대한한의학원전학회지
2021년 2월 34권 1호

pISSN 1229-8328 eISSN 2713-850X
J Korean Med Classics 2021:34(1) 037-045
https://doi.org/10.14369/jkmc.2021.34.1.037

JKMC
Journal of Korean Medical Classics

한의학 고문헌 텍스트에서의 인용문 추정과 탐색

[1]한국한의학연구원 책임연구원
오준호[1]*

Detecting Local Text Reuse in the Texts of East Asian Traditional Medicine

Oh Junho[1]*

[1]Senior Researcher. Korea Institute of Oriental Medicine

Objectives : The purpose of this paper was to examine quantitative methods for estimating and detecting local text reuse in the texts of East Asian Traditional Medicine.

Methods : We introduce techniques that estimate the volume of local text reuse with n-gram and those that directly detect the reuse with the Smith-Waterman algorithm (SW algorithm). Based on this, the estimation and detection of local text reuse were carried out for 『Donguibogam』 and 『Huangdineijing·Suwen』.

Results : Estimates with n-gram had more errors than methods with SW algorithms. SW algorithms detected suspected strings directly with local text reuse, resulting in more accurate results.

Conclusions : Although n-gram does not accurately find local text reuse, its high speed makes it a preferable method for certain purposes, such as screening similar documents. On the other hand, SW algorithms have the advantage of being relatively good at finding similar phrases suspected as local text reuse even if the strings do not completely match. However, due to its excessive consumption of time and computing resources, its benefits are limited to cases where precise results are required.

Key words : Text Reuse, n-gram, Smith-Waterman Algorithm, Korean Medical Classics, East Asian Traditional Medicine.

* Corresponding Author : Oh Junho
Senior Researcher. Korea Institute of Oriental Medicine. 1672 Yuseong-daero, Yuseong-gu, Daejeon, 34054
Tel +82-42-868-9317, E-mail : junho@kiom.re.kr
Received(January 20, 2021), Revised(February 3, 2021), Accepted(February 3, 2021)

I. 서론

동아시아 전통사회에서의 '인용(引用)'은 다른 문헌에 이미 기술된 내용을 자신의 글 속에 사용하는 글쓰기의 한 가지 기법이었다.[1] 한문으로 저술된 동아시아 고문헌에서 광범위하게 나타나는 이 기법은, 권위를 가진 선대 문헌의 내용을 인용함으로써 저자 자신의 지적 역량을 드러내는 한편, 자신이 말하고자 하는 내용의 정당성을 확보하기 위해 빈번히 사용되었다.[2]

동아시아 전통사회에서 저술된 한의학 고문헌에서도 인용은 중요한 서술 방식이었다. 특히 인용은 한의학 고문헌 편찬 과정에서 빼놓을 수 없는 방법이었다. 새로 저술된 서적이라 하더라도 기존 의학 이론과 치법에 기반을 두지 않을 수 없었기 때문에 이전 문헌의 의학 지식을 거의 그대로 전사하는 경우가 많았다. 이 때문에 책의 모두(冒頭)에 인용 서목을 직접 밝히는 경우가 적지 않았으며, 채록한 인용문 각각에 관련된 출처를 병기하는 경우도 보인다. 이처럼 한의학 고문헌은 기존 지식을 담고 있는 '인용문'과 새로운 견해를 제시한 '창작문'을 씨줄과 날줄로 하는 하나의 직조물에 비유할 수 있다.

이러한 이유로 오늘날 한의학 고문헌을 분석할 때에도 대상 문헌이 사용한 인용 서적과 인용문을 분석하는 것은 주요한 연구 방법이다. 서적에 대한 초기 연구는 대체로 여기에서 시작되는 경우가 많다. 예를 들어 조선 시대 주요 의서인 『향약집성방』[3], 『의방유취』[4], 『동의보감』[5], 『제중신편』[6], 『동의수세보원』[7], 『방약합편』[8] 등에 대한 초기 현대 연구들은 모두 이들 서적의 인용서적과 인용문을 정리하거나 분석하는 작업에서 출발하였다.

인용서적과 인용문을 분석하는 일은 고문헌 연구에서 필요한 과정이지만, 상당히 많은 연구자원이 소모된다. 따라서 대부분의 연구들은 인용문 자체를 직접 비교하기 보다는 인용문에 병기된 인용 서적의 명칭을 근거로 분석을 진행하는 경우가 많았다. 그러나 박상영의 지적처럼[9], 병기된 인용 서적을 그대로 분석하다보면 잘못된 결론에 다다르기도 하기 때문에 텍스트 자체를 비교하는 것이 보다 근본적인 방법이 된다.

이러한 문제에 착안하여 본고에서는 한의학 고문헌 텍스트를 상호 비교하여 인용문의 분량을 추정하고, 더 나아가 인용문으로 의심되는 텍스트의 범위를 직접 탐색하는 계량적인 방법을 설명하고자 한다.[10] 본고에서 소개한 방법은 이미 다른 분야에서 사용 중인 방법을 한의학 고문헌에 적용한 것으로, 적절히 활용한다면 연구 자원을 절약하는 이점을 얻을 수 있을

1) 현대의 '인용'은 출처를 밝히는 경우에 한정되며, 그렇지 않은 경우 표절로 취급된다. 그러나 본고에서 다루고 있는 동아시아 전통사회에서의 인용은 이러한 기준을 그대로 적용할 수 없다. 따라서 현대의 인용 개념과는 다소 차이가 있다.

2) 이와 더불어 사상적인 이유도 있었다. 조선 중기 고문가(古文家)들은 글을 지을 때 유가(儒家)의 경전(經典)을 전고(典故)로 삼음으로써 현실에서 쇠락한 유도(儒道)를 확립하고자 하였다. 박영호 저, 김도련 편. 한국 고문의 이론과 전개(조선 중기 고문론 연구). 경기도. 태학사. 1998. pp.103-104.

3) 관련 연구는 다음과 같다. 이하 같음. 김남일. 『鄕藥集成方』의 인용문헌에 대한 연구. 진단학보. 1999. 87. 김중권. 『鄕藥集成方』의 引用文獻 分析. 書誌學硏究. 2006. 35.

4) 최환수, 신순식. 의방유취의 인용서에 관한 연구(1). 한국한의학연구원논문집. 1997. 3(1). 안상우, 김남일. 『醫方類聚』 總論의 體制와 引用方式 分析. 경희한의대논문집. 1999. 22(1).

5) 김중권. 『동의보감』의 문헌적 연구: 인용문헌을 중심으로. 서지학연구. 1995. 11.

6) 지창영. 제중신편의 인용방식에 대한 연구. 한국의사학회지. 2008. 21(1). 이정화. 제중신편의 인용문헌 연구. 서지학보. 2010. 35.

7) 박성식, 송일병. 사상의학의 의학적 연원과 이제마 의학사상에 대한 연구: 동의수세보원 인용문을 중심으로. 1993. 5(1). 이필우, 윤창렬. 동의수세보원 인용문에 대한 연구. 한의학연구소 논문집. 2004. 12(2).

8) 윤용갑, 강순수. 방약합편에 수록된 처방의 주치별 계통분류와 인용문헌에 대한 고찰. 원광한의대논문집. 1986. 4.

9) 고문헌에 병기된 인용서는 현대 문헌에서 참고한 자료를 밝히려는 목적보다는 해당 지식의 기원이 어디인가를 적시하는 것을 목적으로 하고 있다. 그러므로 병기된 서적이 직접 인용처가 아니라 간접 인용처인 경우가 많다. 예를 들어 『인제지(仁濟志)』의 경우, 인용문에 매우 다양한 인용서가 병기 되었음에도 불구하고 사실상 『동의보감』과 『본초강목』의 내용을 모태로 하고 있다. (박상영. 『인제지』의 조선후기 의사학적 위상과 의의: 미키 사카에의 재인용[孫引] 지적과 학술가치 평가에 대한 재검토. 한국실학연구. 2013. 25. pp.531-575.)

10) 이에 대한 선행연구는 적은 편이지만 Donald Sturgeon의 깊이 있는 연구가 있다. Donald Sturgeon. Unsupervised identification of text reuse in early Chinese literature. Digital Scholarship in the Humanities. 2018. 33(3).

것이다. 본고에서는 방법 소개와 함께 활용 예시의 하나로 『동의보감』에 인용된 『황제내경소문』의 인용문을 찾아내고 그 결과를 아울러 보고하고자 한다.

Ⅱ. 본론

1. 연구방법

인용문을 찾는 과정은 연속하여 나열된 글자(string)에서 비슷한 부분(substring)을 찾는 문제로 환원할 수 있다. 이러한 방법은 인터넷 상에 중복된 문서를 찾아 검색 효율을 높이거나, 문서 사이의 유사성을 통해 표절 여부를 검토하기 위해 주로 자연어처리(NLP)나 정보학(information retrieval) 분야에서 연구되어 왔다.[11] 본고에서는 인용문의 분량을 추정하는 방법과, 실제로 인용문으로 의심되는 텍스트를 탐색하는 방법 2가지를 나누어 살펴보고자 한다. 전자는 n-gram 기반의 유사성 검토를 통해서, 그리고 후자는 Smith-Waterman 알고리즘을 이용하여 수행 가능하다.

2. 인용문의 추정

인용문(text reuse)에 대한 선행 연구들은 주로 인용문의 분량을 측정하는 측면에서 수행되었다. 이는 인용문 찾기의 목적이 인용문을 특정하기보다는 온라인에서 중복된 문서를 탐지하여 검색 엔진의 효율성을 높이거나, 텍스트의 표절 여부를 찾아내는 것을 목적으로 하였기 때문이다. 이를 위해서는 문서의 유사성을 측정하는 것만으로 충분하다. 이는 비교하고자 하는 텍스트가 얼마나 동일한가를 통해 추정할 수 있다.[12]

문서를 상호 비교하기 위해서는 비교 단위의 설정이 필요하다. 일반적인 언어의 경우 문단이나 문장이 그러한 단위가 될 수 있다. 이는 문단이나 문장이 이미 구조적으로 의미의 경계를 나타내 주고 있기 때문이다. 만약 비교 단위를 문장으로 설정하였다면, 문장의 유사도(similarity)를 측정하여 전체 문서에서 유사한 문장의 비율을 계산함으로써 문서 전체가 얼마나 유사한지 알아낼 수 있다.

그러나 한의학 고문헌을 포함한 고대 한문 텍스트에서는 이러한 방법을 사용할 수 없다. 문장뿐만 아니라 문단에 대한 구분이 명확하지 않기 때문이다. 현대에는 고대 한문 텍스트를 정리할 때 표점을 추가하여 의미 파악을 돕고 있지만, 동일한 텍스트라고 하더라도 표점의 결과는 사람에 따라 달라질 수 있기 때문에 이를 기준으로 삼기 어렵다.[13]

이 경우 n-gram을 이용한 방법을 적용할 수 있다. n-gram은 하나의 텍스트를 n개의 글자(혹은 단어) 단위로 나누는 방법이다. 예를 들어 "高者抑之下者擧之"를 3-gram으로 표현하면, ['高者抑', '者抑之', '抑之下', '之下者', '下者擧', '者擧之']와 같이 나타낼 수 있다. 이렇게 비교하고자 하는 한의학 고문헌 텍스트를 n-gram으로 분해한 다음, 양자에 모두 포함된 n-gram이 얼마나 존재하는가, 즉 양자 사이에 얼마나 동일한 n-gram을 공유하고 있는가로 인용된 텍스트의 양을 대략 추정할 수 있다. 더 정밀한 추정을 위해 n-gram마다 가중치를 달리하는 방법도 적용 가능하다.

이를 공식으로 표현하면 아래와 같다. 문서a와 문서b가 있다고 하였을 때, S_a와 S_b는 문서a와 문서b 각각의 n-gram 집합이고, $C()$는 집합 요소의 개수를 의미한다. 목적에 따라 여러 가지 공식을 응용할 수 있다.

$$\frac{C(S_a \cap S_b)}{C(S_a)}, \frac{C(S_a \cap S_b)}{C(S_b)}, \frac{C(S_a \cap S_b)}{C(S_a \cup S_b)}$$

이 방법은 n-gram 사이에 글자 중복으로 인해 비교해야 할 데이터의 양이 많아진다는 단점을 가지

11) D. A. Smith, R. Cordel, E. M. Dillon, N. Stramp and J. Wilkerson. Detecting and modeling local text reuse. IEEE/ACM Joint Conference on Digital Libraries. 2014. p.183.

12) Jangwon Seo , W. Bruce Croft. Local text reuse detection. Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2008. pp.3-4.

13) Donald Sturgeon. Unsupervised identification of text reuse in early Chinese literature. Digital Scholarship in the Humanities. 2018. 33(3). pp.671-672.

고 있지만, 텍스트가 문단이나 문장으로 구분되지 않는 경우에도 사용할 수 있으며, 더 나아가 언어와 문자가 가지는 특성에 관계없이 보편적으로 사용할 수 있다는 장점을 가지고 있다.

3. 인용문의 탐색

인용문의 양을 추정하는 문제에서 더 나아가 실제 인용문으로 의심되는 텍스트를 찾는 문제를 생각해 보자. 이 문제 역시 비교 단위를 문단이나 문장으로 정할 수 있는 경우에는 비교적 간단한 문제가 된다. 낱낱의 비교 단위를 각각 짝지어 비교하고, 유사성이 높은 비교 단위를 인용문으로 의심된다고 결론지으면 되기 때문이다.

그러나 한문과 같이 비교 단위가 명확하지 않은 경우에는 인용문으로 의심되는 텍스트의 시작점과 끝점을 파악해야하기 때문에 앞의 방법을 사용할 수 없으며, 전혀 다른 접근 방법이 요구된다. 게다가 한자는 다양한 이체자(異體字)가 존재하며, 인용할 때 원문을 그대로 초록하기도 하지만 때때로 글자를 생략하거나 추가하거나 수정하는 등 변형을 가하는 경우가 많다. 게다가 전사의 오류가 발생하는 경우도 적지 않다. 따라서 텍스트 사이에서 단순히 일치하는 문자열을 찾는 것만으로는 부족하며, 텍스트의 변형을 고려하여 상호 비교를 할 수 있어야 한다. 이러한 문제를 해결하기 위해 Smith-Waterman 알고리즘(이하 SW알고리즘)을 적용할 수 있다.

SW알고리즘은 1981년 T.Smith와 M.Waterman에 의해 개발되었다.[14] 먼저 ①문자열 a와 b에 대하여 점수 행렬 H를 초기화 한다. 다음으로 ②동적 프로그래밍 기법(Dynamic Programming)에 따라 점수를 계산하여 행렬 H를 갱신한다. ③마지막으로 추적 분석을 통해 가장 적합한 부분의 문자열을 찾아낸다.

동적 프로그래밍에서 점수 행렬 H에 대한 공식은 다음과 같다. 여기서 $s(a_i,b_j)$는 a_i와 b_j가 일치하였을 때의 상점(match score)이고, μ는 글자 사이에 빈 공간이나 다른 글자가 있을 때의 벌점(gap cost)이다.[15][16]

$$H_{i,j}=\max\begin{cases}H_{i-1,j-1}+s(a_i,b_j),\\ H_{i,j-1}-\mu,\\ H_{i-1,j}-\mu,\\ 0\end{cases}$$

예를 들어 텍스트a가 "黃帝問曰, 肺之令人咳, 何也。岐伯對曰, 五藏六府, 皆令人咳, 非獨肺也。"[17] 이고, 텍스트b가 "經曰, 五藏六府, 皆能使人咳, 非獨肺也。各以其時主之, 而受病焉。"[18]이라고 하자. 이때 상점을 3점, 벌점을 2점으로 하였을 때 아래와 같은 점수 행렬을 얻을 수 있다(그림 1 참조). 이 점수 행렬에서 가장 최고점은 25점이며, 25점을 끝점으로 하여 인접한 점수 가운데 최고점을 역추적하면 시작점까지의 경로를 알아낼 수 있다(그림 1 붉은색 바탕 부분). 이를 통해 텍스트a의 "曰, 五藏六府, 皆令人咳, 非獨肺也"(13자)와 텍스트b의 "曰, 五藏六府, 皆能使人咳, 非獨肺也"(14자)가 서로 인용 관계에 있다는 결론에 이르게 된다.

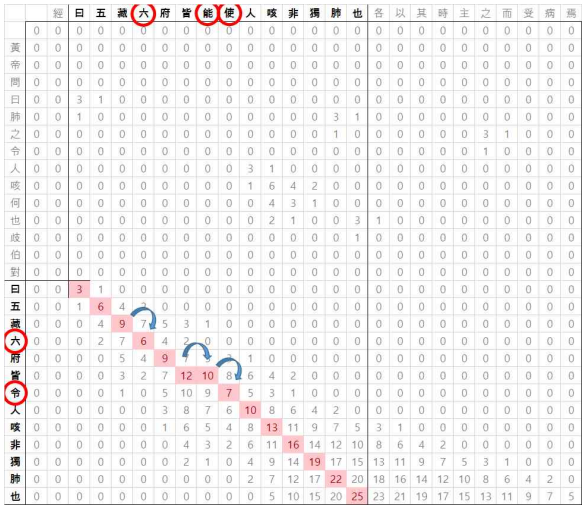

그림 1. 점수 행렬 H의 예시 (붉은색 바탕: 추적 분석 경로, 붉은색 동그라미: 텍스트에서 차이가 있는 영역, 파란색 화살표: 텍스트 차이로 점수가 줄어든 부분)

14) Smith, Temple F., and Michael S. Waterman. Identification of common molecular subsequences. Journal of molecular biology. 1981. 147(1). pp.195-197.

15) 정요상, 이명호, 최동훈. Smith-Waterman 알고리즘을 위한 GPU 상에서의 Multi-Stream 기반 병렬화. 한국정보과학회 2014년 동계학술발표회 논문집. 2014. pp.57-58.

16) 이계성. 다중 지역 정렬 알고리즘. 문화기술의 융합. 2019. 5(3). pp.340-341.

17) 『황제내경소문(黃帝內經素問)』 〈해론편(玆論篇)〉

18) 『동의보감(東醫寶鑑)·잡병편(雜病篇)』 〈해수(咳嗽)〉

이 텍스트 사이에는 2가지 차이점이 존재한다. 첫 번째는 텍스트a의 '六'과 텍스트b의 '六'이다. 전자와 후자는 육안으로 동일하게 보이지만 한자의 다중코드자 가운데 하나로 유니코드로 표시하면 각각 "F9D1"과 "516D"이다. 따라서 컴퓨터는 서로 다른 글자로 인식한다. 이는 한문 텍스트의 전산화 과정에서 자주 발생하는 문제이다. 두 번째는 텍스트a에서 '令'인 부분이 텍스트b에서는 '能使'으로 되어 있다. 이는 고문헌 자체에서 보이는 텍스트 변형이다. 이 2가지 차이점 때문에 점수 행렬 H에서 점수가 줄어든 것을 관찰할 수 있다(그림 1 파란색 화살표 부분). 그러나 이러한 글자 차이에도 불구하고 앞뒤로 동일한 텍스트가 이어지기 때문에 인용문으로 의심되는 텍스트를 찾아낼 수 있게 된다.

4. 『동의보감』에 인용된 『황제내경소문』

앞서 살펴본 방법을 이용하여 『동의보감』과 『황제내경소문』 사이의 인용문을 추정하고 탐색하여 비교해 보았다.[19]

인용문 추정에서는 5-gram을 이용하여 중복되는 텍스트 양을 측정하였다. n-gram에서 n을 너무 작게 잡으면 인용과 무관하게 관습적으로 사용되는 문구가 많이 측정되고, n을 너무 크게 잡으면 작은 차이까지 지나치게 걸러내게 된다. 한문 텍스트는 1구가 4자로 이루어진 경우가 많으므로 4-gram 이하로 하였을 경우 관습적인 표현들까지 추출될 가능성이 높아진다. 본 연구에서는 이러한 문제를 줄이고자 5-gram을 적용하였다.

인용문 탐색에서는 상점(match score)을 3점, 벌점(gap cost)을 2점으로 하였다. 또 일정 길이 이하의 텍스트는 우연의 일치로 볼 수 있으므로 8자(4자 2구절) 보다 큰 10자 이상의 텍스트만을 인용문으로 인정하였다. 이렇게 찾아진 인용문들을 통계치를 중심으로 살펴보도록 하자. 결과 표 안의 숫자는 전체 텍스트 대비 인용문 텍스트의 비율을 의미한다. 본문에서는 괄호 안에 제시하였다.

19) 결과의 정확성을 높이기 위해 전처리를 통해 비교 대상 텍스트에서 '六'(F9D1)-'六'(516D)과 같은 다중코드자를 통일해 주었다. 다른 변형은 가하지 않았다.

우선 『동의보감』에 인용된 『황제내경소문』의 편별 텍스트를 살펴보자(표 1 참조). 『동의보감』에 가장 많이 인용된 편은 어떤 것일까. SW알고리즘을 이용한 탐색 결과 〔SW(unique)〕를 기준으로 보았을 때, 23宣明五氣(100.00%), 38玆論(93.32%), 31熱論(89.24%), 65標本病傳論(87.50%), 1上古天眞論(86.93%), 52刺禁論(82.35%), 12異法方宜論(75.62%), 34逆調論(74.65%), 22藏氣法時論(72.86%), 43痺論(72.45%) 등이 『동의보감』에 많이 인용된 것으로 나타났다. n-gram을 이용한 추정치 〔nGram〕를 기준으로 하였을 때의 결과는 23宣明五氣(82.26%), 31熱論(76.33%), 1上古天眞論(73.53%), 38玆論(68.97%), 22藏氣法時論(61.03%), 34逆調論(58.51%), 2四氣調神大論(57.61%), 65標本病傳論(57.59%), 43痺論(56.86%), 8靈蘭秘典論(53.13%)으로, 조금 다르게 나타났다.

〔nGram〕과 〔SW(unique)〕가 거시적으로는 유사한 경향을 보였으나, 세부적으로는 차이가 적지 않았다. 실제로 후자가 전자에 비해 평균 6.65% 정도 적게 측정되었다. n-gram 추정은 한 글자라도 다르면 제외되므로 이를 보정할 수 있는 SW알고리즘 이용 결과보다 결과가 작게 나타나게 된다.[20] 그렇다 하더라도 일부 편에서는 무시하기 어려운 차이를 보였다. 12異法方宜論의 경우, 추정치 〔nGram〕에서는 33.24%가 나타났지만, 인용문 탐색 결과 〔SW(unique)〕 75.62%의 텍스트가 인용된 것으로 나타나 그 차이가 42.38%에 달했다. 52刺禁論(30.82%), 65標本病傳論(29.91%), 38玆論(24.35%) 등도 적지 않은 차이를 보였다. 〔SW(unique)〕가 본질적인 의미의 인용문 추출에 가깝다고 할 수 있으므로, 〔nGram〕이 어느 정도의 오차를 가지고 있음을 보여준다.

인용문의 양 뿐만 아니라 인용문의 인용 횟수 역시 해당 텍스트의 중요성을 보여준다. 실제로 『황제내경소문』의 일부 텍스트는 『동의보감』에 1회 이상 중복되어 인용되었다. 특히 일부 편은 매우 빈번히

20) 그림 1 예시의 경우, 〔SW(unique)〕를 이용한 결과에서는 인용문 길이가 13~14자로 측정된다. 그러나 글자 차이에 민감한 〔nGram〕으로 추정할 경우, 양자 사이에 동일한 5-gram은 "人咳非獨肺"과 "咳非獨肺也" 뿐이기 때문에 인용문 길이는 "人咳非獨肺也"의 6자로 측정된다.

표 1.『동의보감』에 인용된『황제내경소문』의 편별 텍스트 분량 (nGram: n-gram을 통한 인용문 분량 추정 결과, SW: SW알고리즘을 통한 인용문 탐색 결과. unique: 중복을 무시한 분량. overlap: 중복을 허용한 분량. 녹색: 상위 10개, 빨간색: 하위 10개)

구분	nGram	SW(unique)	SW(overlap)	구분	nGram	SW(unique)	SW(overlap)	구분	nGram	SW(unique)	SW(overlap)
1上古天眞論	73.53%	86.93%	105.54%	28通評虛實論	14.96%	20.73%	23.36%	55長刺節論	16.59%	20.40%	20.40%
2四氣調神大論	57.61%	44.66%	56.96%	29太陰陽明論	37.76%	38.37%	60.61%	56皮部論	27.77%	33.02%	33.02%
3生氣通天論	40.09%	44.62%	60.93%	30陽明脈解	49.08%	66.79%	80.81%	57經絡論	2.08%	0.00%	0.00%
4金匱眞言論	12.41%	14.73%	14.73%	31熱論	76.33%	89.24%	94.40%	58氣穴論	6.34%	7.96%	7.96%
5陰陽應象大論	34.27%	41.43%	94.35%	32刺熱	5.28%	5.43%	15.98%	59氣府論	1.73%	0.00%	0.00%
6陰陽離合論	1.77%	0.00%	0.00%	33評熱病論	40.72%	46.95%	48.48%	60骨空論	9.57%	12.96%	16.05%
7陰陽別論	10.95%	7.67%	10.33%	34逆調論	58.51%	74.65%	90.10%	61水熱穴論	22.98%	33.11%	44.87%
8靈蘭秘典論	53.13%	58.21%	69.55%	35瘧論	38.19%	48.21%	50.39%	62調經論	18.23%	29.46%	42.31%
9六節藏象論	14.27%	17.37%	19.26%	36刺瘧	6.75%	7.41%	7.41%	63繆刺論	7.61%	9.53%	13.57%
10五藏生成	34.74%	43.54%	46.24%	37氣厥論	42.67%	61.64%	84.48%	64四時刺逆從論	11.61%	18.71%	29.52%
11五藏別論	32.18%	39.66%	43.68%	38欬論	68.97%	93.32%	96.12%	65標本病傳論	57.59%	87.50%	97.92%
12異法方宜論	33.24%	75.62%	78.95%	39擧痛論	42.67%	52.68%	60.87%	66天元紀大論	5.26%	3.65%	3.65%
13移精變氣論	1.28%	0.00%	0.00%	40腹中論	18.09%	28.87%	28.87%	67五運行大論	7.21%	10.87%	10.87%
14湯液醪醴論	1.28%	0.00%	0.00%	41刺腰痛	1.32%	0.00%	0.00%	68六微旨大論	5.96%	6.55%	9.04%
15玉版論要	1.06%	0.00%	0.00%	42風論	13.03%	21.57%	22.93%	69氣交變大論	5.37%	9.95%	11.32%
16診要經終論	18.10%	25.52%	27.60%	43痺論	56.86%	72.45%	75.38%	70五常政大論	7.22%	8.56%	8.98%
17脈要精微論	31.36%	40.07%	55.48%	44痿論	42.01%	47.11%	76.02%	71六元正紀大論	2.43%	3.09%	3.24%
18平人氣象論	24.14%	28.76%	29.77%	45厥論	43.72%	55.90%	81.03%	72刺法論	2.70%	7.11%	8.56%
19玉機眞藏論	39.20%	57.65%	75.14%	46病能論	27.79%	50.58%	50.92%	73本病論	1.14%	0.00%	0.00%
20三部九候論	29.90%	40.02%	42.87%	47奇病論	26.61%	41.00%	43.19%	74至眞要大論	13.53%	17.19%	21.19%
21經脈別論	38.92%	44.13%	52.89%	48大奇論	8.20%	10.43%	10.43%	75著至教論	0.00%	0.00%	0.00%
22藏氣法時論	61.03%	72.86%	81.60%	49脈解	7.61%	10.38%	10.38%	76示從容論	5.01%	6.55%	8.36%
23宣明五氣	82.26%	100.00%	122.62%	50刺要論	20.61%	21.05%	21.05%	77疏五過論	11.13%	13.76%	13.76%
24血氣形志	20.60%	29.00%	29.00%	51刺齊論	0.57%	0.00%	0.00%	78徵四失論	0.55%	0.00%	0.00%
25寶命全形論	2.66%	1.88%	1.88%	52刺禁論	51.53%	82.35%	149.65%	79陰陽類論	0.00%	0.00%	0.00%
26八正神明論	16.22%	26.25%	26.25%	53刺志論	50.22%	65.64%	72.25%	80方盛衰論	7.59%	15.04%	15.04%
27離合眞邪論	10.50%	14.79%	14.79%	54鍼解	6.48%	7.22%	7.22%	81解精微論	4.37%	3.24%	3.24%

인용되었는데, 5陰陽應象大論의 경우 중복을 무시한 인용 탐색 결과〔SW(unique)〕가 41.43% 였지만, 중복을 고려한 인용 탐색 결과〔SW(overlap)〕는 94.35% 였다. 이는 인용문들이『동의보감』에 평균 2.28번 인용되었다는 것을 의미한다. 이처럼 인용문 중복의 변수가 있으므로 중복이 많이 된 텍스트를 검토하는 것도 인용문의 분량과 함께 고려해야 한다.

다음으로『동의보감』을 기준으로『황제내경소문』을 인용하고 있는 텍스트를 살펴보자(표 2 참조).『황제내경소문』을 가장 많이 인용하고 있는 문(門)은 어디일까. SW알고리즘을 이용한 탐색 결과〔SW(unique)〕를 기준으로 보았을 때, 42辨證(39.59%), 41審病(29.80%), 34骨(23.34%), 43診脈(12.92%), 40天地運氣(11.68%), 32脈(11.42%), 1身形(10.06%), 44用藥(9.97%), 64痎瘧(9.15%), 31肉(8.59%) 순으로 나타났다. n-gram을 이용한 추정치〔nGram〕를 기준으로 하였을 때의 결과는 42辨證(25.65%), 41審病(24.45%), 34骨(21.59%), 1身形(9.49%), 44用藥(8.96%), 40天地運氣(8.48%), 32脈(8.17%), 43診脈(7.87%), 64痎瘧(7.59%), 31肉(5.81%)으로, 다소 차이가 있었다.

『동의보감』을 기준으로 보았을 때도 〔nGram〕과 〔SW(unique)〕가 거시적으로는 유사한 경향을 보였으나, 평균 0.88% 정도 적게 측정되었다.『황제내경소문』을 기준으로 보았을 때보다 차이가 크게 줄었는데, 이는 인용된 텍스트에 비해 전체 텍스트의 분량이 크기 때문인 것으로 해석할 수 있다. 중복 인용된 텍스트〔SW(overlap)〕의 경우 중복을 무시한 결과〔SW(unique)〕와 큰 차이가 없었다. 이는『황제내경소문』내에서 유사하거나 동일한 조문이 반복되는 경우에만 성립하므로 당연한 결과이다.

SW알고리즘을 통해 탐색한『동의보감』과『황제내경소문』의 인용문 전체 결과는 지면의 한계를 고려하여 웹어플리케이션 형태로 온라인에 게시하였다(그림 2 참조).[21]

21) 텍스트 우측 상단의 어깨번호는 인용문으로 의심되는 텍스트 쌍을 의미한다. 해당 어깨번호를 클릭하거나 왼쪽 상단 드롭다운 메뉴에서 숫자를 클릭하면 해당 텍스트 쌍으로 이동할 수 있다. Detecting Local Text Reuse Demo App 2020 [Internet]. Available from: https://pinedance.github.io/demo#/detecting-local-text-reuse

표 2.『동의보감』에 보이는 황제내경 소문의 텍스트 분량 (약어 및 기호: 앞의 표와 동일함)

구분	nGram	SW(unique)	SW(overlap)	구분	nGram	SW(unique)	SW(overlap)	구분	nGram	SW(unique)	SW(overlap)
1身形	9.49%	10.06%	12.45%	27臍	0.21%	0.00%	0.00%	53火	1.27%	1.94%	2.53%
2精	1.11%	2.05%	2.21%	28腰	0.50%	0.73%	0.73%	54內傷	1.56%	1.94%	1.94%
3氣	3.56%	4.01%	4.02%	29脇	1.35%	2.07%	2.07%	55虛勞	0.50%	0.56%	0.56%
4神	3.67%	4.82%	4.82%	30皮	2.85%	3.05%	3.23%	56霍亂	0.43%	0.58%	0.58%
5血	0.90%	0.84%	0.84%	31肉	5.81%	8.59%	8.59%	57嘔吐	0.29%	0.17%	0.17%
6夢	4.17%	8.04%	8.04%	32脈	8.17%	11.42%	12.68%	58咳嗽	2.17%	2.78%	2.82%
7聲音	2.99%	3.48%	3.48%	33筋	4.51%	5.65%	5.69%	59積聚	1.19%	1.71%	2.23%
8言語	2.77%	3.70%	3.70%	34骨	21.59%	23.34%	23.34%	60浮腫	2.64%	3.24%	3.24%
9津液	2.55%	3.16%	3.16%	35手	4.95%	5.07%	6.10%	61腹滿	0.88%	0.96%	0.96%
10痰飮	0.05%	0.00%	0.00%	36足	4.57%	7.47%	9.13%	62消渴	0.79%	1.03%	1.03%
11五臟六腑	4.80%	6.23%	10.28%	37毛髮	0.25%	0.00%	0.00%	63黃疸	0.12%	0.00%	0.00%
12小便	0.68%	1.06%	1.06%	38前陰	1.01%	1.35%	1.35%	64痰癖	7.59%	9.15%	9.15%
13大便	0.52%	0.59%	0.59%	39後陰	0.36%	0.56%	0.56%	65瘧疫	0.70%	1.45%	1.45%
14頭	1.20%	1.84%	1.84%	40天地運氣	8.48%	11.68%	12.26%	66邪祟	0.06%	0.00%	0.00%
15面	2.20%	2.56%	2.56%	41審病	24.45%	29.80%	29.80%	67癲疝	0.44%	0.57%	0.57%
16眼	0.45%	0.39%	0.39%	42辨證	25.65%	39.59%	40.66%	68諸瘡	0.23%	0.26%	0.26%
17耳	0.28%	0.21%	0.21%	43診脈	7.87%	12.92%	13.55%	69諸傷	0.00%	0.00%	0.00%
18鼻	0.94%	1.46%	1.46%	44用藥	8.96%	9.97%	12.81%	70解毒	0.00%	0.00%	0.00%
19口舌	0.68%	1.14%	1.14%	45吐	0.32%	0.32%	0.32%	71救急	3.20%	4.63%	4.63%
20牙齒	0.50%	0.82%	1.07%	46汗	0.71%	0.00%	0.00%	72怪疾	0.00%	0.00%	0.00%
21咽喉	0.11%	0.00%	0.00%	47下	0.00%	0.00%	0.00%	73雜方	0.01%	0.00%	0.00%
22頸項	2.02%	2.02%	2.02%	48風	2.18%	2.62%	2.75%	74婦人	0.13%	0.18%	0.18%
23背	2.78%	5.01%	5.01%	49寒	2.00%	2.26%	2.26%	75小兒	0.03%	0.00%	0.00%
24胸	0.43%	0.56%	0.79%	50暑	0.79%	0.79%	0.79%	76湯液篇	0.57%	0.72%	0.76%
25乳	0.02%	0.00%	0.00%	51濕	0.76%	1.24%	1.24%	77鍼灸篇	2.45%	3.49%	5.12%
26腹	4.85%	6.73%	6.89%	52燥	0.00%	0.00%	0.00%				

그림 2.『동의보감』과『황제내경소문』의 인용문 탐색 결과 웹어플리케이션 모습 (노란색 바탕: 인용문으로 의심되는 텍스트들, 초록색 바탕: 인용문으로 의심되는 텍스트 쌍)

Ⅲ. 결론

지금까지 한의학 고문헌 텍스트에서 인용문을 추정하고 탐색하는 계량적인 방법을 소개하고, 『동의보감』과 『황제내경소문』 사이의 인용문 분석을 진행해 보았다.

동아시아 전통사회에서 한문으로 저술된 한의학 고문헌은 당시의 학술적 경향으로 인해 권위를 지닌 경전을 빈번히 인용하였을 뿐만 아니라 치료 이론과 방법에 대한 지식을 이전 서적에서 채록하기 위해 많은 서적에서 다양한 내용을 인용하였다. 오늘날 한의학 고문헌을 연구하기 위해 해당 서적에 수록된 인용문과 인용 서적을 알아내고 분석하는 일은 기초적이고 중요한 연구 방법으로 여겨지고 있다. 그러나 이 방법은 많은 연구 자원이 소모되는 일이다. 다행히 전산 장비를 이용한 계량적인 방법을 통해 인용문을 추정하고 탐색하는 일이 어느 정도 가능하다. 이 방법을 사용하면 사람이 수작업으로 수행하기 힘든 대량의 텍스트 비교가 가능하다.

본 연구에서는 n-gram을 이용하여 인용문의 분량을 추정하는 방법과, Smith-Waterman 알고리즘을 이용하여 인용문을 직접 탐색하는 방법을 소개하였다. 『동의보감』과 『황제내경소문』을 대상으로 진행한 분석 결과, 인용문 분량 측정에서 전자와 후자는 유사한 경향성을 보였으나 때때로 적지 않은 차이를 나타내기도 했다. 이는 n-gram을 이용한 인용문 추정 방법이 정확한 결과를 위한 목적으로 사용하기 어렵다는 점을 보여준다. 그럼에도 이 방법은 매우 짧은 시간 내에 결과를 도출할 수 있기 때문에 유사한 문서를 선별검사 하는 등 특정한 목적 아래서는 여전히 유용한 방법이 될 수 있다.

SW알고리즘을 이용한 인용문 탐색 방법은 문자열이 완전히 일치하지 않는 경우에도 인용문으로 의심되는 유사한 구절을 비교적 잘 찾아낼 수 있다는 장점을 가진다. 그러나 몇 가지 한계를 지닌다. 먼저 한의학 고문헌의 인용문은 텍스트를 유사하게 전사하는 경우 이외에도 텍스트 내에서 구절의 순서가 일부 도치되거나 의미에 따라 텍스트가 축약되는 형태로 이루어지는 경우가 있다. 이를 찾기 위해서는 의미를 비교해야 한다. 그러므로 SW알고리즘으로 이러한 경우까지 찾아내기는 힘들다. 다음으로 SW 알고리즘은 글자와 글자를 하나하나 비교해 나가므로 연산 시간이 매우 오래 걸린다. n-gram을 이용한 방법과는 대조적으로 많은 시간과 컴퓨팅 자원을 소모하기 때문에 정밀한 결과가 필요할 때는 유용하지만 대량의 정보를 빠르게 처리해야 하는 경우에는 적당하지 않다.[22]

한의학 고문헌의 전산화가 빠르게 진행되고 있는 만큼 계량적인 방법을 통한 인용문 추정과 탐색을 더 발전시키고 적절히 활용해 나간다면 연구 활동의 효율성을 높이고 더 정교한 연구 결과를 도출하는 데 좋은 방편이 될 수 있을 것이다.

감사의 말씀

본 연구는 한국한의학연구원 주요사업 “AI 한의사 개발을 위한 임상 빅데이터 수집 및 서비스 플랫폼 구축(KSN2012110)”의 지원을 받아 수행되었습니다.

Reference

1. 김남일. 『鄕藥集成方』의 인용문헌에 대한 연구. 진단학보. 1999. 87.
2. 김중권. 『동의보감』의 문헌적 연구: 인용문헌을 중심으로. 서지학연구. 1995. 11.
3. 김중권. 『鄕藥集成方』의 引用文獻 分析. 書誌學硏究. 2006. 35.
4. 박상영. 『인제지』의 조선후기 의사학적 위상과 의의: 미키 사카에의 재인용[孫引] 지적과 학술가치 평가에 대한 재검토. 한국실학연구. 2013. 25.
5. 박성식, 송일병. 사상의학의 의학적 연원과 이제마 의학사상에 대한 연구: 동의수세보원 인용문을 중심으로. 1993. 5(1).
6. 박영호 저, 김도련 편. 한국 고문의 이론과 전

22) 양자를 함께 사용하여 서로의 단점을 보완하는 방법도 생각해 볼 수 있다. D. A. Smith 등의 연구가 하나의 예이다.

개(조선중기 고문론 연구). 경기도. 태학사. 1998.
7. 안상우, 김남일.『醫方類聚』總論의 體制와 引用方式 分析. 경희한의대논문집. 1999. 22(1).
8. 윤용갑, 강순수. 방약합편에 수록된 처방의 주치별 계통분류와 인용문헌에 대한 고찰. 원광한의대논문집. 1986. 4.
9. 이계성. 다중 지역 정렬 알고리즘. 문화기술의 융합. 2019. 5(3).
10. 이정화. 제중신편의 인용문헌 연구. 서지학보. 2010. 35.
11. 이필우, 윤창렬. 동의수세보원 인용문에 대한 연구. 한의학연구소 논문집. 2004. 12(2).
12. 정요상, 이명호, 최동훈. Smith-Waterman 알고리즘을 위한 GPU 상에서의 Multi-Stream 기반 병렬화. 한국정보과학회 2014년 동계학술발표회 논문집. 2014.
13. 지창영. 제중신편의 인용방식에 대한 연구. 한국의사학회지. 2008. 21(1).
14. 최환수, 신순식. 의방유취의 인용서에 관한 연구(1). 한국한의학연구원논문집. 1997. 3(1).
15. D. A. Smith, R. Cordel, E. M. Dillon, N. Stramp and J. Wilkerson. Detecting and modeling local text reuse. IEEE/ACM Joint Conference on Digital Libraries. 2014.
https://doi.org/10.1109/JCDL.2014.6970166.
16. Donald Sturgeon. Unsupervised identification of text reuse in early Chinese literature. Digital Scholarship in the Humanities. 2018. 33(3). https://doi.org/10.1093/llc/fqx024
17. Jangwon Seo , W. Bruce Croft. Local text reuse detection. Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2008.
18. MEDICLASSICS [homepage on the Internet]. Korea Institute of Oriental Medicine; 2015 [cited 30 Jan 2020]. Available from:
19. Smith, Temple F., and Michael S. Waterman. Identification of common molecular subsequences. Journal of molecular biology. 1981. 147(1).