다.

Fig. 4-a의 경우처럼 전체 용어를 사전으로 등록해 분석할 경우, 실질적 의미를 가지지 않으면서 자주 사용된 용어들이 네트워크상의 허브로 부각되어 의미구조 파악에 혼란을 줄 가능성이 있다. 따라서 텍스트 마이닝을 통해 한문으로 구성된 문헌을 분석하고자 할 때는 연구 목표에 따라 사전에 포함될 문법 용어들의 수준을 조정하는 작업이 필요할 것으로 생각된다.

2.1.2 陰陽

Fig. 4-b를 중심으로 살펴보면 의미구조의 중심에 위치한 용어는 '氣'이며, 그와 긴밀하게 연관된 용어들이 크게 4개의 군집으로 묶이는 것을 확인할 수 있다. 그중 A로 표시한 영역을 살펴보면, '氣'가 '陰', '陽'과 각각 연결되며 동시에 둘 사이의 강도가 높게 표현되는데, 이는 『黃帝內經』이 氣를 설명할 때 陰陽이라는 상대성에 기반을 두었음을 반영한다. 또한 D로 표시한 영역을 살펴보면 두 용어씩 짝을 이룬 '淸'-'濁', '精'-'神', '補'-'瀉', '終'-'始', '營'-'衛', '寒'-'熱', '經'-'絡', '天'-'地', '內'-'外', '左'-'右'는 모두 陰陽 속성이 상대되는 용어끼리 대응되었다. 이를 통해 『黃帝內經』의 서술에 사용된 용어들이 陰陽의 상대성을 바탕에 두고 있다는 점을 확인할 수 있다.

2.1.3 經絡

Fig. 4-b의 A로 표시한 영역에서 '陰', '陽'은 다시 '太', '少', '手', '足', '明'과 연결되었는데, 이는 『黃帝內經』이 太少나 多少와 같은 방식으로 陰陽의 상태나 변화를 인식했다는 것을 보여준다. 이 군집에 '手', '足', '明'이 포함된 것은 陰陽과 太少의 용어가 經絡의 명칭에 자주 사용되었기 때문이라 할 수 있다.

한편 B로 표시한 영역에서는 '氣'가 '下'와 연결되며, 下는 다시 '上', '中'과 연결된 것, '行'과 연결된 것, '大'와 연결된 것으로 나뉘었다. 이중 '上', '中', '下'는 부위나 위치를 설명할 때 사용된 것으로 보이며, '行'을 연결점으로 이어진 숫자들은 氣가 행하는 시간이나 길이를 설명한 것으로 생각된다. '小', '節'과 연결된 것은 주로 經絡의 流注部位를 설명할 때 사용된 것으로 보인다. 종합해볼 때, 下를 통해 연결된 용어들은 주로 氣의 운행 방향과 부위를 설명하는 데에 사용되었음을 알 수 있다.

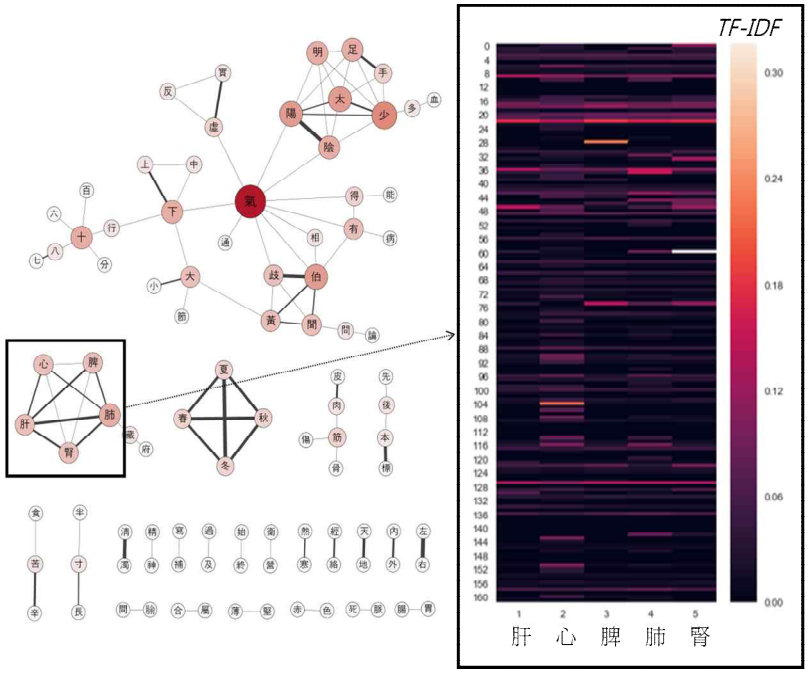
2.1.4. 五臟, 四時, 五體

Fig. 4-b의 C로 표시한 영역을 살펴보면, 五臟과 四時는 각각 소규모 군집을 이루었으며 구성 요소간의 관계도 매우 밀접하게 표현되었다. 이는 『黃帝內經』의 여러 편들이 生理, 辨證, 攝生 등을 설명할 때 五臟과 四時를 기준으로 삼았음을 반영한다고 할 수 있다.

좀 더 자세히 살펴보면, 五臟의 경우 Fig. 4-a에서는 '心'이 '肝'를 제외한 용어와 연결되지 않았고, Fig. 4-b에서도 '心'에 연결된 선의 굵기가 나머지에 비해 가늘게 나타났다. 心은 五臟 중 하나일 뿐만 아니라 마음이나 신체 부위를 지칭하는 경우가

26) 연결 중심성(degree centrality)은 노드에 연결된 선의 갯수, 즉 연결된 노드의 수로 측정된다. 매개 중심성(betweenness centrality)은 최단경로에 기반해서 측정된 중심성 척도 중 하나로 네트워크 상의 서로 다른 두 노드 s,t에 대해 s-t간의 최단경로에 v가 포함되어 있을 확률이 v의 매개중심성이 된다. v노드를 제외한 모든 노드들의 최단경로에 v가 항상 포함되면 그 값은 1이 되고 하나도 v가 포함되지 않으면 그 값은 0이 된다.

Fig. 5. Heatmap composed of five viscera vectors (*Column1~5 : liver(肝), heart(心), spleen(脾), lung(肺), kidney(腎))



있으므로, 나머지에 비해 빈도 분포의 유사도가 상대적으로 낮게 계산되었기 때문이라 추측할 수 있다. 실제 어떠한 요인이 이러한 차이를 유발했는지는 아래에 수록한 히트맵을 통해 보다 분명히 확인할 수 있다.(Fig. 5)

비슷한 예로 '皮', '肉', '筋', '骨'은 五體를 구성하는 요소로서 군집을 이루었는데, '脈'은 이들 사이에 보이지 않고 '死'와 별도의 관계로 연결되었다. 이는 '脈'이 五體 중 하나가 아닌 脈狀의 의미로 사용되어 生死를 구분하는 기준으로 쓰인 경우가 많기 때문이라 할 수 있다. 이 같은 예들은 하나의 용어가 중의적 의미로 사용되는 漢字의 특성에 기인한다고 볼 수 있으며, 그로 인해 현재 적용한 방법으로는 의미 구조를 정확하게 반영하지 못할 가능성이 있음을 확인할 수 있다.

2.2 히트맵(heatmap)

용어 네트워크에서 확인했듯이 五臟은 독립적으로 연관된 군집을 형성했으며, 그중 '心'은 나머지에 비해 상대적으로 낮은 연관성을 보였다. 이에 '肝', '心', '脾', '肺', '腎'의 TF-IDF 값을 문서(篇) 단위로 시각화하여, 유사하거나 차이가 있는 것은 어떤 편인지 직접 확인해보았다.[27]

Fig. 5의 세로축은 『素問』(0~80)부터 『靈樞』(81~161)까지의 篇을, 가로축은 1부터 5까지 각각 '肝', '心', '脾', '肺', '腎'을 의미하며, TF-IDF 값이 높을수록 밝게 표시되었다.

그 결과, 『素問』의 「靈蘭秘典論」, 「藏氣法時論」, 『靈樞』의 「本藏」, 「九鍼論」등에서는 五臟이 모두 밝게 나타났는데, 이들은 인체의 생리나 병

27) 五臟의 관계는 사전1보다 사전2를 사용한 네트워크(Fig. 4-b)에서 보다 명확히 드러나므로 히트맵 역시 사전2를 대상으로 작성했다.

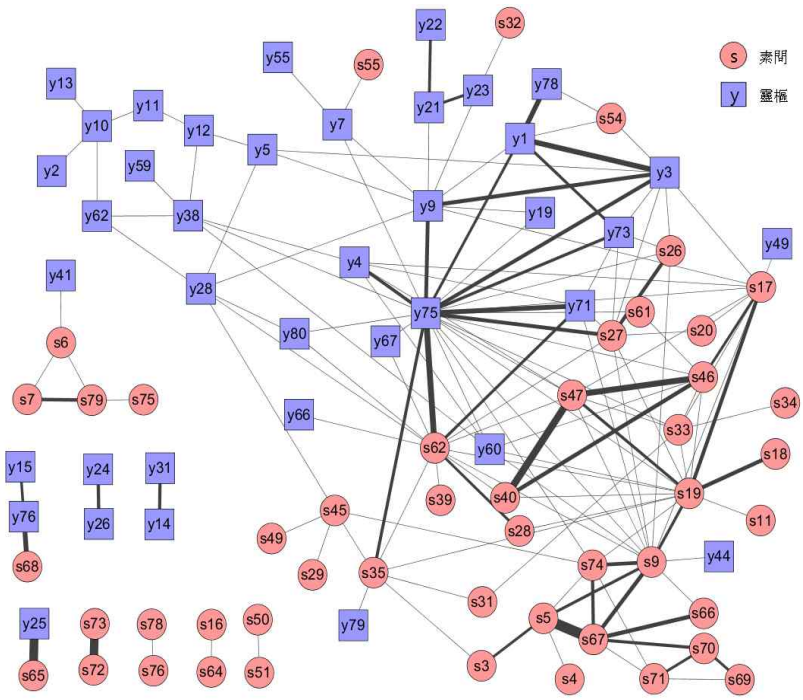
증을 오장으로 구분한 서술이 포함된 篇들이다.[28] 반대로 오장이 모두 어두운 경우도 여럿 보이는데, 오장을 기준으로 설명한 편들과 그렇지 않은 편들이 비교적 분명히 나뉘기 때문에 五臟 사이의 유사성이 높게 계산되어 네트워크상에 군집이 형성된 것이라 판단할 수 있다.

한편 『靈樞』의 「厥病(24)」, 「雜病(26)」, 「口問(28)」, 「五亂(34)」, 「五癃津液(36)」, 「五味論(63)」, 「憂恚無言(69)」, 「大惑論(79)」 등에서는 유독 '心'의 TF-IDF 값이 높게 나타났다. 이는 '心'과 나머지 臟의 유사도를 약화시키는 요인으로 작용했을 것이라 추측할 수 있다. 각 편의 내용을 살펴보면 24편과 26편은 心痛에 관한 설명에 많은 분량을 할애했으며, 34편은 邪氣의 침입과정을 心肺 위주로 설명했다. 이편들에서 '心'은 五臟의 하나이지만 四臟과는 무관하게 설명되었거나 部位를 가리키는 의미로 쓰였다. 반면 28편, 36편, 29편은 감정에 따른 인체의 변화를 언급했으며, 79편에서는 心神의 변화를 기준으로 정신질환을 서술했다. 여기서의 '心'은 감정을 뜻하거나, 감정 변화가 심장에 영향을 미치는 경우에 사용되었다. 이러한 점을 통해 네트워크상에서 '心'의 유사도가 나머지 臟에 비해 낮게 표현된 것은 '心'이 五臟 중 하나가 아닌 심리나 신체 부위를 지칭하는 용어로도 사용되기 때문이라는 가설을 확인할 수 있다.

2.3 문서 네트워크(Document Network)

앞서 용어 네트워크가 『黃帝內經』에 등장한 용어 간의 관계를 보여줬다면 문서 네트워크는 篇 사이의 관계를 시각화한 것으로 주제, 서술 방식, 서술 범

Fig. 6-a. Document Network
(* Dictionary_1, Edge_cutoff=0.55)



28) 이러한 篇들이 주로 『素問』의 전반부에 집중되어 있다는 점도 확인할 수 있었다.

위의 유사성이 네트워크에 반영되어 나타난다.

Fig. 6-a는 사전1(모든 용어), Fig. 6-b는 사전2(문법적으로 사용된 용어 제외), Fig. 6-c는 사전3(최소 세 편 이상에 등장한 용어로 한정)을 대상으로 빈도 패턴을 분석하고 유사도를 시각화한 것이다.

2.3.1 [사전1]로 구성한 문서 네트워크

Fig. 6-a는 상대적으로 높은 역치 값을 지정했음에도 Fig. 6-b와 Fig. 6-c에 비해 훨씬 복잡한 형태의 네트워크를 나타낸다. 이는 사전1에 '也', '矣', '曰' 등과 같은 문법적 용어들이 다수 포함된 것이 실제 내용과 무관하게 편들의 유사도를 상승시키는 요인으로 작용하기 때문이라 생각된다. 용어와 마찬가지로 문서들 사이의 유사도를 확인할 시에도 사전 내 문법적 용어들의 포함 여부를 조정하는 작업이 필요하다고 할 수 있다.

Fig. 6-a를 살펴볼 때 『素問』의 편들끼리 연결된 선과 『靈樞』의 편들끼리 연결된 선의 개수는 『素問』과 『靈樞』의 편이 서로 연결된 수에 비해 상대적으로 많다. 따라서 두 문헌이 서술 대상이나 방식에 있어 나름의 경향성을 가진다고 판단할 수 있으며, 계통적 차이가 있음을 추측하는 근거로 사용될 수 있다.

2.3.2 [사전2]로 구성한 문서 네트워크

Fig. 6-b는 Fig. 6-a에 비해 문서(篇)들 사이의 관계를 비교적 분명하게 보여준다. 따라서 사전1보다는 사전2가 『黃帝內經』에 수록된 편들 사이의 유사도를 분석하기에 보다 적절하다고 말할 수 있다. 설명의 편의를 위해 그림 상에 나타난 편들의 집합을 A부터 E로 구분했다.

가장 많은 연결을 보이는 편은 S5(陰陽應象大論), S62(調經論), S67(五運行大論), Y3(小鍼解), Y9(終始), Y75(刺節眞邪) 등이다. 이들은 공통적으로 다

Fig. 6-b. Document Network
(* Dictionary_2, Edge_cutoff=0.5)

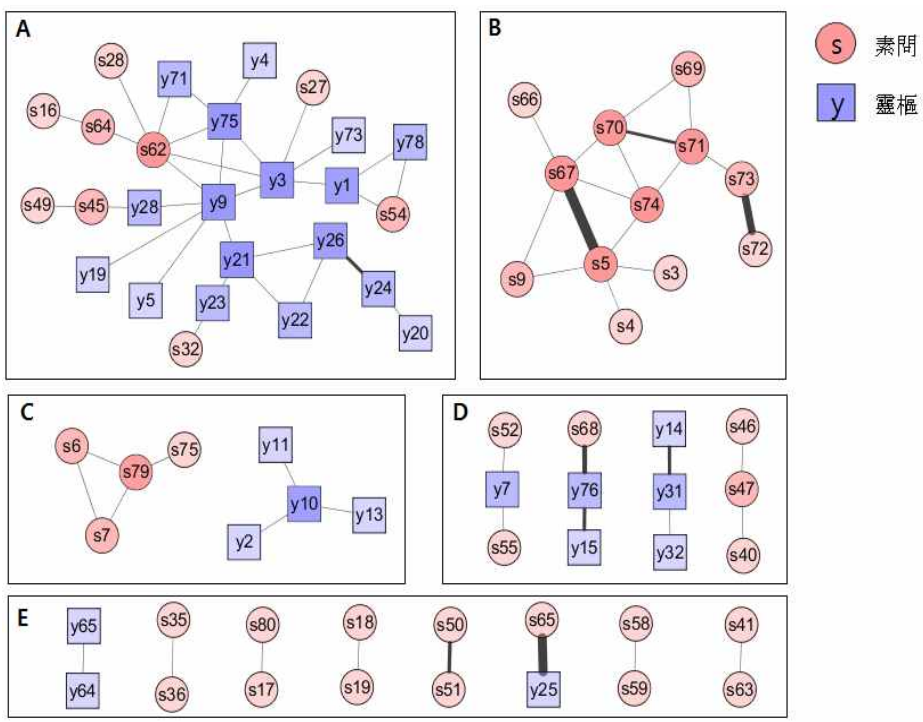
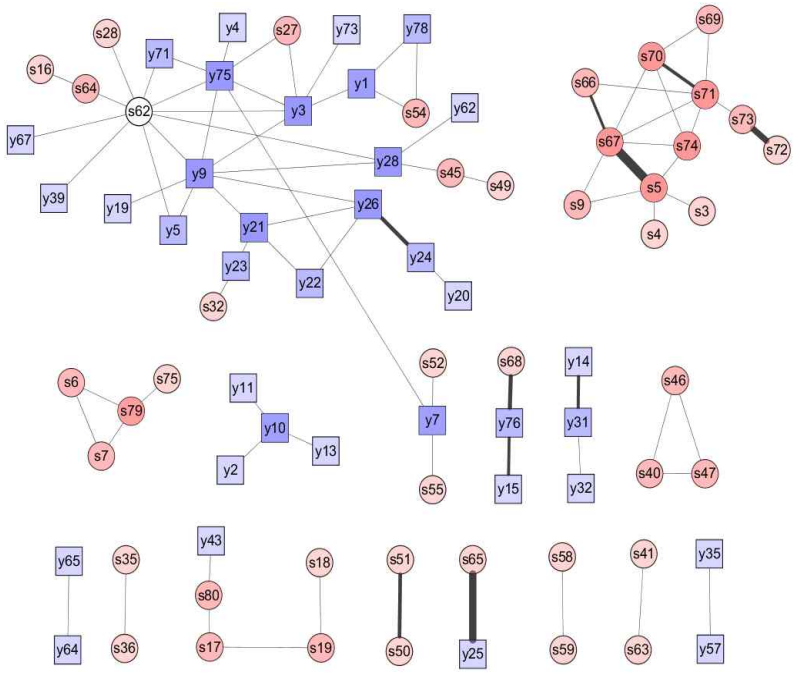


Fig. 6-c. Document Network (* Dictionary_3, Edge_cutoff=0.5)



양한 주제에 대해 논설하였으며, 대체로 길이가 길다. 반대로 연결이 적은 편들 중에는 S3(生氣通天論), S4(金匱眞言論), S9(六節藏象論), Y4(邪氣臟腑病形) 등과 같이 『黃帝內經』의 대표적인 공리를 서술했다고 여겨지는 편들이 포함되어있다. 이러한 점을 통해 문서 네트워크상에서는 노드의 연결선 수(degree)가 해당 문서의 중요도와 반드시 일치하지는 않는다는 사실을 알 수 있다.

A로 표시한 영역은 S62(調經論), Y3(小鍼解), Y9(終始), Y75(刺節眞邪)의 4편을 중심으로 네트워크가 형성되었다. 해당 편들은 刺針이라는 공통적인 키워드를 가지며, 그와 관련된 인체의 經脈, 氣, 邪氣를 함께 설명했다. A에 표현된 편들이 네트워크상 가장 큰 집합을 형성하고 있다는 것은 刺針이 『黃帝內經』의 중심 주제임을 반영한 결과라 이해할 수 있다.

B로 표시한 영역은 12편이 네트워크를 구성했는데 내용상 運氣篇을 중심으로 한 관련 편들과, 陰陽氣 및 五行歸類에 대한 내용을 서술한 편들로 구분된다. 두 군집은 陰陽五行에 대한 폭넓은 설명이 나타나는 S67(五運行大論)과 S5(陰陽應象大論)를 매개로 연결되었다.

C로 표시한 영역 중 좌측에 군집을 이룬 4편은 공통적으로 三陰三陽脈에 대해 서술했다. S79(陰陽類論)와 S7(陰陽別論)은 三陰三陽을 '一陽', '二陽', '三陽', '一陰', '二陰', '三陰'으로 표현하고 脈을 통해 병증을 기술했다는 점에서 유사하다. 그러나 S6(陰陽離合論)이 서술한 삼음삼양은 경맥의 공간 구조 및 開闔樞를 표현한 것이며, S75(著至教論)는 주로 醫道의 전수에 관한 내용이며 三陽脈이 이르는 경우를 추가로 언급했을 뿐이다.[29] 이러한 경우를

29) 같은 예로, D에 속한 Y76(衛氣行)과 Y15(五十營)는 시간 변화에 따른 營衛氣의 순환을 설명했다는 내용적 유사성이 있는 반면, Y76과 S68(六微旨大論)은 주제의 연관성은 크지 않지만 '刻'이라는 단위를 반복적으로 사용함에 따라 서로 연결된 것으로 추측된다. 또한 腸胃의 길이와 용적을 수치화하여 설명한 Y31(腸胃)과 平人이 水穀을 수납하는 용량을 설명한 Y32(平人絶穀)는 주제가 유사하지만, Y14(骨度)는 인체 부위의 길이를 수치화하여 설명하는 과정에서 Y31과 유사한 술어를 반복해 사용했음을 확

통해 본 논문이 적용한 방법론으로는 의미적 유사성과 형식적 유사성의 차이를 구분하기 어렵다는 한계를 발견할 수 있다.

E로 표시한 영역을 살펴보면 Y64-Y65, S35-S36, S18-S19, S50-S51, S58-S59처럼 순서가 가까운 편들이 연결된 경우가 많다. 이러한 경향성은 B, C, D에서도 확인되며 특히 『素問』의 경우 그러한 경향성이 높다. 이는 편집자인 王冰의 의도가 『素問』의 목차 구성에 반영된 결과라 생각된다.

2.3.3 [사전3]으로 구성한 문서 네트워크

사전2를 기준으로 분석한 네트워크(Fig. 6-b)와 사전3을 기준으로 분석한 네트워크(Fig. 6-c)의 형태는 큰 차이를 보이지 않는다. 이를 통해 문서 간 유사성을 확인하기 위한 사전 구성에서 해당 용어가 다수의 편에 등장하는지는 중요한 변수가 아님을 알 수 있다. 소수의 편에 국한되어 사용된 용어들의 경우 문서 간 유사성을 높이는 데에 별다른 영향을 미치지 못하기 때문으로 생각된다.

Ⅲ. 고찰

앞서 본론에서는 텍스트 마이닝을 통한 문헌의 분석 단계를 살펴보고, 『黃帝內經』을 대상으로 구현해보았다. 빈도 분포를 기준으로 용어(字)와 문서(篇)의 유사도를 계산하여 그 결과를 시각화하였으며, 네트워크와 히트맵으로 표현된 정보를 『黃帝內經』에 관한 기존의 지식들과 비교해볼 수 있었다.

그 결과 용어 네트워크에서는 '氣'가 가장 높은 중심성을 보임으로서, 氣가 『黃帝內經』 의학이론의 중심 용어로 활용되었다는 것을 확인할 수 있었다. 또한 陰陽, 經絡, 五臟, 四時와 같은 이론체계의 특성이 네트워크에 반영된 것도 볼 수 있었다. 문서 네트워크에서는 『素問』과 『靈樞』의 편들이 서로 구분되는 양상을 보였으며 編次가 가까운 것들끼리 연결된 경우가 많았는데, 이는 문헌의 계통성 확인과 같은 연구 주제에 텍스트 마이닝이 활용될 수 있는 가능성을 보여준다.

인할 수 있다.

이러한 결과들을 통해, 본 연구가 시도한 방법론이 실제 한의학 문헌 분석에 적용 가능함을 어느 정도 확인해볼 수 있었다. 이어서는 본론에서 발견한 내용을 토대로 원전 분야의 텍스트 마이닝 활용 방안을 논의하고, 그 과정에서 발생할 수 있는 문제점과 보완책을 검토해보도록 하겠다.

1. 텍스트 마이닝의 장점과 원전학 분야의 활용 가능성

1.1 연구 분야

일반적으로 한의학의 원전 연구는 연구자가 직접 문헌을 확인하고 분석하여 텍스트에 담긴 의학적 맥락을 발견하고 연구자의 가설을 검증하는 방식으로 이루어진다. 이에 반해 텍스트 마이닝은 컴퓨터가 사람을 대신해 텍스트를 정량화하고, 미리 프로그래밍 된 일련의 연산과정을 거쳐 결과를 도출한다. 이처럼 텍스트 마이닝의 방법론은 기존의 원전 연구 방법론과 성격이 상이하지만, 역으로 이전의 방법들로는 해결하기 어려웠던 작업을 가능하게 한다는 점에서 보완적 의미의 활용 가치가 있다. 컴퓨터의 빠르고 정확한 데이터 처리 능력을 활용해 그동안 한정된 인력과 시간으로 인해 시도하기 어려웠던 대량의 자료를 분석할 수 있으며, 문헌의 종류가 다양하고 형식이 불규칙한 경우에도 특정한 속성을 기준으로 수치화함으로써, 일괄적인 비교를 시도할 수 있다. 분석 결과를 직관적이고 이해하기 쉬운 다양한 형식으로 시각화 할 수 있다는 점도 또 하나의 장점이다.

텍스트 마이닝을 원전 연구에 적용할 때 가장 유용할 것으로 예상되는 방식은 문헌의 스크리닝(Screening)[30]이다. 연구 주제의 탐색 과정에서 대상 문헌들의 관계를 사전에 파악함으로써 새로운 가설 및 문제의식을 도출하는 계기로 삼을 수 있다. 예를 들어 텍스트 마이닝을 통해 문헌들을 시대별, 지역별로 분류하고 비교하여 경향성 여부를 파악하

30) 일반적으로 문헌 연구 분야에서 '스크리닝(Screening)'이라는 용어를 자주 사용하지는 않지만, 여기에서는 '넓은 범위의 문헌으로부터 연구 주제를 탐색하는 과정'을 의미하는 것으로 보았다.

거나, 『黃帝內經』에서 기존에 잘 인식되지 않던 편들 사이의 관련성을 발견함으로써 이를 조명할 기회를 얻을 수 있다. 이러한 시도는 직접 열람하기 어려운 분량의 문헌을 대상으로 미처 알지 못했던 경향성을 파악하고, 기존 견해로 생긴 선입관 때문에 인식되지 못했던 부분들을 관찰 가능하도록 한다는 점에서 유용하다.

이와 반대로 이미 선정된 연구 주제의 현실성과 효용성, 이미 수립된 가설의 타당성 등을 일차적으로 검토하는 파일럿 스터디(pilot study)에도 활용할 수 있다. 본격적인 연구에 앞서 현재 설정된 주제나 가설의 타당성을 검토하는 수단으로 활용하여 타당성이 낮은 가설을 효율적으로 제거할 수 있다. 같은 원리로, 정성적 연구를 통해 도출된 결론을 다수의 문헌에 일반화하여 적용할 수 있는지 확인하는 목적으로도 활용될 수 있다.

텍스트 마이닝의 활용 방안을 분야별로 구체화시켜보면 다음과 같다. 먼저 의사학 방면에서는 학술 유파를 구성하는 학자 혹은 문헌 간의 관련성을 검증하는 방법으로 활용할 수 있다. 예를 들어 『素問病機氣宜保命集』의 저자가 劉完素인지 張元素인지에 대한 논란[31]을 문헌 간 유사도 평가를 통해 검증해 볼 수 있다. 서론에서 언급한 오준호의 논문은 『東醫寶鑑』과 『醫學入門』, 『景岳全書』 사이의 연관성을 평가한 것으로 이러한 예에 해당한다.

본초와 침구 분야에서는 문헌 내에서 언급된 증상, 변증, 장부 등과 본초, 경혈의 共起(co-occurrence) 패턴을 분석함으로써 연관법칙을 확인할 수 있다. 예를 들어 본초의 경우 歸經理論의 근거를 확인해볼 수 있으며, 침구분야에서는 경락과 그에 속한 경혈들 간의 관계, 絡穴과 인접 경락과의 관련성 등을 검증해볼 수 있을 것이다. 이어서 같은 처방이나 본초, 경혈의 활용에 대한 학술 유파, 문헌, 시대 별 관점을 비교하는 파생 연구도 기대할 수 있다.

醫案 분석 역시 활용가치가 높은 분야로 예상된다. 역대 의안에 기록된 증상, 변증 내용, 처방의 유사도를 분석할 수 있으며, 최근 발표된 증례, 한방병원의 진료기록들이 문헌의 내용과 얼마나 부합하는지 검증해볼 수 있다. 뿐만 아니라 처방에 대한 문헌근거를 확보함으로서 임상진료지침의 기반 자료로 삼거나, 차후 임상 의사결정 지원시스템 개발의 가능성도 타진할 수 있다. 하지만 본초, 침구, 의안 분야에서 보다 정확한 결과를 도출하기 위해서는 관련 용어를 포괄하는 사전을 구성하고 질환과 처방의 범주를 지정하는 작업이 우선되어야 할 것으로 생각된다.

앞서 제시한 연구 주제들은 최종적으로 전문 연구자의 종합적 판단을 필요로 하며, 텍스트 마이닝 적용 과정에서의 세부적 조정 역시 전문 연구자의 지식과 경험에 의존할 수밖에 없다. 그러나 연구 주제의 획득과 가설 검토에 효율을 더하고 연구 문헌의 폭을 확장하고자 하는 경우, 기존의 방법을 보조하거나 한계를 극복하는 데에 일조할 수 있을 것이라 생각된다. 나아가 원전학 연구에 정량적, 객관적 속성을 가미하려는 시도는, 그간 정성적 연구와 정량적 연구라는 차이로 인해 이격된 양상을 보여 온 문헌연구와 실험연구, 임상연구 사이를 잇는 가교가 될 수 있다는 점에서 그 가능성을 높게 평가할 수 있다.

1.2 교육 분야

텍스트 마이닝을 활용한 연구 결과들은 교육에도 다방면으로 활용될 수 있다. 교육 분야에서 발휘될 수 있는 가장 큰 이점은 연구 결과를 다양한 방식으로 시각화할 수 있다는 것이다. 본론에 수록한 그림과 같이 『黃帝內經』의 여러 편들 사이의 관계를 조망하거나 문헌, 의가, 학파들의 계통을 네트워크로 제시할 수 있으며, 한의학의 주요 용어와 개념들 사이의 관계를 설명하는 학습 자료의 개발도 가능하다. 시각 자료의 활용은 학습 시 이해에 보탬이 될 뿐 아니라 흥미 유발 효과도 기대할 수 있다.

문헌 간 관련성을 평가한 연구 결과는 교재 편찬을 위한 기반 자료로도 이용될 수 있다. 類編 형식의 『黃帝內經』 교재, 시대별 문헌을 발췌한 형태의 의학한문 교재, 학술 유파 중심의의사학 교재 등의

31) 조대진. 素問病機氣宜保命集의 著者에 關한 考察. 경희대학교 석사학위논문. 1998.

내용과 차례 구성에 참고자료로 활용될 수 있다.

2. 한계와 가능성

2.1 한의학 원전의 특성에 따른 한계

원전학 분야의 연구 대상인 한의학 문헌은 대부분 한문으로 기록되어 있다. 한문은 단어, 구절, 문장의 구분이 명확하지 않고 조사가 발달하지 않았으며, 虛辭의 용법 역시 매우 다양하다. 때문에 의미 구조를 일괄적으로 파악하기란 쉽지 않다. 또한 뜻글자의 특성상 하나의 글자가 다양한 의미를 가지므로 모두 구분해서 정의하기 어렵다. 앞서 본론에서는 복수의 의미로 사용되는 '心'과 '脈'를 구분하지 못함에 따라 유사도가 왜곡되는 경우를 볼 수 있었다. 이처럼 텍스트 마이닝을 이용해 한의학 문헌을 분석하고자 할 때 일차적인 난점은 한문 특유의 의미 구조를 반영하기 어렵다는 것이다. 하지만 최근에는 보다 고차원적인 의미구조를 파악하기 위한 다양한 방법론이 개발되고 있으며, 차후 어느 정도 보완이 가능할 것이라 기대할 수 있다.[32]

문제의 범주를 한의학으로 좁혀보면, 한의학 문헌 내용의 많은 부분은 상징적인 언어를 통해 기술되었고, 논리 구조가 문장에 직접적으로 표현되지 않는 경우가 많다. 때문에 그것을 수치화하여 비교하기란 쉽지 않다. 본론의 분석 결과에도 이러한 예를 볼 수 있었는데, 문서 네트워크에 드러난 연결의 수와 강도가 내용의 중요도와 완전히 일치하지는 않았다. 반대로 三陰三陽이라는 형식적 유사성은 주제와 무관하게 유사도를 높이는 요인으로 작용하기도 했다.

분석 대상인 한의학 문헌들이 매우 오랜 기간에 걸쳐 생성되었다는 특징도 간과할 수 없다. 연구 범위를 통시적으로 넓게 선정할 경우 해당 용어의 의미가 달라질 수 있는데, 예를 들어 『黃帝內經』에서 命門은 '눈(目)'을 가리키며, 『難經』으로부터 현재와 같은 의미가 부여되었다. 『黃帝內經』과 『難經』의 문장을 모두 인용한 후대 문헌의 경우 문제가 더욱 복잡해질 수 있다.

2.2 기술적 한계

연구 과정 중에는 앞서 논의한 한의학 원전의 특성을 제외한 기술적인 부분의 문제도 발견할 수 있었다. 가장 원초적인 문제는 원문 파일이 구성되는 단계에서 글자 코드가 서로 다르게 부여되어 동일한 글자로 인식하지 못하는 경우이다. 또한 異體字와 通用字가 혼재되어 빈도가 부정확하게 계산될 가능성도 있다. 이는 유니코드의 정립과 같은 기술적 노력으로 해결할 수 있는 부분이다.

텍스트 마이닝의 적용 과정 중 빈도계수에 선행되어야 할 작업은 용어를 정의하고 사전을 구성하는 것이다. 용어 정의 단계에서 본 연구는 모든 글자를 용어로 지정하고 계수했지만, 이 경우 글자들의 조합으로 생성되는 의미 단위는 반영할 수 없다. 이처럼 사전을 구성하거나 문장의 의미를 분석할 때 전문가의 검토 없이 기술적으로만 구현하기에는 현실적인 한계가 있다.

사전 구성 단계에서 당면하게 되는 문제는 한의학 용어 체계의 표준화 및 전산화 미비이다. 본 연구는 진행 과정에서 총 3가지 종류의 사전을 설정하였고, 그에 따른 결과의 차이가 관찰되었다. 이를 통해 사전의 구성 방식이 분석 결과에 적지 않은 영향을 미친다는 것을 확인할 수 있는데, 연구마다 개별 사전을 구성하기에는 많은 시간과 노력이 요구될 뿐 아니라 자동화 분석의 장점도 그만큼 희석될 수 밖에 없다. 따라서 병증, 경혈, 본초, 방제 등에 관한 표준화된 용어체계 개발이 한의학 문헌 연구의 텍스트 마이닝 활용을 위한 선결과제라 할 수 있다. 이를 위해서는 기술적 개선 노력도 필요하지만, 한의학에 사용된 한자와 용어 및 논리 구조의 특성을

32) 대표적으로 2013년 구글 에서 발표한 Word2vec 모델은 내부의 인공신경망 알고리즘을 통해 전체 텍스트를 학습함으로써, 용어 간 의미관계를 보다 정확히 연산해낼 수 있다. 그 예로 충분히 학습된 Word2vec에 "woman + king - man=?" 이라는 질문을 입력하면 답은 "queen" 일 확률이 가장 높다고 대답한다. 이는 단순히 덧셈과 뺄셈의 연산이 아니라 king과 man의 관계가 queen과 woman의 관계에 대응됨을 컴퓨터가 이해하고 있음을 보여주는 예시라고 할 수 있다. 『黃帝內經』 원문 데이터만으로 Word2vec을 학습시켰을 때에도 대표적인 의미연산으로서 "陰+寒-陽=?"을 입력했을 때, 92%의 확률로 "熱"이라는 답을 얻었다. 이는 陰-寒, 陽-熱의 의미적 대응관계를 컴퓨터가 인식할 수 있다는 것을 보여 준다.

반영할 수 있어야 하므로 원전 연구자들의 적극적인 참여가 요구된다.

살펴본 한계점들을 감안할 때, 아직까지는 정량화된 방법만으로 한의학 문헌의 내용을 완벽히 분석해내거나 새로운 이론을 생산하기에는 한계가 있다. 그러나 제시한 문제점 중 상당 부분은 향후 기술 발전에 따라 해결될 수 있을 것이라 기대되며, 그 외의 문제에 대해서는 원전 분야의 전문 연구자가 연구 목표의 설정, 대상 문헌의 범주 설정, 정량화 방법에 대한 설계, 연구 결과 해석 등에 참여하는 방식으로 보완할 수 있을 것이라 생각된다.

Ⅳ. 결론

본 연구는 텍스트 마이닝(text mining)을 적용한 한의학 원전 연구의 가능성을 검토하고자 한의학의 대표 문헌인 『黃帝內經』에 해당 방법론을 적용해 분석하였으며, 그 결과를 토대로 향후 연구자들의 활용방안과 현재로서의 한계점을 고찰하였다. 적용 과정, 분석 결과, 활용 가능성, 한계 및 보완점을 다음과 같이 정리할 수 있다.

1. 『黃帝內經』에 대한 텍스트 마이닝의 적용은 불필요한 문자를 제거하는 데이터 전처리 과정, 용어(字)와 문서(篇) 단위의 텍스트 분할, 분석에 사용할 사전 구성, 빈도를 기반으로 한 용어의 속성 규정, 문서-용어의 행렬 구성, 문서 및 용어 사이의 유사도의 계산, 시각화 등의 단계로 진행할 수 있다.
2. 시각화 자료에 표현된 정보를 기존의 지식들과 비교한 결과, 중심 용어, 서술체계 등 『黃帝內經』의 문헌 속성이 텍스트 마이닝의 분석 결과에 일정 수준 반영되어 나타남을 확인할 수 있었다. 이를 통해 텍스트 마이닝을 적용한 분석이 문헌으로부터 유의미한 정보를 추출해내는 데에 활용 가치가 있음을 인정할 수 있다.
3. 텍스트 마이닝은 컴퓨터의 자동화된 데이터 처리 능력을 통해 광범위한 문헌을 다룰 수 있다는 점에서 원전학 연구 분야에 새로운 접근법을 제시할 수 있으며, 정성적 연구의 보완책으로 활용될 수 있다. 대표적으로, 문헌 스크리닝(screeing)을 통한 문제의식과 가설 도출에 활용될 수 있으며, 이미 도출된 결론의 일반화 가능성을 검증하는 방법으로도 사용될 수 있다.
4. 원전학 연구에서 텍스트 마이닝의 활용 분야를 구체적으로 예상해 보면, 학술 유파, 학자, 문헌 간의 계통성 확인과 같은 의사학적 연구, 본초 및 경혈과 병증과의 연관성 분석, 醫案을 통한 증상, 변증, 처방의 연관 패턴을 분석하고 이를 한방병원 진료기록과 비교 하는 등 파생적인 연구에 적용될 수 있다. 교육 분야에서는 시각화 자료를 활용한 수업자료와 교육과정의 개발, 교재 편찬 등에 활용될 것으로 기대된다.
5. 한의학 원전 연구에 텍스트 마이닝 적용 시 예상 가능한 문제는 크게 두 가지로 구분할 수 있다. 첫째는 한의학 문헌 본래의 특성에 기인하는 것으로, 주 언어인 한문의 의미단위 파악이 어렵다는 점, 형식으로 표현되지 않은 논리 구조를 수치화하기 어렵다는 점, 용어의 의미가 시대와 문헌에 따라 변화하는 경우가 있다는 점 등이다. 둘째는 기술적 부분으로, 異體字와 通用字, 원문 파일에 사용된 문자 코드의 부정확성, 한의학 용어 체계의 표준화 및 전산화 미비 등이 있다. 제기한 문제의 많은 부분은 향후 기술 발전을 통해 점차 해결될 것으로 기대되며, 한의학 문헌 자체의 속성에 의한 한계점은 원전 분야의 전문가가 연구 설계와 결과 해석에 참여함으로써 보완할 수 있다.

이번 연구를 통해 한의학 문헌을 대상으로 텍스트 마이닝의 활용 가능성을 검토한 결과, 원전학 연구의 효율성을 제고하고 연구 방법과 범위를 확대하는 데에 보탬이 될 수 있다는 것을 확인하였다. 그러나 이 방법론의 가치는 단순히 원전학 분야의 기술적 효용에만 국한되지 않는다. 문헌 연구를 정량화 할 수 있다는 장점을 통해, 텍스트 마이닝을 활용한 문헌 연구 결과가 실험 연구 및 임상 연구 분야와 소통할 수 있는 가교 역할을 담당함으로써 한

의학 연구 분야 전반의 단절을 해결하는 단초가 될 수 있다는 점에서 보다 큰 의미가 있다고 생각된다.

References

1. Hong WS ed.. Jeonggyohwangjenaegyeong Somun. Seoul. Publisher of Institute of Oriental Medicine. 1985.
洪元植. 精校黃帝內經素問. 서울. 동양의학연구원출판사. 1985.
2. Hong WS. Jeonggyohwangjenaegyeong YoungChu. Seoul. Publisher of Institute of Oriental Medicine. 1985.
洪元植. 精校黃帝內經靈樞. 서울. 동양의학연구원출판사. 1985.
3. Christofides N. Graph theory: An algorithmic approach (Computer science and applied mathematics). Orlando, USA. Academic Press, Inc. 1975.
4. Kim JW, et al.. Inferring Disease-related Genes using Title and Body in Biomedical Text. KIISE Transactions on Computing Practice. 2017. 23(1).
김정우 外 4人. 생물학 문헌 데이터의 제목과 본문을 이용한 질병 관련 유전자 추론 방법. 정보과학회 컴퓨팅의 실제 논문지. 2017. 23(1).
5. Lee HC, et al. Extraction of the protein-protein interaction using text mining technique. Journal of the Research Institute for Computer and Information Communication. 2004. 12(1)
이현철 外 4人. 단백질 상호작용 추출을 위한 텍스트 마이닝 기법. 컴퓨터정보통신연구, 2004. 12(1).
6. Park HS, et al.. Data Engineering : TF-IDF Based Association Rule Analysis System for Medical Data. Korea Information Processing Society review. 2016. 5(3).
박호식 外 3人. 데이터 공학: 의료 정보 추출을 위한 TF-IDF 기반의 연관규칙 분석 시스템. 정보처리학회논문지. 소프트웨어 및 데이터 공학. 2016. 5(3).
7. Park CY, et al.. A Study on Terminology in ZhenJiuJiaYiJing. The Journal Of Korean Medical Classics. 2013. 26(3).
박찬영, 이병욱, 김기욱.『鍼灸甲乙經』의 用語體系에 관한 연구. 대한한의학원전학회지. 2013. 26(3).
8. Song IW, Lee BW. Method for improving search efficiency using relation of anatomical structure from Donguibogam. The Journal Of Korean Medical Classics. 2012. 25(4).
송인우, 이병욱.『동의보감』에 기재된 인체 용어 관계를 이용한 검색효율성 향상 방법. 대한한의학원전학회지. 2012. 25(4).
9. KIM MH, et al.. A Study of classification the predicate in『Biwiron(脾胃論)』. The Journal Of Korean Medical Classics. 2010. 23(1).
김명희, 이병욱, 김은하. 비위론에 기재된 술어의 분류에 관한 연구. 대한한의학원전학회지. 2010. 23(1).
10. KIM MG, et al.. The Comparative Study of the Nominal Terms between『Biwiron(脾胃論)』and『Soayakjeungjikgyeol(小兒藥證直訣)』. The Journal Of Korean Medical Classics. 2010. 23(1).
김민건, 이병욱, 김은하. 小兒藥證直訣과 脾胃論에 기재된 용어 비교에 관한 연구. 대한한의학원전학회지. 2010. 23(1).
11. Beak JU, Lee BW. A study on the frequencies of medicinal herb combinations in the prescriptions of『Bangyakhappyeon(方藥合編)』. The Journal Of Korean Medical Classics. 2011.24(4).

백진웅, 이병욱.『方藥合編』收錄 處方 內의 藥物 조합 頻度 연구. 대한한의학원전학회지. 2011.24(4).

12. Wu YH, et al. Analysis of Prescriptions from Taepyeonghyeminhwajegukbang, Somunsunmyungronbang and Nansilbijang. The Journal Of Korean Medical Classics. 2014. 27(4).
오월환 외 3人.『太平惠民和劑局方』과『素問宣明論方』과『蘭室秘藏』의 방제구성 비교. 대한한의학원전학회지. 2014. 27(4).
13. Kim KW, et al.. Automatic Extraction Method of Compositional Herb Using Herb List. The Journal Of Korean Medical Classics. 2014. 27(3).
김기욱, 김태열, 이병욱. 본초 목록을 이용한 방제의 본초 구성 자동 추출 방법. 대한한의학원전학회지. 2014. 27(3).
14. Oh JH. Can Similarities in Medical thought be Quantified? -Focusing on Donguibogam, Uihagibmun and Gyeongagjeonseo-. The Journal Of Korean Medical Classics. 2018. 31(2).
오준호. 의학 사상의 유사성은 계량 분석 될 수 있는가. 대한한의학원전학회지. 2018. 31(2).
15. Lee TH, et al. A Structural Analysis of Acupuncture & Moxibustion Points in the NaeGyeong Chapter of DongUiBoGam Using Text Mining. Korean Journal of Acupuncture. 2013. 30(4).
이태형 외. 텍스트마이닝을 이용한 동의보감 질병인식방식과 내경편 경혈 분석. 대한경락경혈학회지. 2013. 30(4).
16. Vishal Gupta and G.S. Lehal. A Survey of Text Mining Techniques and Applications. Journal of Emerging Technologies in Web Intelligence. 2009. 1(1).
17. Salton G, Wong A, Yang CS. A vector space model for automatic indexing. Commun ACM. 1975. 18(11)
18. Zellig S. Harris. Distributional Structure. WORD. 1954. 10.
19. Wu HC et al.. Interpreting TF-IDF term weights as making relevance decisions. ACM Trans Inf Syst. 2008. 26(3).
20. P. D. Turney, P. Pantel. From Frequency to Meaning: Vector Space Models of Semantics. Journal of Artificial Intelligence Research. 2010. 37.
21. Ye J. Cosine similarity measures for intuitionistic fuzzy sets and their applications. Mathematical and Computer Modelling. 2011. 53(1).
22. Shannon P, et al. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Research. 2003. 13(11).
23. Jo DJ. A Study on the Writer of Bao Ming Shi. Kyunghee Univ. 1998.
조대진. 素問病機氣宜保命集의 著者에 關한 考察. 경희대학교 석사학위논문. 1998.