J Physiol & Pathol Korean Med 34(4):159~169, 2020

텍스트마이닝 기법을 이용한 『상한론』 내의 증상-본초 조합의 탐색적 분석

장동엽 · 하윤수[1] · 이충열* · 김창업*

가천대학교 한의과대학 생리학교실, 1 : 가천대학교 한의과대학

Analysis of Symptoms-Herbs Relationships in Shanghanlun Using Text Mining Approach

Dongyeop Jang, Yoonsu Ha[1], Choong-Yeol Lee*, Chang-Eop Kim*

Department of Physiology, College of Korean Medicine, Gachon University, 1 : College of Korean Medicine, Gachon University

Shanghanlun (Treatise on Cold Damage Diseases) is the oldest document in the literature on clinical records of Traditional Asian medicine (TAM), on which TAM theories about symptoms-herbs relationships are based. In this study, we aim to quantitatively explore the relationships between symptoms and herbs in Shanghanlun. The text in Shanghanlun was converted into structured data. Using the structured data, Term Frequency - Inverse Document Frequency (TF-IDF) scores of symptoms and herbs were calculated from each chapter to derive the major symptoms and herbs in each chapter. To understand the structure of the entire document, principal component analysis (PCA) was performed for the 6-dimensional chapter space. Bipartite network analysis was conducted focusing on Jaccard scores between symptoms and herbs and eigenvector centralities of nodes. TF-IDF scores showed the characteristics of each chapter through major symptoms and herbs. Principal components drawn by PCA suggested the entire structure of Shanghanlun. The network analysis revealed a 'multi herbs - multi symptoms' relationship. Common symptoms and herbs were drawn from high eigenvector centralities of their nodes, while specific symptoms and herbs were drawn from low centralities. Symptoms expected to be treated by herbs were derived, respectively. Using measurable metrics, we conducted a computational study on patterns of Shanghanlun. Quantitative researches on TAM theories will contribute to improving the clarity of TAM theories.

keywords : Shanghanlun, Text mining, Symptoms-herbs relationships

서 론

『상한론』은 『황제내경』과 더불어 한의학의 가장 오래되고, 가장 중요한 고전이다. 특히 『황제내경』은 한의학의 근간이 되는 개념들에 대해 서술한 이론서인 반면, 『상한론』은 환자를 직접 치료하고 그 경과를 관찰한 내용을 기록한 최고(最古)의 임상서이다[1]. 특히 『상한론』의 내용을 어떻게 이해하고 해석하는가에 따라 다양한 임상 유파가 생겨나기도 하였다[2]. 『상한론』의 이론과 처방은 현대에도 그대로 이용되고 있어[3,4], 『상한론』을 이해하는 것은 현대 한의학을 이해하기 위해서도 매우 중요하다.

텍스트마이닝이란 텍스트에서 의미 있는 규칙과 패턴을 발견하는 기법을 말한다[5]. 텍스트는 방대하고 풍부한 정보를 표현하지만 컴퓨터가 해독하기 어려운 비정형적인 구조를 갖고 있다. 따라서 텍스트마이닝은 텍스트를 컴퓨터가 처리할 수 있는 형태로 구조화하고 구조화 된 데이터 내에서 패턴을 도출하는 과정으로 구성된다. 현재 텍스트마이닝은 다량의 텍스트 데이터에서 자동으로 의미를 도출하기 위해 IT산업, 사회연구 등 여러 분야에서 사용된다[6-8]. 텍스트를 중요시하는 한의학의 특성을 고려하였을 때, 한의학의 텍스트를 바탕으로 텍스트마이닝을 진행할 경우 여러 한계로 발견하지 못했던 새로운 지식을 발굴할 수 있다.

현대 한의학의 가장 큰 특징 중 하나는 한의학의 과학화이다. 20세기 접어들어 동아시아에 서구식 근대화가 가속화되면서 전통의

* Corresponding author

Choong-Yeol Lee, College of Korean Medicine, Gachon University, 1342 Seongnamdaero, Sujeong-gu, Seongnam-si, Republic of Korea

·E-mail : cylee@gachon.ac.kr ·Tel : +82-31-750-5419

Chang-Eop Kim, College of Korean Medicine, Gachon University, 1342 Seongnamdaero, Sujeong-gu, Seongnam-si, Republic of Korea

·E-mail : eopchang@gachon.ac.kr ·Tel : +82-31-750-5493

·Received : 2019/10/29 ·Revised : 2020/07/27 ·Accepted : 2020/07/28

pISSN 1738-7698 eISSN 2288-2529 http://dx.doi.org/10.15188/kjopp.2020.08.34.4.159

Available online at https://kmpath.jams.or.kr

학 체계에도 합리성을 요구하기 시작하면서[9], 한의학에 대한 과학적 연구가 증가하고 있다. 현재 대세를 이루고 있는 한의학의 과학화는 한의학 치료의 안전성과 유효성을 과학적인 방법을 이용하여 규명하는 것으로, 동물실험이나 임상시험이 주를 이루고 있는 반면[10] 한의학의 이론체계에 대한 과학적 논의는 상대적으로 부족하다는 한계가 있다. 특히 한의학 개념들은 많은 경우 추상적이고, 정량적인 지표에 기반하지 않으며, 과학적 합리성 및 반증가능성이 부족하다는 평가를 받는다. 한의학 이론 연구의 비정량성은 한의학 이론 체계가 과학과의 접점을 쉽게 만들지 못하는 이유가 된다. 지금의 추세는 한의학 치료와 이론이 괴리되게 만들며, 심하게는 한의학 이론이 배제된 채 양의학적 진단에 한의학의 치료가 활용되는 '폐의존약(廢醫存藥)' 경향이 고착화되도록 만들 수 있다[10]. 따라서 한의학 이론을 정량화하는 것은 한의학의 과학화에 있어 필수적이며, 궁극적으로는 정량화된 한의학 이론을 바탕으로 한의학의 치료를 과학적으로 이해할 수 있는 교두보가 되게 한다.

한의학의 원전을 정량적으로 연구한 선행 연구는 다음과 같다. 『방약합편』의 경우 대변과 약물 간의 연관성을 Apriori rule mining을 바탕으로 분석하거나 등장하는 한의학 용어 개념을 정량화한 연구가 있었다[11,12]. 또 『황제내경』에 등장하는 한자 간의 관계와 소문과 영추 각 81편 간의 관계를 Term Frequency - Inverse Document Frequency (TF-IDF)와 cosine similarity를 활용하여 밝힌 연구도 이루어졌다[13]. 『동의보감』의 경우 『동의보감』에 등장하는 경혈의 공간적 패턴을 텍스트마이닝으로 접근하여 시각화한 연구가 진행되었으며[14], 『동의보감』, 『의학입문』, 『경악전서』의 유사성을 계량 분석하고자 한 연구가 존재한다[15]. 한편 『상한론』에 대한 연구의 경우 『상한론』 조문을 데이터베이스화하여 정형화한 연구와[16] 『상한론』에 등장하는 증상을 계통별로 분류하고 유사한 표현을 통합한 연구[17] 등 『상한론』을 구조화하는 연구가 진행되었다.

본 연구에서는 『상한론』에 등장하는 증상과 본초가 어떤 관계를 맺는지 정량적으로 인식하기 위해 『상한론』의 탕증을 바탕으로 단원과 증상, 본초의 관계를 분석하였다. 이를 위해 『상한론』의 탕증 텍스트에 등장하는 증상과 본초를 구조화된 데이터로 변환하고, TF-IDF, principal component analysis (PCA), Jaccard score, eigenvector centrality와 같은 정량적인 기준을 바탕으로 각 단원, 증상, 본초의 특징을 분석하였으며 결과를 데이터과학적 기법을 활용하여 시각화하였다.

연구방법

1. 텍스트 수정 및 증상 및 처방 사전 제작

『상한론』의 조문에 증상이 등장하는 횟수를 집계하기 위해, 선행연구를 바탕으로[17] 『상한론』 증상 사전을 제작하였다. 선행연구를 간략하게 요약하자면, 『상한론』의 조문에서 증상을 추출하여 증상들을 계통별로 분류하고 유사 표현을 통합하여 『상한론』 증상을 체계적으로 정리한 연구로, 임상에서 상한병에 대해 증상을 기록할 때에 그 기준으로 활용하고자 하였다는 의의가 있다. 본 연구는 위 선행논문에서 제시하는 대분류 및 세부분류와 증상을 집계하는 기준을 따랐다. 또한 선행논문의 기준에 따라 원문에 등장하는 증상의 표현을 일부 변경하였다. 예를 들어, 28번에 등장하는 頭項强痛이란 표현은 頭痛과 項强으로 각각 분류될 수 있으므로 따라서 사전에 수록되어 있는 '頭痛'과 '項强'으로 컴퓨터가 인식할 수 있게 '頭項强痛'을 '頭痛項强'으로 변경하는 작업을 진행하였다. 한편 증상의 분류를 세밀화하고 선행논문이 다루는 증상보다 더 넓은 범위의 증상을 다루기 위해 선행논문의 기준을 일부 개선하였는데, 이는 다음과 같다. 첫 번째로, 증상의 대분류 중 신경정신과적 증상과 심흉부의 증상을 분리하였다. 선행논문에서는 언어곤란, 수면장애 등 신경정신과적 증상과 심번(心煩), 심계(心悸), 심통(心痛), 호흡기계 증상 등 심흉부의 증상을 같은 범주로 분류하였는데, 서양의학적 기준에서 이는 뇌신경계의 증상과 흉부의 증상으로 나누어지기 때문에 본 연구에서는 이를 구분하였다. 두 번째로, 이전에 어떤 치료를 받고 부작용이 생겼는지(오치(誤治))에 대한 정보를 증상의 범주에 포함하였다. 본 연구에서의 15개의 대분류 중 '이전에 받았던 치료'는 선행논문에서 다루지 않았던 내용이다. 이는 엄밀히 말하면 증상이 아니지만, 오치의 유무는 망문문절을 통해 환자로부터 객관적으로 파악할 수 있는 정보이며 특히 『상한론』에서는 많은 조문에서 오치를 중요한 정황 근거로 다루기 때문이다. 따라서 조문에서 설명하는 바를 최대한 반영하기 위해 오치를 광의의 증상에 포함시켰다. 세 번째로, 두항부의 증상 중 '어지러움'은 『상한론』에 등장함에도 선행 논문에서 증상에 포함시키지 않아, 본 연구에서 추가하였다. 마지막으로, 본 연구에서는 특정 증상이 증상이 없다고 언급된 경우를 별도의 증상으로 분류하고 이를 구분하여 집계하였다. 이와 같이 개선한 기준을 바탕으로 『상한론』의 증상을 정리한 결과 총 702개의 증상을 15개의 대분류와 82개의 세부분류로 나누었다(Table 1). 한편 본초의 경우, 각 조문의 처방을 본초로 치환하기 위해 각 처방을 구성하는 본초들의 목록으로 처방 사전을 제작하였다.

Table 1. The categorization of symptoms in Shanghanlun

대분류	세부분류
열 관련 증상	오한, 오한이 없음, 오풍, 발열, 발열이 없음, 조열, 수족궐냉, 수족궐냉이 없음, 한열왕래, 한열왕래가 없음, 오열, 기타 열 증상
땀 관련 증상	한출, 한출 없음, 두면부 한출, 도한, 수족부 한출
전신의 증상	신통, 신통 없음, 신중, 신체 떨림, 기타 전신 증상
맥상	부맥, 부맥 없음, 침맥, 삭맥, 지맥, 실맥, 허맥, 장단맥, 촌관척의 맥, 음양맥, 욕절맥, 화맥, 기타 맥
안이비인후의 증상	눈, 코, 비출혈, 귀, 입, 혀, 인후, 갈증, 갈증 없음
대변 관련 증상	대변난, 하리, 하리 없음, 대변당, 대변출혈, 기타 대변 증상
비뇨기 관련 증상	소변난, 소변불난, 소변삭, 소변불삭 소변적, 소변청 기타 비뇨기 증상
언어, 정신, 수면, 번조 등의 증상	언어곤란, 정신 이상, 수면장애, 다수면, 기타 정신 증상
구토, 복부 증상 등 소화기 관련 증상	구토, 구토 없음, 헛구역질 딸꾹질 트림, 심하부 불편감, 복부창만 및 복부통증, 소복부 증상, 실기, 실기 없음, 소화부진, 소화 좋음
흉협부의 증상	흉부의 통증, 심흉부 번조, 심흉부 번조 없음, 심흉부 계, 협부의 증상, 호흡기계의 증상, 호흡기계 증상 없음
피부 관련 증상	얼굴색의 변화, 전신 피부 변화
두항부 증상	두부 통증, 두부 통증 없음, 어지러움, 경항부 강직
사지관절의 증상	사지 통증, 사지 당김 및 경련
기생충 관련 증상	기생충 관련 증상
이전에 받았던 치료	한법 사용, 토법 사용, 하법 사용, 소침 사용

2. 분석 대상 조문의 선정

본 연구에서는 『상한론』 송판(宋板)을 대상으로 분석하였으며, 『상한론』 398개 조문 중 이론에 대한 설명, 금기증 등 처방의 적응증이 아닌 내용은 제외한 처방의 적응증을 설명하는 218개 조문만을 연구 대상으로 선정하였다. 예를 들어, 19번 조문인 '凡服桂枝湯吐者 其後必吐膿血也'은 계지탕(桂枝湯)을 복용한 후 나타나는 반응에 대해 서술하고 있으므로, 증상과 그 증상을 치료하기 위한 본초 간의 관계를 연구한다는 주제에 포함되지 않아 분석에서 제외하였다. 그러나 처방에 대해 직접적으로 서술하지는 않지만 중요한 개념을 정의하는 조문은 분석에 반영될 수 있게 하였다. 예를 들어 1번 조문은 처방에 대한 설명이 아닌 태양병을 정의하는 조문이므로 해당 조문을 직접 분석하지는 않았지만, '太陽病'이 등장하는 32번 조문의 증상을 집계할 때 1번 조문에서의 태양병의 정의에 포함된 '脈浮', '頭痛', '項强', '惡寒' 증상을 추가로 집계하여 주요 개념을 분석에 활용하였다. 한편 처방의 적응증을 설명하는 조문 중 한 조문 내에서 여러 처방의 적응증을 설명하는 경우에는 해당 조문을 분리하여 독립적인 조문으로 다루었다. 예를 들어 29번 조문에서는 감초건강탕(甘草乾薑湯), 작약감초탕(芍藥甘草湯), 조위승기탕(調胃承氣湯), 사역탕(四逆湯)을 활용해야 하는 경우를 다루는데, 각 처방에 대한 조문으로 나눠 총 4개의 세부조문으로 분리하였다. 이와 같은 과정을 통해 218개의 처방 조문을 245개의 세부조문으로 분리하였다.

한편 본 연구에서는 조문을 '태양병 상편', '태양병 중편', '태양병 하편', '양명병', '소음병', '궐음병' 총 6개의 '단원'으로 분류하였다. 태양병의 내용은 상중하편에 따라 이질적인 내용이 같은 태양병 범주로 묶여 있다는 문제가 있다. 따라서 태양병의 조문을 태양병 상편, 태양병 중편, 태양병 하편으로 재분류하였다. 또한 소양병, 태음병은 해당되는 조문의 갯수가 5개 미만으로 연구결과에 비뚤림(bias)을 만들 우려가 있어 본 연구의 분석 대상에서 제외하였다. 또한 곽란병, 음양역차후노복병은 일반적인 육경의 범주에 포함되지 않아 본 연구에서는 분석 대상에서 제외하였다. 그 결과 태양병 상편 조문 20개, 태양병 중편 조문 74개, 태양병 하편 조문 43개, 양명병 조문 47개, 소음병 조문 25개, 궐음병 조문 19개의 조문을 분석하였다(Table 2).

Table 2. Counts and percentages of sentences in each chapter

단원		탕증 조문 수	비율(%)	비고
태양병	상편	20	8.1	
	중편	74	30.1	
	하편	43	17.5	
양명병		47	19.1	
소양병		1	0.4	분석에서 제외
태음병		4	1.6	분석에서 제외
소음병		25	10.2	
궐음병		19	7.7	
곽란병		7	2.8	분석에서 제외
음양역차후노복병		6	2.4	분석에서 제외
총합		245	100	

3. 데이터 정형화

컴퓨터가 연산할 수 있는 형태의 데이터를 제공하기 위해 증상, 처방 사전을 바탕으로 각 조문을 702차원의 증상 벡터와 89차원의 본초 벡터로 표현하였다. 증상 벡터는 해당 증상이 있으면 1, 없으면 0인 이진형의 702차원의 벡터이며, 본초 벡터는 처방에 해당 본초가 포함되면 1, 없으면 0인 이진형의 89차원의 벡터이다. 그러나 증상 벡터의 경우 0이 많은 희소 벡터(sparse vector)로 과적합(overfitting)의 원인이 될 수 있으므로, 사전의 세부분류를 활용하여 비슷한 증상은 같은 증상으로 통합하였다. 그 결과 각 증상 벡터를 78차원의 이진형 벡터로 재구성하였다.

4. TF-IDF

태양병 상편, 태양병 중편, 태양병 하편, 양명병, 소음병, 궐음병 총 6개 단원을 문서로, 각 증상과 본초를 단어로 TF-IDF를 계산하여 각 단원을 대표하는 증상을 도출하였다. TF-IDF는 문서군이 주어졌을 때 한 문서에서 어떤 단어가 중요한지 나타내는 수치로, 한 문서에서 각 단어들의 절대빈도인 Term Frequency (TF)와 단어가 얼마나 특이하게 등장하는지를 나타내는 Inverse Document Frequency (IDF)의 곱으로 계산된다. 따라서 단어들의 TF-IDF의 패턴이 비슷한 문서는 내용이 비슷한 문서라고 추론할 수 있다. TF, IDF, TF-IDF를 구하는 식은 각각 다음과 같다.

$$D=\{d_1,d_2,...d_\alpha\},\ T=\{t_1,t_3,...t_\beta\}$$

$$TF_{i,j}=\frac{n_{i,j}}{\sum_k n_{k,j}} (1)$$

($n_{i,j}$: 단어 t_i가 문서 d_j에서 출현한 횟수, $\sum_k n_{k,j}$: 문서 d_j에서 모든 단어가 출현한 횟수)

$$IDF_i=\log\frac{|D|}{|\{d_j|t_i\in d_j\}|} (2)$$

($|D|$: 문서집합에 포함되어 있는 문서의 수, $|\{d_j|t_i\in d_i\}|$: 단어 t_i가 등장하는 문서의 수)

$$TFIDF_{i,j}=TF_{i,j}\times IDF_i$$

이와 같은 계산을 통해 TF는 각 단원에서 특정 단어가 차지하는 비율을 나타내는 반면, TF에 IDF가 곱해진 TF-IDF는 단원 내에서의 단어의 빈도를 반영할 뿐 아니라 해당 단원에서 얼마나 특이적으로 등장하는지를 나타낸다. 본 연구에서는 TF와 TF-IDF 분석 결과를 동시에 제시하여, 절대적인 빈도와 상대적인 빈도를 동시에 비교하였다.

5. PCA

앞서 계산된 각 단원에 대한 증상의 TF를 바탕으로 PCA를 시행하여 고차원 증상 행렬의 정보를 효율적으로 나타내는 차원을 탐색하였으며, 본초의 TF를 PC (Principal component)에 투사하여 조문들의 증상 정보와 본초 정보의 패턴을 시각화하였다. PCA는 주어진 데이터를 가장 잘 설명하는 축을 찾는 방법으로, 각 축을 해석함으로써 데이터에서 가장 설명력이 높은 기준을 찾아낼 수 있다. 본 연구에서는 6차원인 각 증상 및 본초 벡터의 길이가 1이 되도록 정규화한 뒤, 증상 행렬을 2차원으로 축소하여 2차원 평면 상에 각 증상의 정보를 시각화하였다. 이후 본초 벡터를 주성분에 투사하여 각 본초의 정보를 시각화하였다.

6. 네트워크 분석

각 본초가 어떤 증상과 연관되는지 파악하기 위해 각 증상과

본초 간의 Jaccard score를 계산하고, 이를 바탕으로 네트워크 분석을 시행하였다. Jaccard score의 식은 다음과 같다.

$$J(a_i, b_j) = \frac{|a_i \cap b_j|}{|a_i \cup b_j|} \quad (4)$$

(a_i, b_j는 각각 i번째 증상 벡터와 j번째 본초 벡터)

Jaccard score는 증상-본초 쌍은 증상 벡터와 본초 벡터의 교집합을 합집합으로 나눈 것으로, Jaccard score가 높은 증상-본초 쌍은 두 증상과 본초가 밀접한 관계를 맺고 있다고 추론할 수 있다. 계산된 Jaccard score를 바탕으로 bipartite 네트워크 분석을 시행하였으며, 네트워크 구성에 필요한 Jaccard score의 cutoff는 1000배수 permutation test에서 상위 5%의 Jaccard score로 결정하였다. 또한 구성된 네트워크에서 노드의 eigenvector centrality를 계산하여 보편적인 증상 및 본초와 특이적인 증상 및 본초를 구별하였다. 그래프 $G := (V, E)$에 대한 인접행렬 $A = (a_{v,t})$가 주어졌을 때, 노드v의 eigenvector centrality x_v의 식은 다음과 같다.

$$x_v = \frac{1}{\lambda} \sum_{t \in M(v)} x_t = \frac{1}{\lambda} \sum_{t \in G} a_{v,t} x_t \quad (5)$$

($M(v)$는 노드 v와 연결된 노드들의 집합이며, λ는 상수)

Eigenvector centrality는 해당 노드와 연결된 이웃 노드들이 얼마나 영향력이 있는지를 바탕으로 해당 노드의 영향력을 측정하는 지표로, 값이 클수록 다양한 노드들과 영향을 주고 받으며, 값이 작을수록 제한된 노드에 영향을 주고 받는다는 의미를 갖는다. 특히 해당 노드가 단순히 얼마나 많은 노드와 연결되어 있는지를 나타내는 차수(degree)와는 달리, eigenvector centrality는 해당 노드와 연결되어 있는 노드들의 영향력을 반영하여 계산되기 때문에 네트워크의 전체 형상(topology)을 반영할 수 있다는 장점이 있다.

7. 소프트웨어

본 연구의 데이터 분석 및 시각화는 범용 프로그래밍 언어인 Python(v3.6.5)을 이용해 수행되었으며, 네트워크 시각화는 Cytoscape (v3.6.0)를 이용해 수행하였다.

결 과

1. 단원 별 핵심 증상 및 본초 분석

각 단원의 핵심 증상과 본초를 탐색하기 위해 TF 및 TF-IDF를 계산하였다(Fig. 1). 태양병은 상편, 중편, 하편의 증상에 차이가 있었다. 구체적으로 태양병 상편과 중편은 오한, 부맥(浮脈) 두부통증과 같이 풍한사(風寒邪)에 감응되었을 때 나타나는 증상의 비중이 높았다. 그러나 상편에서는 표허(表虛)의 대표적인 증상인 한출(汗出)이, 중편에서는 발열이 높은 TF를 보여, 상편과 중편이 다루는 상태에 차이가 있음을 나타냈다. 또한 중편과 하편에서는 하법(下法)을 사용하였다는 과거력에 대한 TF가 높아, 병의 기간이 길어지거나 오치했을 때 나타나는 증상을 중편 및 하편에 배치하고 있다는 사실과 일치한다. 특히 하편에서는 심하부의 불편감에 대한 증상의 TF가 높았는데, 이는 태양병 하편에서 결흉(結胸)을 다루는 것과 연관이 있다. 한편 양명병에서는 발열, 한출, 오열(惡熱), 대변난(大便難)과 같은 열사로 인한 증상의 TF가 높고, 소음병과 궐음병은 허맥, 하리, 수족궐냉, 다수면과 같은 허증(虛證)을 구성하는

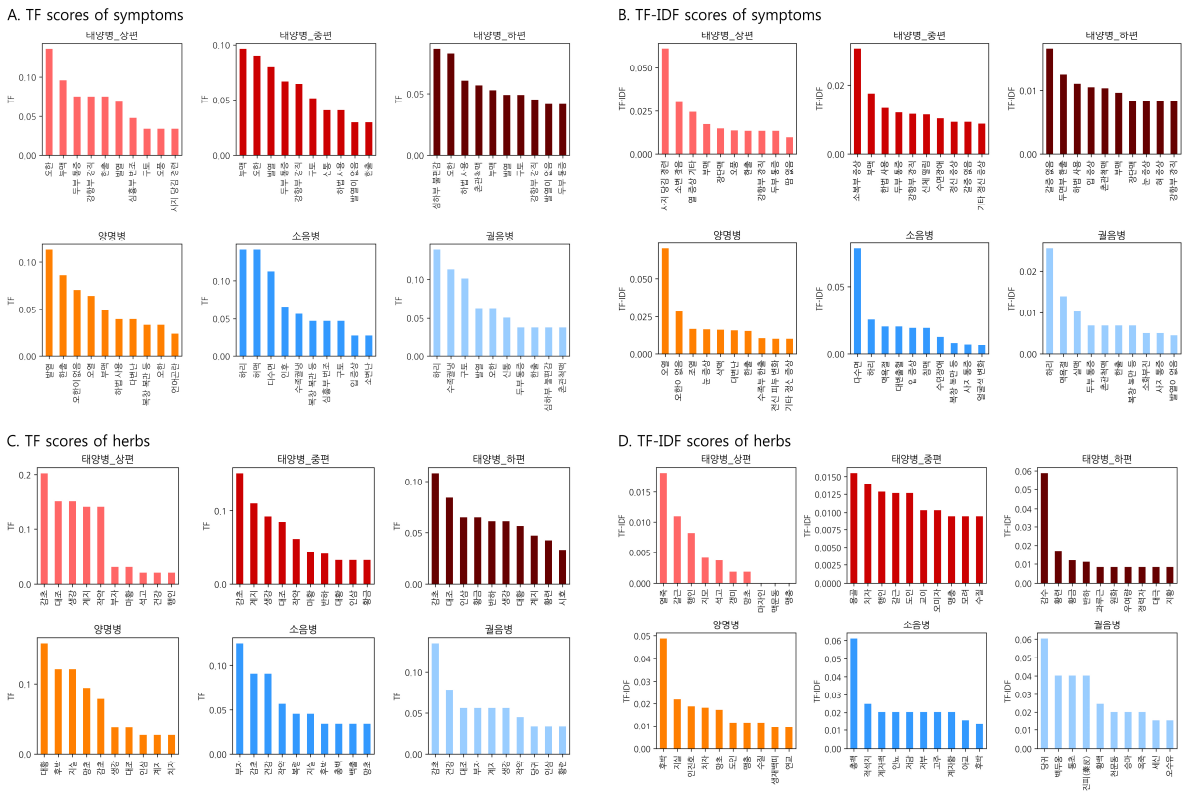

Fig. 1. Symptoms and herbs with high term frequency (TF) and term frequency-inverse document frequency (TF-IDF) scores in each chapter.

症의 TF가 높았다. 이는 양병(陽病)과 음병(陰病)의 차이가 허실(虛實)에 있음을 나타낸다. 다음으로 각 단원에 대해 특이적인 증상을 알아보기 위해 TF-IDF값을 분석하였다. 태양병 상편은 경련이 높은 TF-IDF값을 나타냈는데, 이는 태양병 상편에서 계지탕으로 오치 후 양허증(陽虛證)이 생긴 경우를 다루기 때문이다. 태양병 중편에서는 소복부 증상이 특이적이며 이는 태양병 중편이 표사가 입리(入裏)하여 발생하는 축혈증(蓄血證) 등을 다루기 때문이다. 태양병 하편에서는 '갈증이 없는' 경우가 특이적으로 나타났는데, 이는 해당 증상을 양명병 및 소양병과 구분하기 위해 언급하기 때문이다. 양명병, 소음병, 궐음병의 경우 각각 오열, 다수면, 하리가 특이적인 증상으로 등장해 각 단원의 특징을 보인다.

본초의 경우 태양병 상편과 중편에서는 계지탕의 구성 본초인 감초, 대조, 계지, 생강, 작약이 높은 빈도로 등장하였고, 특히 중편에서는 마황, 반하 등의 본초의 빈도가 높았다. 이와 같은 결과는 태양병 상편과 중편에서 계지탕을 기본으로 가감한 처방을 빈용하기 때문이다. 반면 태양병 하편에서는 인삼-황금-반하 등의 본초 비율이 높아, 울열(鬱熱)을 해소하기 위해 사심탕(瀉心湯)류의 처방을 빈용하는 것을 확인하였다. 양명병에서는 대변 배출을 원활하게 하기 위해 대황, 후박, 지실 등의 본초를 다용하며, 소음병과 궐음병에서는 보양을 위해 부자, 건강 등을 빈용하였다. 다음으로 각 단원의 특이적인 본초를 파악하기 위해 TF-IDF를 분석하였다. 태양병 중편에서는 용골이 높은 TF-IDF를 보이며, 태양병의 초기에 나타나는 열증의 해열에 빈용된다. 갈근과 행인은 태양병 상편 및 중편에서 특이적으로 활용되는 것으로 나타나, 표사 감응 초기에 활용되고 있음을 확인하였다. 이밖에 태양병 하편에서는 감수, 양명병에서는 후박, 소음병에서는 총백, 궐음병에서는 당귀가 높은 TF-IDF를 보였다.

2. 단원 간의 관계 분석

증상 패턴의 다양성을 포착하는 최선의 축을 탐색하기 위해, 증상들의 빈도로 정의되는 6차원의 단원 공간에 PCA를 수행하였다(Fig. 2). PC1에서 가장 크게 분리되는 단원은 태양병 상편과 양명병으로, 태양병 상편에 등장하는 증상과 양명병에 등장하는 증상이 가장 이질적인 것으로 나타났다. 다음으로 PC2를 통해 태양병 상편과 중편에서 등장하는 증상과 소음병 및 궐음병에서 등장하는 증상이 두 번째로 이질적인 것으로 나타났다. 증상을 바탕으로 위의 결과를 해석하였을 때, PC1은 오열, 조열(潮熱), 대변난, 무오한(無惡寒) 등 실열(實熱)로 인한 증상과 계지탕증을 포함한 허증(虛證)을 나누는 축으로 기능한다고 해석할 수 있다. 즉 PC1으로 분리되는 오열, 대변난, 조열, 황달과 같은 실열증의 증상들과 계지탕증으로 대표되는 표허의 증상은 한 단원 내에서 동시에 다루는 경우가 비교적 적다는 의미로, 『상한론』에서는 실열로 인한 증상과 표허로 인한 증상을 구분하여 서술하고 있다고 추측할 수 있다. 더 나아가 『상한론』이 다양한 임상경험을 망라했다는 전제 하에 실열의 증상들과 표허의 증상들은 임상에서 동시에 나타나는 경우가 드물다고 해석할 수 있다. 한편 PC2는 허맥(虛脈), 다수면(多睡眠), 하리(下利), 침맥(沈脈), 수족궐냉 등 허증(虛證)을 구성하는 증상과 부맥, 오풍과 같은 표증(表證)을 구성하는 증상을 구분하는 축으로 기능한다고 해석할 수 있다. PC2로 분리되는 허맥, 맥욕절(脈欲絶), 하리, 수족궐냉과 같은 양허증은 표사(表邪)로 인한 증상과 함께 서술하지 않으므로, 표사로 인한 증상과 양허증의 증상은 동시에 나타나지 않는다고 추측할 수 있다. 또한 증상 공간에서 도출한

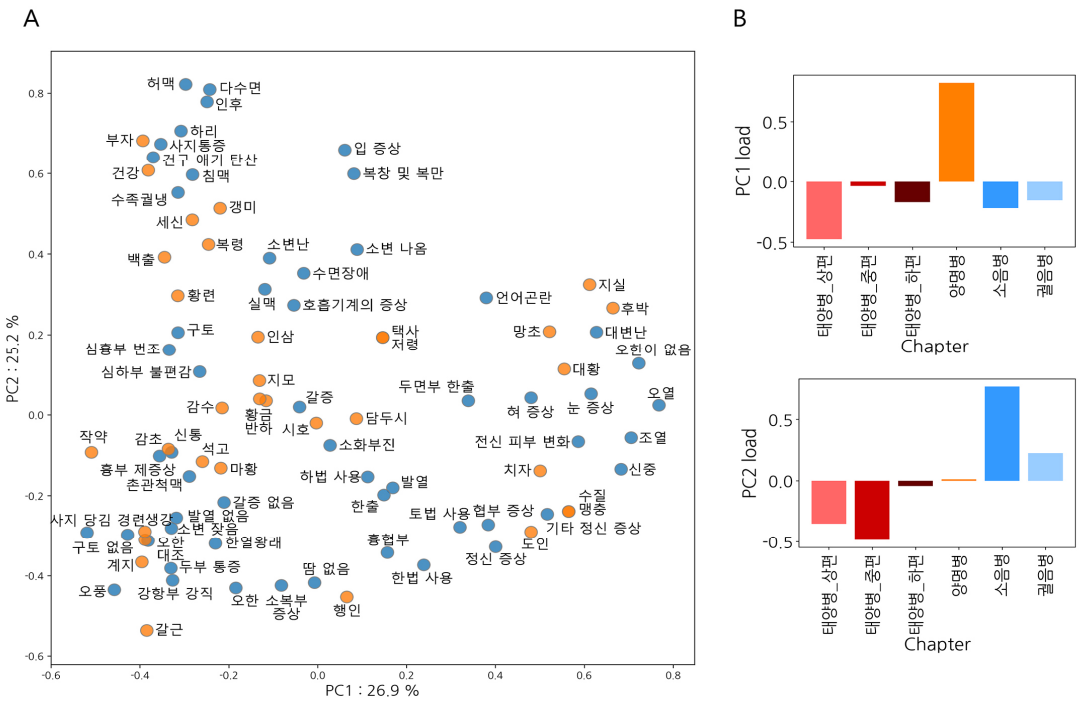

Fig. 2. The principal component analysis is conducted for the 6-dimensional chapter space of symptoms into 2 principal components (PC) space. (A) Symptoms and herbs are projected into PC space. Blue and orange colors indicate symptoms and herbs, respectively. (B) Load vectors of the PC1 (top) and PC2 (bottom) are represented.

PC가 본초 공간에도 적용될 수 있는지 알아보기 위해, 본초 공간을 증상 공간에서 도출한 PC에 투사하여 시각화하였다. 그 결과 본초와 해당 본초의 적응증이 유사한 위치에 배치되었으며, 이는 『상한론』의 증상-본초 관계에 대한 기존 지식과 상당부분 일치하였다. 이를 통해 증상 데이터에서 얻은 PC의 의미가 본초 데이터에서도 적용될 수 있음을 확인하였다.

3. 증상과 본초와의 관계 분석

각 본초와 증상들의 노드(node)로, 본초와 증상의 동시출현(co-occurrence)을 엣지(edge)로 네트워크를 구성하였다. 엣지의 cutoff를 정하기 위해 『상한론』 전체에 대한 증상 및 본초의 빈도, 각 조문이 포함하는 증상 및 본초의 개수를 보존하되, 각 조문에 해당하는 증상 및 본초는 임의로 선택하여 Jaccard score를 구하는 permutation test를 1000배수 진행하였다. Permutation test를 통해 계산된 Jaccard score 중 상위 5%에 해당하는 Jaccard score는 0.09였으며, 이를 cutoff로 정했다. 구성된 네트워크는 전반적으로 끊어지지 않고 이어져 있으며, 증상과 본초가 다대다 관계를 이루고 있다. 이는 하나의 증상에 하나의 본초가 치료약물로 대응되는 단순한 구조가 아닌, 여러 증상의 유무를 동시에 고려하여 본초를 선택하는 복잡한 관계임을 시사한다. 또한 본 시각화를 통해 기존에 알려진 『상한론』에 대한 지식이 네트워크에 반영되어 있음을 확인하였다. 예를 들어 실열증에 청열지제(淸熱之劑)로 활용되는 백호탕(白虎湯)의 구성 본초인 석고-지모-갱미는 근접한 위치에 군집을 이루고 있다. 마찬가지로 오령산(五苓散)의 구성 본초인 저령-택사-복령, 치자시탕(梔子豉湯)의 구성 본초인 치자-담두시, 각종 사하약(瀉下藥)의 구성 본초인 지실-후박-망초-대황이 각각 군집을 이루고 있어 약대에 대한 정보를 반영하고 있음을 확인하였다. 또한 이들이 치료하는 증상들과 직접적으로 엣지를 형성하고 있어, 본초의 적응증에 대한 정보가 네트워크에 반영되어 있음을 시각적으로 확인하였다.

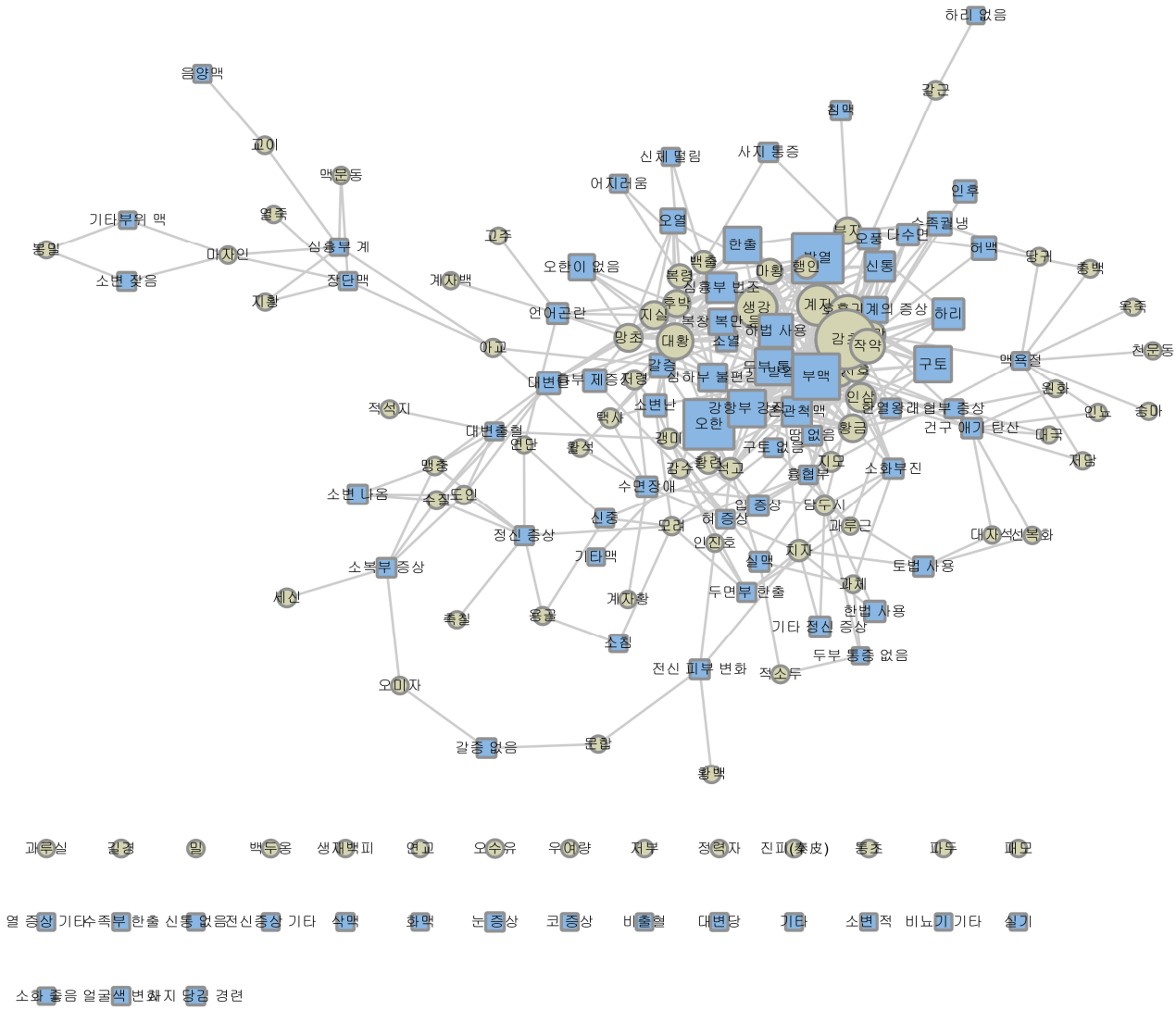

Fig. 3. Visualization of a bipartite network based on Jaccard scores between symptoms and herbs. Blue squares and olive circles indicate symptoms and herbs, respectively. Sizes of nodes indicate frequencies of symptoms or herbs. An edge between nodes indicates a co-occurrence of the symptom and the herb in the same sentences.

Eigenvector centrality 분석을 통해 보편적인 증상과 특이적인 증상을 정량적인 지표를 기준으로 분류하고, 마찬가지로 범 증상에 사용할 수 있는 본초와 특수한 증상에 대해 사용되는 본초를 분류하였다. 발열, 오한 등의 증상은 높은 centrality를 보여 보편적인 증상으로 분류되었는데, 해당 증상들은 태양병의 제강(提綱) 및 상한에 해당되는 증상이다. 또한 갈증이 없거나, 소변이 잘 나오는 상황 등은 낮은 centrality을 나타내고 있어 특이적인 증상으로 분류할 수 있는데, 해당 증상들은 『상한론』에서 예외적인 경우로 나타나는 증상에 해당된다. 본초의 경우 감초, 반하, 인삼, 황금 등이 높은 centrality를 보였으며 이러한 본초는 다양한 증상에 사용되는 본초로 분류될 수 있다. 실제로 감초는 많은 처방에서 화약(和藥)으로 활용되고 반하-인삼-황금은 소시호탕에 포함되는 본초로 임상에서 반표반리증을 포함한 다양한 증상군에 활용되고 있어, 기존의 본초 지식과 일치하는 결과를 보였다. 한편 복령-저령-택사는 『상한론』에서 수습불리(水濕不利)라는 제한된 상황에서만 활용되고 있어 낮은 centrality를 보였다. 『상한론』이 외사에 감응된 뒤 인체의 반응을 기록한 서적이고 외사의 개념이 현대의 외부 항원의 개념과 유사하다는 점을 반영한다면, 증상의 centrality가 높을수록 다양한 감염 상황에서 공통적으로 발생하는 증상일 확률이 높다. 반면 증상의 centrality가 낮을수록 특정 질환에서 나타나는 증상이거나 예외적인 상황일 확률이 높다. 마찬가지로 centrality가 높은 본초는 감염으로 인한 다양한 증상에 보편적으로 활용될 수 있는 약물인 반면, centrality가 낮은 본초는 특정 증상이 있을 경우 활용할 수 있는 약물이라고 추측할 수 있다.

4. 본초별 적응증 탐색

각 본초가 어떤 증상과 높은 상관관계를 갖는지 알아보기 위해 각 본초별로 Jaccard score가 높은 10개의 증상을 시각화하였다(Fig. 4). 계지와 마황은 모두 표사에 감수되었을 때 나타나는 증상에 대해 이용한다는 공통점이 있으나, 계지는 오한과 발열에 연관, 마황은 무한(無汗)과 호흡기계의 증상과 더 연관된다는 차이점이 나타난다. 이를 통해 계지와 마황이 각각 표허증 및 표실증에 이용되는 경향을 확인하였다. 생강과 대조는 상관관계를 가지는 증상이 유사한데, 이는 생강과 대조가 처방을 구성할 때 흔히 "강삼조이(薑三棗二)"로 배합되어 사용되기 때문으로 추측된다. 석고와 지모는 모두 갈증과 구강의 증상(입마름 등)에 사용되고 있는데, 특히 석고는 청열약(淸熱藥)임에도 '발열이 없음'과 높은 연관을 보이고 표사와 관련된 증상들과 연관이 있는 반면 지모는 그렇지 않다는 차이점을 보였다. 시호와 반하는 『상한론』에서 배합되는 경우가 많은데, 시호는 반표반리증, 흉협부 증상과 특히 연관이 있는 반면 반하는 소화기계의 증상과 특히 연관이 있었다. 생강, 인삼, 건강, 부자와 같이 함께 온리약으로 분류되는 본초도 연관이 있는 증상군에 차이가 있었다. 생강은 해표약으로 기능하였으며, 인삼은 구토 및 오한, 발열, 신통과 같은 외증(外證)을 대표하는 증상과 연관이 있었다. 한편 건강과 부자는 하리, 수족궐냉, 허맥과 같은 한증(寒證)을 구성하는 증상과 연관이 높았으며 건강은 주로 소화기계의 증상과 관련이 있었다. 수습에 작용한다고 여겨지는 복령, 택사, 백출은 모두 갈증 및 소변불리와 높은 연관을 보였으며, 그 중 복령과 택사는 갈증과 소변불리(小便不利)에 대해 특이적으로 높은 Jaccard score를 보이는 반면 백출은 보다 다양한 증상에 높은 Jaccard score를 보여 백출이 여러 용도로 활용된다고 추측할 수 있다. 대황, 망초, 후박은 모두 대변난, 오열, 발열의 상황에 사용되고 있으며 특히 대황과 망초는 하법을 사용한 뒤 부작용이 나타날 때에도 다시 사하를 유발시키기 위해 사용한다고 추측된다. 청열해독약(淸熱解毒藥) 중 삼황(三黃)이라고 일컬어지는 황련, 황백, 황금도 그 적응증에 다소 차이가 있었다. 황련은 심하부의 불편감, 황백은 전신 피부 변화(황달)에 특이적으로 연관되는 반면, 황금은 심하부 불편감을 포함하여 상초(上焦)-중초(中焦)의 다양한 증상에 활용되었다. 그러나 본 결과를 통해 『상한론』의 제한적인 인식 또한 확인할 수 있다. 현대에는 인삼이 보비폐기(補脾肺氣)하는 대표적인 온보약(溫補藥)으로 알려져 있지만, 『상한론』에서의 인삼은 허증과 연관성이 부족하며 심비(心痺)에 반하 등과 배합되어 사용되고 있어 본초의 효능에 대한 제한적인 인식 역시 확인할 수 있다.

Table 3. Eigenvector centralities of symptom and herb nodes in the network

a) Symptom nodes

증상	centrality	증상	centrality	증상	centrality	증상	centrality
발열	0.222	한열왕래	0.080	혀 증상	0.020	기타맥	0.003
오한	0.215	호흡기계 증상	0.077	허맥	0.019	두부 통증 없음	0.003
촌관척맥	0.212	소화부진	0.070	실맥	0.018	정신 증상	0.002
두부 통증	0.205	협부 증상	0.067	입 증상	0.017	전신 피부 변화	0.002
경항부 강직	0.200	갈증	0.065	인후	0.015	소복부 증상	0.001
부맥	0.190	건구 애기탄산	0.057	신중	0.012	소변 나옴	0.001
신통	0.175	오열	0.054	수면장애	0.008	소침	0.001
구토	0.164	흉협부	0.048	신체 떨림	0.007	하리 없음	0.001
발열이 없음	0.163	대변난	0.041	어지러움	0.007	심흉부 계	0.001
심하부 불편감	0.145	언어곤란	0.041	두면부 한출	0.007	장단맥	0.001
한출	0.142	오한이 없음	0.040	사지 통증	0.007	갈증 없음	0.000
하법 사용	0.132	수족궐냉	0.039	구토 없음	0.006	기타부위맥	0.000
복창 복만 등	0.129	다수면	0.038	토법 사용	0.005	소변 잦음	0.000
심흉부 번조	0.115	흉부 제증상	0.029	기타 정신증상	0.004	음양맥	0.000
하리	0.089	땀 없음	0.024	한법 사용	0.004		
조열	0.086	소변난	0.022	침맥	0.004		
오풍	0.081	맥욕절	0.021	대변출혈	0.003		

b) Herb nodes

본초	centrality	본초	centrality	본초	centrality	본초	centrality
감초	0.233	복령	0.050	인뇨	0.007	천문동	0.002
반하	0.220	부자	0.045	저담	0.007	적소두	0.002
인삼	0.218	지모	0.045	아교	0.006	용골	0.001
계지	0.204	감수	0.036	활석	0.006	계자황	0.001
황금	0.194	백출	0.032	대자석	0.005	적석지	0.000
대황	0.189	담두시	0.026	선복화	0.005	속칠	0.000
대조	0.189	갱미	0.025	당귀	0.005	문합	0.000
생강	0.189	치자	0.020	연단	0.005	황백	0.000
건강	0.173	저령	0.018	인진호	0.005	마자인	0.000
작약	0.163	택사	0.018	도인	0.004	맥문동	0.000
마황	0.162	황련	0.013	맹충	0.004	지황	0.000
석고	0.129	과루근	0.012	수질	0.004	오미자	0.000
시호	0.121	대극	0.011	계자백	0.004	세신	0.000
행인	0.108	원화	0.011	고주	0.004	교이	0.000
망초	0.107	모려	0.008	총백	0.003	열죽	0.000
지실	0.086	과체	0.008	승마	0.002	봉밀	0.000
후박	0.076	갈근	0.007	옥죽	0.002		

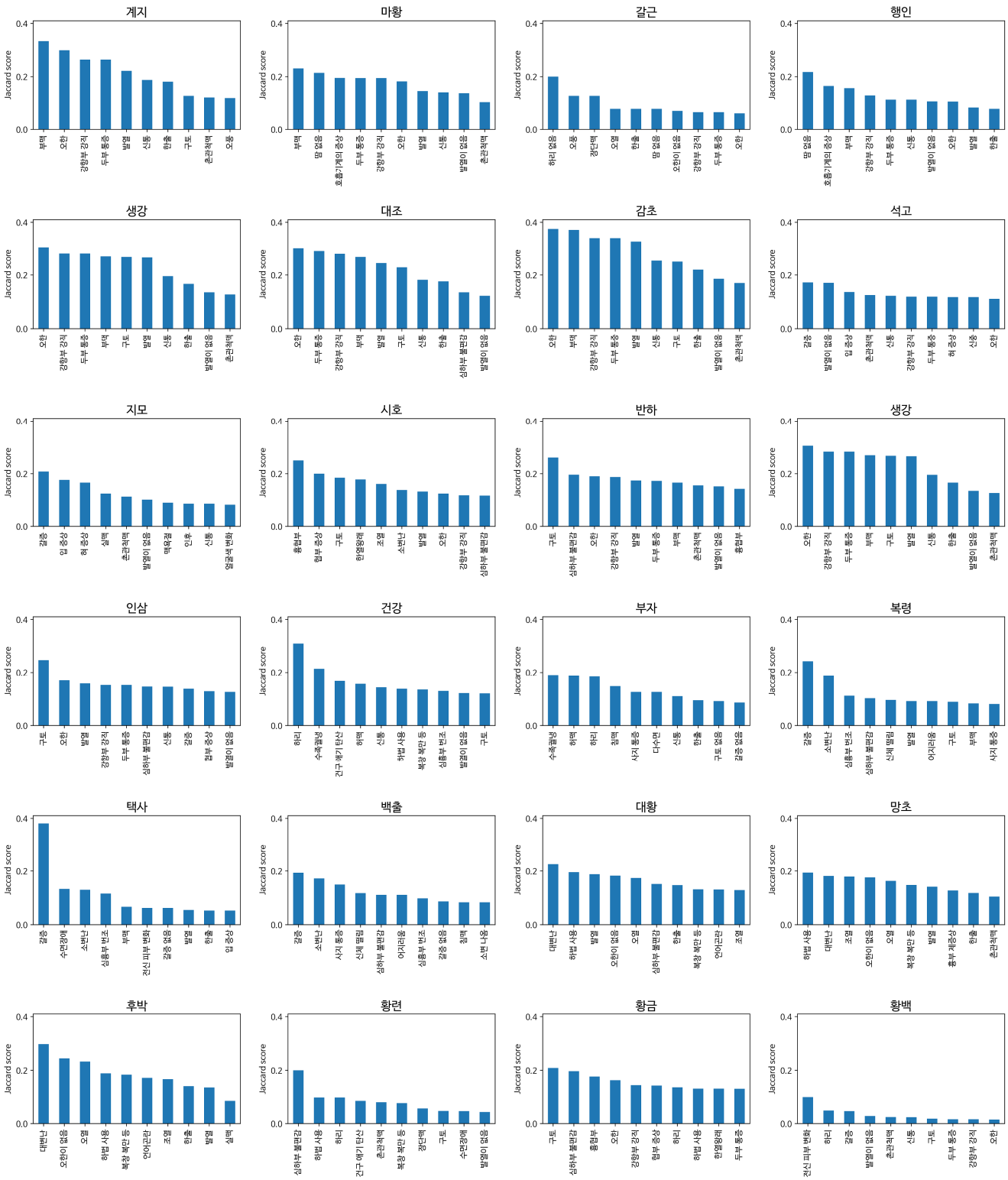

Fig. 4. Symptoms with high co-occurrence for each herb. Co-occurrence is calculated with the Jaccard score.

고 찰

본 연구에서는 『상한론』의 처방 조문에 대한 정량적 접근을 통해 증상과 본초간의 관계를 분석하였다. 한의학의 이론체계는 기혈진액(氣血津液), 오장육부(五臟六腑)와 같이 많은 부분이 정량화 할 수 없는 개념들로 구성되어 있어 수학적인 언어로 구조를 설명하기 어렵다. 그러나 『상한론』은 구체적인 증상을 언급하고 그에 따른 처방을 제시하는 임상서로, 모호한 개념이 상대적으로 적고 직접적인 경험을 바탕으로 서술되어 있어 정량적인 연구방법을 적용하기에 상대적으로 용이하다. 또한 『상한론』은 『황제내경』과 더

불어 한의학의 기반을 이루는 현재까지 임상에 직접적으로 활용되는 문헌으로[3,4] 현대 한의학에서의 중요도가 매우 높다. 따라서 본 연구에서는 『상한론』의 처방 조문에 대해 정량적 연구방법을 적용하여 『상한론』의 증상-본초 구조를 객관화하였다.

문헌 분석에의 텍스트마이닝의 도입은 정량적인 기준을 바탕으로 다량의 문서를 빠른 시간 내에 분석할 수 있게 하며, 그 결과 원전 연구의 객관성을 확보하고 효율성을 높일 수 있다. 본 연구에서는 『상한론』을 대상으로 텍스트마이닝을 진행하였는데, 분석 과정에 있어 연구자의 주관을 최소한으로 개입하면서 『상한론』의 증상-본초 구조를 분석하고 시각화하였다는 특징이 있다. 특히 텍스트마이닝은 인간이 직접 분석하기 어려운 양의 문헌을 분석할 때 더욱 높은 효용을 보인다. 예를 들어 e-commerce의 동향을 파악하기 위해 약 3만여 건의 소셜네트워크 기록을 분석하거나[18], 미국 대통령 선거 후보에 대한 여론을 파악하기 위해 2만여 건의 소셜네트워크 기록을 분석하는 연구는[19] 사람이 직접 분석할 수 없는 양과 복잡도를 텍스트마이닝 분석방법으로 해결한 연구이다. 특히 방대한 연구 동향을 텍스트마이닝을 활용하여 파악하는 연구도 다수 진행되고 있는데, 13만건의 논문을 분석하여 어떤 뇌의 영역이 통증과 연관되어 연구되고 있는지 파악한 연구[20], 1,700만여 건의 문헌에서 노화와 관련된 유전자와 질병을 탐색한 연구[21] 등이 그 예이다. 한의학 또한 본 연구에서 다룬 『상한론』 이외에 다른 한의학 고전이나, 과거에서부터 현재까지 누적된 의안 등 대량의 텍스트에 정량적 분석 방법이 적용된다면 기존 방법론의 한계로 인해 발견하지 못했던 암묵적 지식을 객관화할 수 있으며, 이를 바탕으로 한의학의 새로운 측면을 발견할 수 있을 것이다.

특히 본 연구를 통해 한의학 지식 체계를 정량적으로 설명할 수 있는 가능성을 발견하였다. 예를 들어 TF 및 TF-IDF에 따른 단원별 증상 및 본초의 분류는 각 단원을 대표하는 증상 및 본초를 정량적인 기준으로 제시할 수 있게 한다. 구체적으로 TF가 높은 증상은 해당 단원에서 자주 등장하는 증상으로, 임상에서 자주 발견할 수 있는 증상을 의미한다. TF-IDF가 높은 증상은 자주 등장하면서도 다른 단원에서는 보이지 않는 증상으로, 임상에서 해당 증상을 보일 경우 쉽게 변증할 수 있는 기준이 될 수 있다. 한편 TF가 높은 본초는 임상에서 빈용되는 본초로 이해할 수 있으며, TF-IDF가 높은 본초는 특정 단원의 증상을 치료하기 위해 특이적으로 이용되는 본초로 해석할 수 있다. 또한 본 연구에서 Jaccard score를 통해 각 본초들이 어떤 증상과 상관관계를 갖는지 계산하는 과정은 본초의 적응증을 추론하는 과정으로 이해할 수 있다. 특히 본 연구에서 증상과 본초의 상관관계를 계산하는 과정은 한의학 개념을 최소한으로 이용하여 이루어졌는데, 이와 같은 방식은 요시마스 토도(吉益東洞)와 같이 당시의 한의학 이론들을 대부분 배제하고 적응증을 중심으로 처방을 결정하는 방식과 일면 유사하다[22]. 이러한 연구방법은 기존의 한의학적 지식체계의 도움을 최소화하고 데이터를 통해 한의학의 패턴을 도출할 수 있으며, 이를 통해 신속하고 연구자의 주관을 최소화한 결과를 얻을 수 있다는 장점이 있다. 이러한 데이터과학 연구방법은 기존 한의학 지식의 전산화와 임상현장 등에서의 한의 데이터 증가가 가속화될수록 활용도가 높아질 것으로 예상된다.

본 연구는 한의학 지식체계의 정량성을 확보함으로써 궁극적으로는 한의학을 과학의 언어로 설명하는 가능성을 제시한다. 한의학 지식은 동아시아인의 세계관이 반영된 은유로 표현되어 있어, 한의학 용어들에서 실체적 대응 관계가 불분명하다는 특징이 있으며[10] 한의학 용어의 측정불가성은 한의학 이론의 과학적 연구에 장애물로 작용하고 있다. 한의학의 내용을 정량적으로 설명하는 것은 한의학 이론을 보다 선명하게 이해하고 명료화할 수 있으며, 현대과학과 한의학 고전 이론과의 접점을 넓히는 계기를 제공한다. 특히 한의학 이론이 타 전공자가 이해하기 어려운 언어로 서술되어 있다는 지적이 지속적으로 제기되고 있는데[23], 한의학을 보편적인 학문 언어인 수학으로 서술하는 것은 비전공자가 한의학의 체계를 보다 쉽게 이해할 수 있게 하는 방법이 된다. 또한 본 연구를 통해 밝힌 『상한론』의 수학적 구조는 컴퓨터가 처리할 수 있는 형태의 한의학 진단 시스템을 설계하는 데에 기반이 되는 등 한의학의 IT 응용 연구에 기초자료로 활용될 수 있다.

본 연구에는 크게 네 가지의 한계가 있다. 첫 번째로, 각 단원마다 등장하는 처방의 수가 균등하지 않아 분석결과에 비뚤림이 발생할 수 있다. 특히 처방의 수가 적은 소양병, 태음병 등은 독자의 혼동을 막기 위해 본 분석에 포함되지 않아, 『상한론』의 전체 경향을 온전히 반영하였다고 보기는 어렵다. 본 연구에서는 이러한 한계점을 보완하기 위해 TF와 TF-IDF를 동시에 비교하거나 Jaccard score, eigenvector centrality와 같은 metrics를 사용하는 등 절대빈도를 보정할 수 있는 연구방법론을 활용하였다. 두 번째로, 증상 및 본초의 빈도가 중요도로 해석되어 결과에 비뚤림을 발생시킬 가능성이 있다. 본 분석의 핵심 가정 중 하나는 자주 등장하는 증상 및 본초가 더 중요하다는 것이다. 그러나 이 가정이 항상 옳은 것은 아닌데, 예를 들어 한 증상은 보편적인 증상을 설명하기 위해 등장했고 다른 한 증상은 예외적인 증상을 설명하기 위해 등장했다고 하더라도 등장 횟수가 같으면 같은 중요도로 다루기 때문이다. 이에 대해 본 연구에서는 태양병, 상한과 같이 여러 조문에서 빈용되는 개념의 정의를 사전으로 등록하고, 각 조문별 증상을 집계할 때 이러한 개념을 설명하는 증상들을 포함하도록 하여 중요 개념이 분석에 반영될 수 있도록 하였다. 세 번째로, 증상의 구체도가 낮고 증상의 정도가 반영되지 않았다. 본 연구에서는 유사한 증상을 하나의 그룹으로 묶는 등 증상의 범주를 간략화하는 작업을 거쳤고, 大熱, 微熱 등을 熱 범주에 포괄하는 등 증상의 정도를 반영하지 않았다. 본 연구에서 제시한 석고의 적응증 중 '발열이 없음'은 사실 '無大熱'을 의미하여, 일반적인 석고의 적응증과 일치하지 않는다. 그럼에도 불구하고 분석에 앞서 증상 범주를 지금보다 구체화하는 것은 바람직하지 않다. 본 연구에서는 사람의 주관이 반영된 판단을 최소화하는 방법을 제시하고 있으며, 변수를 세분화할수록 전문가의 주관이 개입될 여지가 많아지기 때문에 변수를 무한정 구체화할 수는 없다. 또한 텍스트의 양이 적은 『상한론』에 텍스트마이닝을 적용하기 위해서는 데이터의 차원을 낮출 필요가 있다. 후속연구에서 본 연구방법을 활용하여 『상한론』에서 보다 구체적인 내용을 분석해야 할 경우, 연구대상이 되는 변수의 수를 제한하는 방법 등을 통해 위의 문제를 우회할 수 있다. 마지막으로, 본 연구에서 드러난 연관관계가 반드시 임상적인 상관성을

의미하는 것은 아니다. 본 연구의 방법에 대해 상술한 한계점으로 인해, 인간이 문서를 읽으며 쉽게 파악할 수 있는 행간의 의미를 텍스트마이닝을 적용한 방법으로는 상대적으로 파악하기 어렵다. 특히 축약적으로 기술된 한의학의 고전문헌은 이와 같은 문제가 상존한다. 따라서 본 논문의 독자는 텍스트마이닝을 통해 파악된 문서의 구조를 이해함과 동시에 『상한론』에 대한 기존의 지식을 반영하여 비판적으로 결과를 해석하여야 한다. 또한 텍스트마이닝을 한의학 연구에 적용하고자 하는 연구자는 단순히 빈도를 기반으로 한 패턴을 포착하는 수준에서 더 나아가, 문맥을 반영함으로써 고차적 유추를 가능케 하는 방법을 개발 및 적용하도록 노력하여야 할 것이다.

결 론

『상한론』의 처방 조문을 대상으로 정량적인 지표를 기준으로 각 단원별 핵심 증상 및 본초를 선정하였다. 태양병 상편은 오한 및 부맥, 태양병 중편은 부맥 및 오한, 태양병 하편은 심하부 불편감 및 오한, 양명병은 발열 및 한출, 소음병은 허맥 및 하리, 궐음병은 하리, 수족궐냉, 구토가 빈발 증상으로 나타났다. 각 단원에서 특이적으로 등장하는 증상은 태양병 상편에서 사지 당김과 경련, 태양병 중편에서 소복부 증상, 태양병 하편에서 갈증 없음, 양명병에서 오열, 소음병에서 다수면, 궐음병에서 하리로 나타났다. 한편 단원별 빈용 본초는 태양병 상편과 중편은 계지탕에 포함된 본초, 태양병 하편은 소시호탕에 포함된 본초, 양명병은 사하지제, 소음병과 궐음병은 온리약으로 나타났다. 각 단원에서 특이적으로 이용하는 본초는 태양병 상편에서 열죽, 태양병 중편에서 용골 및 치자, 태양병 하편에서 감수, 양명병에서 후박, 소음병에서 총백, 궐음병에서 당귀로 나타났다.

증상들에 대한 6차원의 단원 공간에 PCA를 적용하여 단원 간의 구조를 분석하였다. 분산이 가장 큰 첫 번째 PC를 통해 태양병 상편에 등장하는 표허증의 증상과 양명병에서 등장하는 실열증의 증상이 구분되며, 두 번째 PC를 통해 태양병에 등장하는 표사로 인한 증상과 소음병에 등장하는 허증의 증상이 구분된다.

본초별로 증상과의 상관관계를 계산하여 증상-본초 네트워크를 구축하고 분석하였다. 증상과 본초가 다대다 관계를 이루고 있으며, 그 중 발열, 오한, 두부 통증과 같은 증상과 감초, 반하, 인삼과 같은 본초는 높은 centrality를 보여 각각 일반적인 증상과 보편적으로 사용되는 본초로 나타났다. 반면 구토 없음, 토법 사용, 어지러움 등의 증상과 당귀, 도인, 지황, 세신과 같은 본초는 낮은 centrality를 보여, 각각 특수한 증상과 특수한 용법의 본초로 나타났다.

본 연구는 한의학 고전 문헌의 내용을 정량적으로 분석함으로써 한의학을 수학적 언어로 설명할 수 있는 가능성을 제시하였다. 한의학 이론의 정량화는 한의학의 명료성을 보강하고 과학 및 공학을 한의학 이론 연구에 응용할 수 있는 기회를 제공한다.

감사의 글

본 연구는 한국한의학연구원 주요사업 AI 한의사 개발을 위한 임상 빅데이터 수집 및 서비스 플랫폼 구축(KSN1911250)의 지원을 받아 수행되었음.

References

1. Shin SW, Kim JB. Diagnosis system in Sanghanlun. Journal of Pathology in Korean Medicine. 1998;12(1):1-18.
2. Shin HM. A Study on the Dose of Prescription in Shanghanlun. Journal of Korean Medicine. 1999;20(3):3-8.
3. Hong SM, Hur IH, Byun HS, Sim SY, Kim KJ. A Case Study on Atopic Dermatitis with the Treatise on Febrile Diseases. The Journal of Korean Oriental Medical Ophthalmology & Otolaryngology & Dermatology. 2007;20(2):230-9.
4. Oh MS, Jeon TD. A Study on the Significance of Sanghanron Prescription in Traffic Accident Patient. Journal of Oriental Rehabilitation Medicine. 2010;20(1):153-66.
5. Hearst MA. Untangling text data mining. Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics. 1999:3-10.
6. Bae JH, Son JE, Song M. Analysis of twitter for 2012 South Korea presidential election by text mining techniques. Journal of Intelligence and Information Systems. 2013;19(3):141-56.
7. Chou CH, Sinha AP, Zhao H. A text mining approach to Internet abuse detection. Information Systems and e-Business Management. 2008;6(4):419-39.
8. Spasic I, Ananiadou S, McNaught J, Kumar A. Text mining and ontologies in biomedicine: making sense of raw text. Briefings in bioinformatics. 2005;6(3):239-51.
9. Lee CY. Understanding Current Traditional Korean Medicine - Preliminary Study for Discussion on the Identity Issue of TKM. J Physiol & Pathol Korean Med. 2010;24(5):758-69.
10. Lee CY. Discussion on the Issues of the Modernization of the Fundamental Theories and Terms in Korean Medicine. J Physiol & Pathol Korean Med. 2013;27(5):540-52.
11. Lee JH, Kim WY, Oh JH. Study on quantization of Korean medicine terminology concept. J Korean Medical Classics. 2014;27(1):099-109.
12. Song YS, Yang D-h, Park YJ, Park YB. A study of relationship between excrement and materia medica in Bangyakhappyeon based on the data mining analysis. The Journal of the Society of Korean Medicine

Diagnostics. 2012;16(2):33-45.
13. Bae HJ, Kim CE, Lee CY, Shin SW, Kim JH. Investigation of the Possibility of Research on Medical Classics Applying Text Mining-Focusing on the Huangdi's Internal Classic. Journal of Korean Medical classics. 2018;31(4):27-46.
14. Jung WM, Lee T, Lee IS, Kim S, Jang H, Kim SY, et al. Spatial patterns of the indications of acupoints using data mining in classic medical text: a possible visualization of the meridian system. Evidence-Based Complementary and Alternative Medicine. 2015;2015.
15. Oh JH. Can similarities in Medical thought be Quantified? The Journal Of Korean Medical Classics. 2018;31(2):71-82.
16. Kim SW, Kim KW, Lee BW. A study on combination of prescription of Shanghanlun using database. J Korean Medical Classics. 2019;32(1):171-89.
17. Kim SU, Lee HK, Jung HJ. Suggestions for writing the medical records based on the symptoms in Sanghanron. The Journal of the Society of Korean Medicine Diagnostics. 2014;18(2):85-109.
18. Shen CW, Chen M, Wang CC. Analyzing the trend of O2O commerce by bilingual text mining on social media. Computers in Human Behavior. 2019;101:474-83.
19. Wang H, Can D, Kazemzadeh A, Bar F, Narayanan S. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. Proceedings of the ACL 2012 system demonstrations. 2012:115-20.
20. Oh J, Bae H, Kim CE. Construction And Analysis Of The Time-Evolving Pain-Related Brain Network Using Literature Mining. Journal of Pain Research. 2019;12:2891.
21. Kwon Y, Natori Y, Tanokura M. New approach to generating insights for aging research based on literature mining and knowledge integration. PloS one. 2017;12(8).
22. LEE JH, Baik YS, Jeong CH. Yoshimasu Todo's medical theory extracted from Yakjing III. The Journal Of Korean Medical Classics. 2006;19(2):66-73.
23. Yeo IS. Korean medicine, see it as science? Research Institute for Healthcare Policy Korean Medical Association. 2011;9(3):70-5.