大韓本草學會誌 제30권 제3호(2015년 5월)
Kor. J. Herbol. 2015 ; 30(3) : 63-68

ISSN 1229-1765(Print), ISSN 2288-7199(Online)
http://dx.doi.org/10.6116/kjh.2015.30.3.63.

HF-IFF: TF-IDF를 응용한 병증-본초 연관성(relevancy) 측정과 본초 특성의 시각화
- 청강의감 방제를 대상으로 -

오준호*

한국한의학연구원

HF-IFF: Applying TF-IDF to Measure Symptom-Medicinal Herb Relevancy and Visualize Medicinal Herb Characteristics
- Studying Formulations in *Cheongkangeuigam* -

Oh Junho*

Korea Institute of Oriental Medicine

ABSTRACT

Objectives : We applied the term weighting method used in the field of data search to quantify relevancy between symptoms and medicinal herbs, and, based on this, we aim to introduce a method of visualizing the characteristics of medicinal herbs.

Methods : We proposed HF-IFF, an adaptation of TF-IDF, which is a term weighting measurement method adapted in the field of data search. Using this method, we deduced relevancy between symptoms and medicinal herbs In *Cheongkangeuigam* that was published in 1984 by organizing the medical theory of *Cheongkang, Kim Younghoon*, and visualized this as a graph in order to compare the characteristics of medicinal herbs used for different symptoms.

Results : HF-IFF is the product of HF and IFF, where HF is the frequency of the relevant medicinal herb for a set of symptoms, and IFF is the inverse of the number of formulations (FF) containing that herb. A total of 251 types of medicinal herb are used in *Cheongkangeuigam*, and 1538 formulations are classified according to 67 types of symptom. The overall mean for HF-IFF was 0.491, with a maximum of 4.566 and a minimum of 0.013.

Conclusions : In spite of several limitations, we were able to use HF-IFF to measure relevancy between symptoms and medicinal herbs, with formulations as an intermediate. We were able to use the quantified results to visually express the characteristics of the herbs used for symptoms by bubble chart and word-cloud from HF-IFF.

Key words : Medicine, East Asian Traditional, Data Mining, TF-IDF, HF-IFF, Herbal Formula

서 론

한의학에 대한 현대적인 탐구가 활발히 이루어지고 있는 오늘날에도 고문헌은 임상, 교육, 연구 현장에서 중요한 텍스트로 활용되고 있다. 임상 현장에서 사용되고 있는 약재 처방들이나 침구 이론 상당수가 고문헌에 근거하고 있으며, 교육 현장에서도 동의보감 등 고문헌의 내용을 교육 전반에 활용하고 있다. 연구 분야에서는 고문헌의 임상 기록을 객관적으로 확인하거나 새로운 발견을 위해 고문헌에서 영감을 얻으려는 노력이 계속되고 있다.

하지만 고문헌의 양은 실로 방대할 뿐만 아니라 오늘날에는 현대 연구 결과까지 폭발적으로 증가하고 있다. 전통적으로 한의학자들은 한의학 고전의 치료 방법을 임상에 적용하고, 다시 임상에서 얻은 경험을 통해 고전 이해의 폭을 넓히는 방식으로 의학을 연구해 왔다. 그러나 오늘날 현대 연구 결과를 소화하면서 고문헌 이해를 병행해 나가기란 전문가들

*Corresponding author : Oh Junho. Korea Institute of Oriental Medicine
· Tel : +82-42-868-9317 · FAX : +82-42-863-9463 · E-mail : junho@kiom.re.kr
· Received : 10 April 2015 · Revised : 22 May 2015 · Accepted : 25 May 2015

에게 조차 버거운 일이다. 한의학 고문헌의 규모를 그 속에 담긴 방제 수로 환원해 보면, 한의학의 대표적 저작인 동의보감에만도 4천여 개의 방제가 실려 있으며, 동아시아 전체 고문헌으로 확장하면 약 9만여 개의 방제가 존재한다[1]. 서적별로 중복되어 언급된 양을 고려한다면 이보다 몇 배는 될 것으로 추정된다.

고무적인 것은 정보검색(information retrieval) 분야, 데이터 분석(data analysis) 분야 등에서 대량의 정보에 우선순위를 부여하고 내용을 분석하는 여러 가지 방법들을 발전시켜 왔다는 점이다. 이 방법들을 적절히 적용한다면 한의학 고문헌 속에서 핵심적인 지식들을 우선적으로 파악해 내는 것도 가능할 것이다. 최근 현대적 분석 기법을 이용하여 미쳐 인식되지 않았던 고문헌의 특성을 탐색하거나, 암묵적으로 통용되어 왔던 지식들을 검토하는 연구들이 수행 되었다[2-7]. 본 연구에서는 이런 노력의 하나로 정보검색 분야에서 활용되고 있는 용어 가중치(term weighting) 방식을 응용하여 병증과 본초 사이의 연관성(relevancy)을 정량화하고, 이를 바탕으로 병증에 활용된 본초의 특성을 시각화하는 방법을 제안하고자 한다.

재료 및 방법

본 연구에서는 본초와 병증 사이의 연관성을 정량화하기 위해 정보검색 분야에서 용어 가중치 정량화에 사용되는 TF-IDF를 변형한 HF-IFF를 소개하고, 이어 청강의감의 분석 예를 통해 해당 내용을 상술하였다. 이를 위해 먼저 HF-IFF의 모태가 된 TF-IDF를 설명하고, 이어 분석 대상 고문헌으로 청강의감을 선정한 이유를 서술하였다.

1. 용어 가중치 측정과 TF-IDF

용어 가중치(term weighting)는 정보검색(information retrieval) 분야에서 검색어와 연관성이 높은 검색 결과를 도출해 내기 위해 고안된 방법이다. 단순히 한 문서 내에서 용어가 있는지 없는지를 기준으로 검색하면 너무 많거나 너무 적은 결과가 도출된다. 대부분은 결과가 너무 많이 도출되는데, 이러한 경우에는 어떤 문서가 중요한 문서인지 순위(ranking)를 정해 줄 필요가 있다. 이를 위해 검색어와 도출된 문서 사이에 유사도를 계산하게 되는데, 이 과정에서 더 중요한 용어와 덜 중요한 용어를 구분하여 용어 별로 가중치를 부여하면 수월한 유의성 결과를 얻을 수 있다.

용어 가중치 측정에는 여러 가지 방법이 있지만, TF-IDF가 가장 널리 사용되고 있다. 이 방법은 TF와 IDF의 값을 곱한 것으로, 한 문서 내에서 해당 용어가 많을수록, 문서 집합 전체에서는 적게 나타날수록 유의미한 용어라는 '경험적 추론'을 바탕으로 한다. 정보의 기본 단위를 '용어(term)', 용어들이 모여 이루어진 텍스트를 '문서(document)', 문서들의 모임을 '문서집합(collection)'이라고 하였을 때, 하나의 문서 안에 자주 등장하는 용어는 해당 문서 내에서 중요한 의미를 가진다고 볼 수 있다. TF(term frequency)는 바로 문서 내에 출현하는 용어를 하나하나 센 값(빈도)이다. 하지만 빈도가 높은 용어들 중에는 '바로', '이것은', '것이다' 처럼 일반적으로 많이 사용되기는 하지만 구체적인 의미를 담고 있는 않은 단어들도 포함되어 있다. 따라서 단순히 단어의 빈도만 가지고는 용어의 중요성을 확정할 수 없다.

이를 보정하기 위해 특정 용어가 나타난 문서의 개수 DF(document frequency)를 측정한다. '바로', '이것은'과 같은 용어들은 모든 문서들에 나타나기 때문에 DF가 크게 나타난다. 반대로 전문 용어와 같은 것들은 특정 문서에서만 나타나기 때문에 DF 값이 작게 나타난다. 용어의 유의미한 정도와 DF는 반비례 관계에 있다고 할 수 있으므로 DF의 역수값인 IDF(inverse document frequency)를 취하면 해당 용어가 문서집합에서 얼마나 특수한 의미로 사용되었는지 알 수 있다. 실제 수치 연산에서 약간의 변형 공식들이 존재하지만 일반적으로는 다음과 같이 계산한다.

문서집합 내 문서의 개수를 N, 문서 j에서 용어 i가 나타난 빈도를 f_{ij} 문서 내에서 가장 빈도가 높은 용어 k의 빈도를 $max_k\ f_{kj}$라고 하였을 때, 문서 j에 대한 용어 i의 $TFij$ 값은 다음과 같다.

$$TF_{ij} = \frac{f_{ij}}{max_k f_{kj}} \quad (1)$$

TF_{ij}는 기본적으로 문서 내에 용어가 포함된 빈도(횟수)를 의미하지만, 문서마다 크기가 다르기 때문에 일반적으로 최빈도 용어의 빈도로 나누어 정규화(normalization) 시킨다. 용어 i가 포함된 문서의 개수를 이라고 했을 때, 용어 i의 IDFi 값은 다음과 같다.

$$IDF_i = \log\left(\frac{N}{n_i}\right) \quad (2)$$

IDF_i 역시 분석 대상에 따라 문서집합 내 문서 개수가 상이하기 때문에 이를 나누어 상대적인 값을 취한다. 값에 log를 취하여 정규화 한 것은 용어마다 출현 빈도가 고르지 않고 지수적인 차이를 보이기 때문이다. 즉 일반적으로 많이 사용되는 용어들은 빈도가 극적으로 높은데 반해 대다수의 용어들은 1~2번만 출현한다[8-10].

2. 분석 대상 문헌의 선정

『청강의감(晴崗醫鑑)』은 송재(松齋) 이종형(李鍾馨, 1929-2008)이 스승인 청강 김영훈(金永勳, 1882-1974)의 의학사상을 정리하여 펴낸 의학 서적으로, 간행 시기는 1984년이지만 19세기 말에서 20세기 후반까지의 한의학을 살펴 볼 수 있는 매우 중요한 임상서이자 사료이다. 『청강의감』에는 몇 가지 주목할 만한 의학사상이 담겨 있다. 먼저, 책의 병증 분류 방식과 병증 설명 내용에는 서양의학의 유입으로 시작된 20세기 동서의학 연구의 단면이 드러나 있으며, 기혈담울(氣血痰鬱)을 질환의 주요 병리로 여겨 향부자(香附子), 반하(半夏), 복령(茯苓), 당귀(當歸), 천궁(川芎), 작약(芍藥) 등의 약재를 중요하게 사용하였다. 약성이 강렬한 약재보다는 온순한 약재를 선호하였으며, 값싸고 효과 좋은 약재로 서민들의 삶에 보탬이 되고자 했던 흔적들도 보인다[11].

청강의감은 한 개인의 처방 사용 기록으로 일관된 경향을 보이며, 단일 방제에 사용된 본초 개수에 편차가 크지 않고 본

초 종류 역시 광범위하지 않다. 또 대표처방 외에 처방 용례들을 부연하고 있어 중요한 방제 및 본초의 경우 반복적으로 등장한다. 이러한 특징은 빈도를 기반으로 한 분석에 유리하다. 본 연구에서 청강의감을 분석 대상으로 선정한 까닭이다.

여러 방제가 포함된 방제군을 분석하고자 할 때에 본초의 개수가 많은 대방(大方)과 본초의 개수가 적은 소방(小方)이 불규칙하게 섞여 있거나 방제에 사용된 본초의 종류가 너무 많은 경우에는 빈도를 기반으로 데이터를 정리하기 어렵다. 또 본초의 이명들이 다양하게 등장하는 경우에도 이를 정규화하기 힘들다. 여러 시대 여러 의방서에 실려 있는 방제들을 섭렵하고 있는 동의보감이나 방약합편의 분석 연구가 어려운 이유이다.

결 과

본 연구의 목적은 고문헌의 방제를 매개로 본초와 병증 사이의 연관성을 정량화하여 한의학 지식을 시각화하는 한 가지 방법을 제안하는데 있다. 따라서 TF-IDF를 차용한 HF-IFF 자체를 연구 결과로 제시하고, 이어 청강의감의 분석 예를 통해 해당 내용을 상술하였다.

1. HF-IFF를 이용한 병증-본초 연관성(relevancy) 측정 방법

앞서 설명한 TF-IDF의 전제, 즉 경험적 추론은 한의학 방제에서도 유의미하다. 상한(傷寒) 치료를 위해 사용된 방제들에 마황(麻黃)이 자주 나타난다면, 마황은 상한 치료에 중요한 본초라고 할 수 있다. 하지만 생강(生薑)이나 감초(甘草)와 같은 약재들은 상한 치료 방제들에 마황보다 더 자주 나타나지만 상한에 중요한 약재라고 할 수는 없다. 상한 이외의 병증 치료 방제들에도 많이 나타나기 때문이다. 따라서 ①특정 병증에 자주 나타나는 본초일수록, 또 ②모든 병증에 고루 나타나지 않을수록 해당 병증에 더 유의미한 본초라고 할 수 있다. 전자(①)는 TF와, 후자(②)는 IDF와 개념적으로 같다. 따라서 정보검색 분야에서 TF-IDF를 통해 용어의 가중치를 측정하였다면, 한의학에서는 방제를 기반으로 병증별로 사용된 본초의 가중치를 도출해 낼 수 있다. 이렇게 도출된 병증에 대한 본초의 가중치는 병증과 본초 사이의 연관성(relevance)에 대한 조작적 정의로 활용할 수 있다. 병증에 대한 본초의 가중치가 높다는 것은 병증과 본초 사이의 높은 연관성을 의미하기 때문이다.

본고에서는 병증-본초 연관성 측정에 TF-IDF의 방법을 한의학 방제에 맞게 변형한 'HF-IFF'라는 방법을 제안하고자 한다. 기본적인 착안점은 앞에서 살펴본 바와 같이 해당 병증에 많이 사용되었을수록, 전체 방제군에는 적게 사용되었을수록 해당 병증과 본초 사이에 유의미한 관계가 존재한다는 경험적 추론에 기반한다. HF-IFF는 다수의 방제가 병증별로 군집되어 존재할 때 병증별로 사용된 본초에 가중치를 측정하는 방법으로 사용될 수 있다. 이러한 가중치는 병증과 본초 사이에 연관성으로 귀결된다.

HF-IFF는 HF와 IFF의 곱으로 측정된다. HF(herb frequency)는 병증에 사용된 본초의 빈도로, 병증마다 사용된 처방의 개수가 다르기 때문에 수치적으로는 처방에 대한 상대빈도를 사용한다. 방제군 내 포함된 전체 방제의 개수를 N, 병증 s에서 본초 h가 나타난 빈도를 f_{hs}, 병증 s에 사용된 방제의 개수를 f_s라고 하였을 때, 문서 j에 대한 용어 i의 HF_{hs} 값은 다음과 같다.

$$HF_{hs} = \frac{f_{hs}}{f_s} \quad (3)$$

FF(formula frequency)는 방제군 전체에서 해당 본초가 포함된 처방의 개수를 의미하며, 역수를 취한 IFF(inverse formula frequency)는 희소성을 통해 방제군 전체에서 해당 본초의 중요성을 수량화 해 준다. 방제군 전체에서 본초 h를 포함하고 있는 방제의 개수를 n_h이라고 했을 때, 본초 h의 IFF_h 값은 다음과 같다.

$$IFF_h = \log\left(\frac{N}{n_h}\right) \quad (4)$$

IDF와 마찬가지로 IFF를 log scale로 설정한 것은 본초를 포함한 처방 개수를 나타내는 FF 값이 지수적으로 나타나기 때문이다. 즉, 생강, 감초와 같은 본초들은 대다수 방제에 포함되어 있지만, 상당수의 본초들은 처방군 전체에서 단지 1~3회 정도만 나타난다. IFF에 사용된 log scale은 이런 편향을 보정하기 위한 것이다(Fig. 1).

HF-IFF를 활용하기 위해서는 본초들은 방제 내에서 상호작용(상수 상반 등) 없이 독립적으로 사용되었다는 가정이 필요하며, 정량화 수단으로 빈도를 사용하기 때문에 본초의 용량이 아니라 출현 빈도를 데이터로 사용해야 한다. 물론 용량을 HF의 가중치로 활용할 수도 있지만 이에 대해서는 추가적인 탐구가 필요하다.

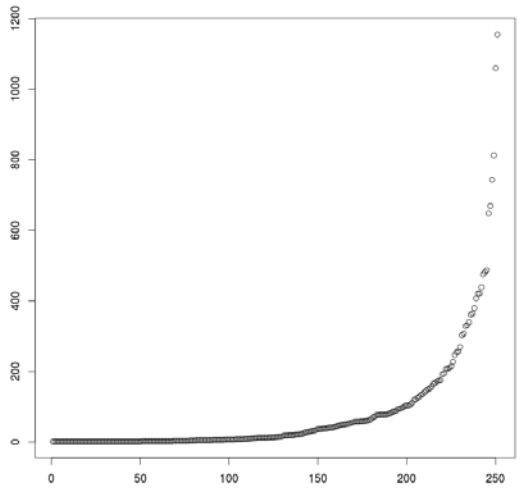

Fig. 1. Formulation frequency(FF) trends for individual herbs. (x-axis: herbs, ordered by frequency. y-axis: formula frequency.)

2. HF-IFF를 이용한 청강의감 병증-본초 연관성 측정

앞서 제안한 HF-IFF의 방법을 통해 청강의감에 대해 병증과 본초 사이의 연관성을 측정하고 정량화된 결과를 바탕으

로 병증별로 사용된 본초 특성을 시각화 하였다. 이를 위해 본초들은 서로 독립적으로 사용되었다고 가정하였고, 용량과 도량형은 무시하고 본초 출현 빈도를 기준으로 처방을 정리하였다.

청강의감에는 방제 구성이 명시된 것을 기준으로 청강 선생이 사용하였던 방제 1538종이 수재되어 있다. 이 방제들은 '일반감모(一般感冒)', '상한장역온병(傷寒瘴疫溫病)', '후풍인종성시(喉風咽腫聲嘶)'에서 '유뇨낭종(遺尿囊腫)', '감병(疳病)', '허약불사식도한위벽오연(虛弱不思食盜汗痿躄五軟)'까지 모두 67종의 병증들에 나누어져 있다. 방제에 사용된 본초는 '가자(訶子)', '생강(生薑)', '갈(葛)', '감국(甘菊)' 등 모두 251종이다.

청강의감 병증-본초 연관성 측정결과를 일반감모(一般感冒)에 사용된 형개(荊芥)와 생강(生薑)의 HF-IFF 도출 예를 통해 살펴보자. 일반감모에는 모두 40종의 방제가 등장하며, 이 가운데 형개는 26번 나타난다. 따라서 일반감모에서 형개의 HF 값은 $\frac{26}{40}$ 즉, 0.65가 된다. 전체 1538종의 방제 가운데 병증과 상관 없이 형개를 포함한 방제는 214개이다. 따라서 형개의 IFF는 $\log(\frac{1538}{214})$ 즉, 1.972가 된다. (소수점 4째자리에서 반올림, 이하 같음) 따라서 일반감모에 사용된 형개의 HF-IFF 값은 이 둘을 곱한 1.282가 된다.

대조적으로 일반감모에 사용된 생강을 보자. 생강은 일반감모에 형개보다 많은 36회 나타난다. 따라서 일반감모에서 생강의 HF 값은 $\frac{36}{40}$ 즉 0.9로 형개보다 크다. 일반감모를 치료하는 90% 처방에 생강이 쓰인 셈이다. 청강의감 전체 방제 가운데 생강을 포함한 방제는 모두 1155개로 생강의 FF는 $\frac{1155}{1538}$ 즉, 0.751이다. 청강의감 전체 방제의 75%에 생강이 포함되어 있다는 의미이다. 일반감모 90%에 비해서는 낮은 편이지만 전반적으로 높은 빈도로 사용되었다. 생강이 일반감모 뿐만 아니라 다른 병증들에도 일반적으로 많이 사용된 본초라는 것을 알 수 있다. IFF는 FF에 역수에 로그값을 취한 것으로 $\log(\frac{1538}{1155})$ 즉, 0.286이 된다. 따라서 일반감모에 사용된 생강의 HF-IFF 값은 형개보다 낮은 0.258이 된다. 단순 빈도에 있어서는 생강이 압도적으로 높지만, 가중치를 측정하면 형개가 더 높게 나타났다. HF-IFF를 병증과 본초 사이의 연관성으로 보았을 때, 일반감모와 형개의 연관성이 생강과의 연관성 보다 상대적으로 높다고 할 수 있다.

청강의감 전체를 HF-IFF 값으로 살펴보자. 청강의감 처방군 전체의 HF 평균(0인 값 제외)은 0.264이다. 1번 이상 등장하는 본초들의 경우, 하나의 본초가 특정 병증에 사용된 처방 가운데 26.4%에 나타난다는 의미가 된다. 물론 이것은 생강이나 감초와 같은 고빈도 약재들 때문에 생긴 결과로, 중앙값을 보면 0.16에 불과하다. 병증 내 상대빈도 최대값은 1, 즉 특정 본초가 모든 방제에 사용된 경우이다. 0을 제외하고 병증 내 상대빈도 가운데 최소값은 0.012이었다. HF-IFF 전체 평균(0인 값 제외)은 0.491, 최대값은 외상어혈(外傷瘀血)에 사용된 소목(蘇木)으로 4.566의 가중치를 보였다. 0을 제외하고 HF-IFF의 최소값은 식체주체서체식독반진채독(食滯酒滯暑滯食毒痰疹菜毒)에 사용된 당귀(當歸)로 0.013을 나타냈다(Fig. 2-3). 이상의 시각화 결과들은 HF-IFF demo page(on-line)[12]에서 전체를 확인해 볼 수 있다.

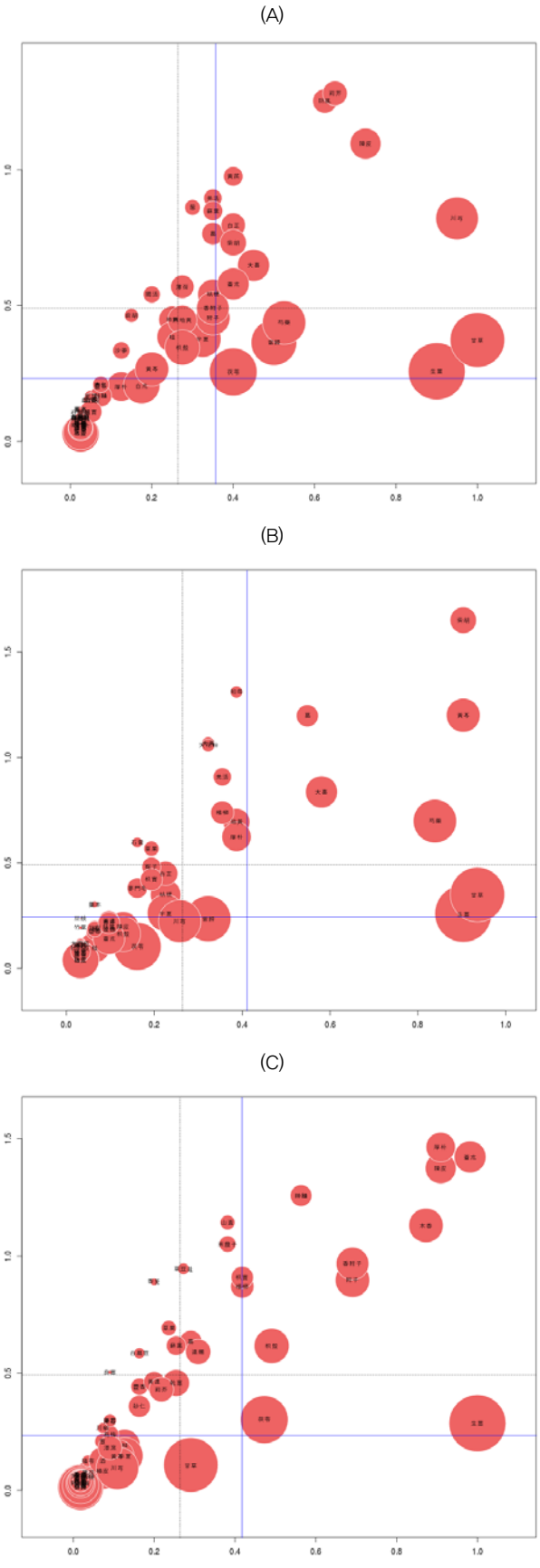

Fig. 2. Characteristics of medicinal herbs used for (A)general cold, (B) epidemic febrile disease, (C) indigestion. (x-axis: HF, y-axis: HF-IFF, circle size: FF.)

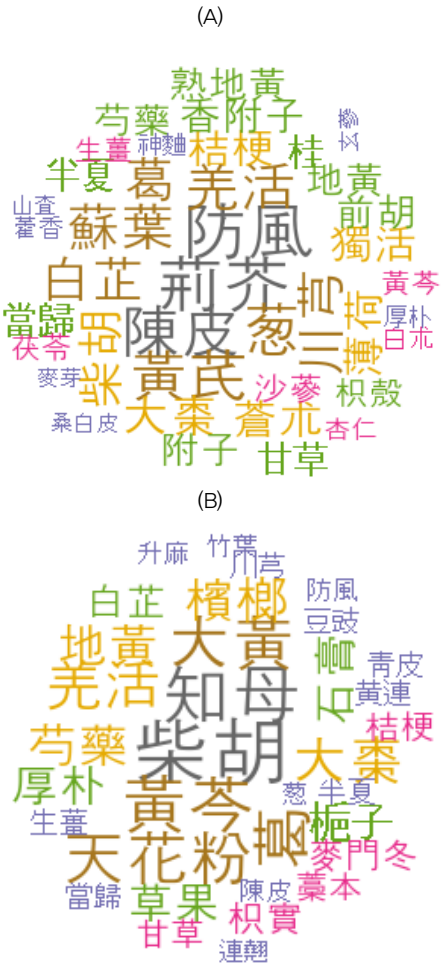

Fig. 3. Word cloud to show the characteristics of medicinal herbs for (A)general cold, (B) epidemic febrile disease.

고 찰

HF-IFF를 통해 도출된 결과는 한의학 이론을 비교적 잘 드러내 주고 있다. 병증별 본초 특성을 보여주는 시각화 결과를 보자. 가로축은 HF, 세로축은 HF-IFF이며, 원의 크기는 FF 즉 방제군 전체에서 해당 본초가 사용된 빈도이다. 생강이나 감초에서 알 수 있듯이 원이 클수록 방제군 전체에서 빈번하게 사용된 본초를 의미한다. 점선은 방제군 전체의 평균값을, 실선은 해당 병증에서의 평균값을 각각 나타낸다(HF-IFF demo page[12] 참조).

시각화 결과는 병증마다 상이하지만, Fig 2a, 2b와 같이 대체적으로 상단(HR, high relevance group), 좌측하단(LRl, low relevance group at the left side), 우측하단(LRr, low relevance group at the right side)을 꼭지점으로 하는 삼각형 형태로 본초들이 분포한다. Y축을 기준으로, 상단은 병증과의 높은 연관성을, 하단은 낮은 연관성을 나타낸다. 상단은 하나의 집단(HR)으로 수렴되는 반면, 하단은 좌측(LRl)과 우측(LRr)의 두 집단으로 나뉘는 경향을 보인다. 일반감모의 경우, 형개 방풍과 같은 본초들이 HR에, 작약 당귀와 같은 본초들이 LRl에, 생강 감초와 같은 본초들이 LRr에 나타났다. 하단의 값들은 모두 병증과 연관성이 적다고 할 수 있지만 함의는 다르다. 좌측하단에 위치한 값들(LRl)은 HF와 HF-IFF 값이 동시에 작은 본초들로서, 빈도와 가중치 두 측면 모두에서 병증과의 연관성이 적다. 반면 우측하단에 위치한 값들(LRr)은 생강이나 감초와 같이 HF는 크지만 HF-IFF는 작다. 병증 내 방제들에 높은 빈도로 사용되었고 처방군 전체에서도 높은 빈도로 사용된 본초들이다.

상단에 위치한 값들(HR)은 HF-IFF가 높은 본초들로서 병증과 연관성이 높다고 할 수 있다. 상단 꼭지점은 보통 중앙부에 위치하지만 Fig 2c와 같이 오른쪽으로 치우친 경우도 있었다. 이런 경우는 HF가 상대적으로 높게 나타난 경우로서 해당 병증 방제 대다수에는 사용되었지만 다른 병증에는 잘 사용되지 않은, 병증과의 연관성이 특히 더 높은 본초들이라고 할 수 있다. Fig 2c의 창출, 후박, 진피와 같은 본초들이 이에 해당한다. 이처럼 시각화 결과 상단 꼭지점의 치우친 정도를 통해 본초들이 병증에 얼마나 독점적으로 사용되었는지 확인할 수 있다.

이러한 시각화는 병증별 본초 특성을 비교하는 수단으로 활용될 수 있다. Fig 2a, 2b를 통해 일반감모와 상한온역온병에 사용된 본초 특성을 비교해 보면, 일반감모에서 HR 그룹은 형개 방풍 진피 등인데 반해, 상한온역온병에서의 HR그룹은 시호 황금 갈근 지모 등이었다. 외감질환으로 나타난 병증들이지만 치료의 접근은 상이했다는 점을 알 수 있다. 또한 상한온역온병에서 HF-IFF 값이 가장 높은 시호는 1.5 이상인데 반해 일반감모에서 가장 높은 형개는 1.5 미만으로 HR에 해당하는 본초들 사이에서도 상대적인 비교가 가능했다.

Fig 3a, 3b과 같이 HF-IFF를 기준으로 word cloud를 그려보면 사용된 본초들의 특성을 더 직관적으로 알 수 있다. HF-IFF이 높은 HR 그룹 본초들이 가운데 크게 위치하는 것을 볼 수 있다.

결 론

본 연구는 정보 검색 분야에서 용어 가중치 측정을 위해 고안된 TF-IDF를 응용하여 한의학 처방을 매개로 병증과 본초 사이의 연관성 측정을 시도하였고, 다음의 결론을 얻을 수 있었다.

1. 정보 검색 분야에서 고안된 용어 가중치 측정 방법 TF-IDF는 빈번히 나오는 용어일수록, 특정 문서에 독점적으로 등장하는 용어일수록 더 중요하다는 경험적인 추론에 근거하고 있다. 이 추론은 한의학의 병증과 본초를 이해하는 데에도 유효하다. 동일 병증에 특정 본초가 많이 사용되었다면 해당 본초는 병증과 연관성이 깊다고 할 수 있으며, 다른 병증에 비교적 적게 사용되면서 해당 병증에만 편중되어 활용되었다면 해당 병증과 특정 본초의 연관성은 크다고 할 수 있다.

2. 이런 유사성을 근거로 본 연구에서는 TF-IDF를 변형한 HF-IFF를 제안하고, 이를 통해 청강의감의 병증-본초 연관성을 도출해 보았다. 그 결과 HF-IFF 전체 평균(0인 값 제외)은 0.491, 최대값은 '외상어혈(外傷瘀血)'에 사용된 '소목(蘇木)'으로 4.566의 가중치를 보였다. 0을 제외하고 HF-IFF의 최소값은 '식체주체서체식

독반진채독(食滯酒滯暑滯食毒痰疹菜毒)'에 사용된 '당귀(當歸)'로 0.013을 나타냈다.

3. HF-IFF로 도출된 정량화 결과는 병증별로 사용된 본초의 특성을 비교 가능하게 한다. 본 연구에서는 병증별로 사용된 본초의 현황을 전문적으로 살펴보기 위한 버블차트(bubble chart)와 직관적으로 개술하기 위한 워드클라우드(word cloud)로 청강의감의 병증별 본초 사용 특성을 시각화 해 볼 수 있었다.

비록 경험적 추론을 전제로 수행되었고 방제에 사용된 본초의 용량이나 본초 상호 관계를 고려하지 못한 한계를 지니고 있지만, HF-IFF를 통해 방제를 매개로 병증과 본초 사이의 연관성을 측정할 수 있으며, 정량화된 결과를 통해 병증에 사용된 본초의 특성을 시각적으로 표현할 수 있었다. HF-IFF는 서적 내에서 병증 간에 본초 사용 특성을 비교할 수 있게 할 뿐만 아니라 동일 병증에 대해 서적간 본초 사용 특성을 비교하여 의학적 특성을 질적으로 판단하는 데에도 활용할 수 있을 것이다. 나아가 한의학 처방이 대량으로 축적된다면 병증과 본초 사이의 연관성을 알아내는 기법으로도 사용할 수 있을 것으로 보인다.

감사의글

본 연구는 한국한의학연구원 주요사업 "모노그래프 고문헌 DB 구축(K154311)"의 지원을 받아 수행되었습니다.

References

1. Peng W. Dictionary of Chinese medicine prescription. 1st ed. Beijing : People's Medical Publishing House. 2005 : 3-4.
2. Baek JU, Lee BW. Extended indications of Four-Constitution Medicinal formula analyzing composition on Dongeuibogam formula Ⅱ. J Korean Med Hist. 2013 ; 26(2) : 23-9.
3. Lee BW, Baek JU. Extended indications of Four-Constitution Medicinal formula analyzing composition on Dongeuibogam formula-The case of Bojungyikgi-tang for So-Eum type-. J Korean Med Class. 2013 ; 26(3) : 99-109.
4. Lee JH, Kim WY, Oh JH. Study on quantization of Korean medicine terminology concept-for disease symptom terms of Compilation of Formulas and Medicinals Addendum-. J Korean Med Class. 2014 ; 27(1) : 99-109.
5. Wu YH, Kim KW, Lee BW, Kim EH. Analysis of Prescriptions from Taepyeonghyeminhwajegukbang, Somunsunmyungronbang and Nansilbijang. J Korean Med Class. 2014 ; 27(4) : 121-31.
6. Kim KW, Kim TY, Lee BW. Prescriptions from Taepyeonghyeminhwajegukbang, Somunsunmyungronbang and Nansilbijang based on Herb weight ratio grade. J Korean Med Class. 2014 ; 27(4) : 73-84.
7. Yang DH. Data mining analysis on relationship between disease pattern and materia medica in Bangyakhappyeon. Seoul : Kyunghee Univ. 2011 : 1-77.
8. Aizawa A. An information-theoretic perspective of tf-idf measures. Inf Process Manag. 2003 ; 39(1) : 45-65.
9. Wu HC, Luk RWP, Wong KF, Kwok KL. Interpreting TF-IDF Term Weights As Making Relevance Decisions. ACM Trans Inf Syst. 2008 ; 26(3) : 13:1-13:37.
10. Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman. Mining of Massive Datasets. Mining of Massive Datasets. New York : Cambridge University Press. 2011 : 7-9.
11. Oh JH. A Study on the Organization and Contents of "CheongKangEuiGam". J Korean Med Hist. 2014 ; 27(2) : 63-74.
12. Oh JH. HF-IFF demo page. Retrieved Mar 14 2015, from http://pinedance.github.io/demo.html#/HF-IFF