



고의서에 나타난 경혈과 병증의 연관성 측정 및 시각화 - 침구자생경 분석 예를 중심으로 -

오준호

한국한의학연구원

Measure of the Associations of Acupoints and Pathologies Documented in the Classical Acupuncture Literature

Junho Oh

Korea Institute of Oriental Medicine

Objectives : This study aims to analyze the co-occurrence of pathological symptoms and corresponding acupoints as documented by the comprehensive acupuncture and moxibustion records in the classical texts of Far East traditional medicine as an aid to a more efficient understanding of the tacit treatment principles of ancient physicians. **Methods :** The Classic of Nourishing Life with Acupuncture and Moxibustion(Zhenjiu Zisheng Jing; hereinafter ZZJ) was selected as the primary reference book for the analysis. The pathology-acupoint co-occurrence analysis was performed by applying 4 values of vector space measures(weighted Euclidean distance, Euclidean distance, Cramér's V and Canberra distance), which measure the distance between the observed and expected co-occurrence counts, and 3 values of probabilistic measures(association strength, Fisher's exact test and Jaccard similarity), which measure the probability of observed co-occurrences. **Results :** The treatment records contained in ZZJ were preprocessed, which yielded 4162 pathology-acupoint sets. Co-occurrence was performed applying 7 different analysis variables, followed by a prediction simulation. The prediction simulation results revealed the Weighted Euclidean distance had the highest prediction rate with 24.32%, followed by Canberra distance(23.14%) and association strength(21.29%). **Conclusions :** The weighted Euclidean distance among the vector space measures and the association strength among the probabilistic measures were verified to be the most efficient analysis methods in analyzing the correlation between acupoints and pathologies found in the classical medical texts.

Key words : association, data mining, Acupuncture therapy

서 론

동아시아 전통의학에서 고문헌은 치료경험 자체 혹은 치료경험에서 얻어진 지견을 정리해 놓은 문서들의 집합이라고 할 수 있다.

오랜 기간 동안 커커이 누적된 고인들의 치료경험은 고의서 안에 잠장되어 지금까지 전해져 왔다. 오늘날 전통의학은 꾸준히 그 외연을 확장하고 있지만, 임상·연구·교육 각 분야에서 여전히 고문헌을 기반으로 삼고 있는 것은 방대한 시간 동안 누적된 치료경험

Received March 2, 2016, Revised March 11, 2016, Accepted March 15, 2016

Corresponding author: Junho Oh

Korea Institute of Oriental Medicine, 1672 Yuseong-daero, Yuseong-gu, Daejeon 34054, Korea
Tel: +82-42-868-9317, Fax: +82-42-863-9463, E-mail: junho@kiom.re.kr

This study was updated from previous poster presentation in the ICCMR2015 (doi:10.1016/j.imr.2015.04.247).

This Research was supported by the "KIOM Customization of TCM Knowledge for R&D(K16670)" funded by Korea Institute of Oriental Medicine.

© This is an open access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

때문이다.

전통의학을 지탱하는 고문헌의 수는 매우 많고, 그 속에 있는 치료경험은 헤아리기 어려울 정도로 방대하다. 고문헌에 담겨 있는 치료경험에 대한 공식적인 측정치는 없지만, 고의서의 수는 실전된 것을 포함하여 대략 2,880여종(『의적고醫籍考』기준, 1819년 일본 타키모토츠구丹波元胤가 춘추전국시대부터 청대까지의 의서에 대한 일종의 해제집)정도¹⁾이며, 본초의 수는 약 2,374종(『중수정화경 사증류비급본초重修政和經史證類備急本草』수록 약재 기준)²⁾, 경혈의 수는 약 930종(『경혈회해經穴匯解』기준, 1803년 일본 원창극 原昌克이 저술한 일종의 경혈 사전)³⁾, 방제의 수는 약 61,739종(『보제방普濟方』기준, 15세기 초 명나라 주숙朱肅 등이 편찬한 대형 의방서)에 달한다.

만약 고문헌에 수록된 이 방대한 치료경험들이 가리키는 바를 효과적으로 요약해 내거나 직관적으로 인식할 수 있다면 우리는 고문헌으로부터 더 많은 지혜들을 얻을 수 있을 것이다. 나아가 고문헌에 정리되어 있는 치료 지침을 오늘날 의료인들이 지니고 있는 암묵지(tacit knowledge)와 비교해 본다면 한의학의 이론적 발전과 임상 치료율 향상에도 직간접적으로 도움을 줄 수 있을 것이다.

그러나 고문헌 텍스트 내에서 의미 있는 결과를 추출하기 위해서는 고문헌 텍스트를 어떻게 디지털 데이터로 만들 것인가에 대한 문제와 이렇게 만들어진 디지털 데이터를 어떻게 분석해 낼 것인가의 문제를 먼저 고민해야 한다. 고문헌의 디지털화는 꾸준히 진행되고 있으나 여전히 미비한 상태여서⁴⁾ 분석 대상이 되는 텍스트를 자유롭게 선택하는데 제약이 많다. 또 유니코드가 보편적으로 사용되면서 한자 텍스트의 디지털화가 보다 손쉬워졌으나, 형태는 같지만 코드 포인트(code point)가 다른 다중코드 문제, 의미가 같아 혼용되지만 모양이 다른 이체자 문제가 여전히 존재한다.

디지털화 이후에도 데이터 분석을 위해서는 텍스트를 최대한 구조화시켜 치료정보별로 구분해 주어야 하고, 동의어 및 유의어 사전을 구축하여 합곡(合谷)과 호구(虎口)같은 용어들의 빈도를 합쳐서 평가해야 한다. 또 치료대상과 치료방법의 연관성 분석을 위해 어떤 모델을 적용시킬 것인가, 그리고 그 결과를 어떻게 시각적으로 표현해 낼 것인가도 큰 주제이다.

데이터 분석과 관련해서는 많은 연구들이 존재하지만, 분석 대상 분야(domain)에 따라 최적화 해 나가야 한다는 것이 정론이다. 따라서 전통의학 고문헌 분석을 위해서는 해당 분야를 잘 아는 전문가들이 이를 고민해 나가야 한다. 최근 침구학 고의서를 대상으로 해당 텍스트의 이론적 특징을 고찰하거나⁵⁾, 경혈의 혈성⁶⁾ 및 조합 방식을^{7,8)} 통찰하기 위한 텍스트 마이닝(text mining) 연구들이 있었다. 본 연구에서는 의서가 담고 있는 치료대상과 치료방법

사이의 관계에 주목하여 병증과 경혈 사이의 연관성을 정량적으로 측정해 보았다. 본고에서는 분석 대상 문헌의 선정부터 시각화까지 전반적인 분석 과정을 개괄하여 향후 전통의학 고문헌 텍스트를 대상으로 분석연구를 진행하기 위해 선결되어야 할 문제점들을 짚어보고 이에 대한 현실적인 방법들을 제안해 보고자 한다.

연구방법

고문헌 텍스트 분석은 크게 분석 대상을 선정하여 디지털 텍스트로 구축하는 단계, 디지털 텍스트를 분석에 용이한 형태로 전처리하고 측정하고자 하는 용어에 대한 용어집을 만드는 단계, 그리고 마지막으로 분석 기법을 선택하고 그 결과를 시각화하는 단계로 구분할 수 있다.

1. 분석 대상 디지털 텍스트 구축

1) 대상 서적 선정: 분석하고자 하는 고문헌이 많은 치료정보를 담고 있는가, 분석 목적에 맞는 보편성 혹은 특수성을 지닌 문헌인가를 우선 고려해야 한다. 텍스트 분석을 수행할 때 기본이 되는 수치형 데이터는 용어의 출현 빈도이다. 따라서 빈도가 충분히 도출되지 않으면 양질의 분석 결과를 도출할 수 없으며, 충분한 양의 치료정보를 담고 있는 텍스트여야 분석을 시도해 볼 수 있다. 또한 분석 목적에 맞는 성격을 가진 의서여야 한다. 전통의학 전반을 살펴보자 한다면 이전 시대 의학 경험을 집대성하고 후대에 많은 영향을 끼친 서적, 즉 관찰서와 같이 보편성을 지닌 의서를 선택하는 것이 유리하다. 반면 지역적 혹은 시대적 특성을 검토해 보고자 한다면 해당 지역이나 시대의 특수성을 대표하는 의서가 좋다.

본 연구에서는 동아시아 전통의학에서 병증과 경혈의 연관성을 보편적으로 가늠해 보고자 하였다. 이런 취지에 따라 치료정보를 풍부하게 싣고 있으면서 영향력 있었던 침구서로 『침구자생경(鍼灸資生經)』을 선택하였다. 『침구자생경』은 동아시아 전체에서 가장 권위 있는 침구학 서적 가운데 하나였다. 남송(南宋)의 왕집중(王執中)이 12세기 말(1180~1195년 사이)에 모두 7권으로 저술한 이 책에는 『소문(素問)』, 『갑을경(甲乙經)』, 『천금요방(千金要方)』, 『명당경(明堂經)』 등 당시까지의 침구 치료 지식들이 폭넓게 집적되어 있어 북송(北宋) 관찰 침구서인 『동인수혈침구도경(銅人臉穴鍼灸圖經)』과 함께 후대에 큰 영향을 끼쳤다.

내용을 요약해 보면, 권1에는 수혈(俞穴)의 위치와 주치가 간략히 설명되어 있는데, 주로 왕유일(王維一)의 『동인수혈침구도경』에서 채록한 것이다. 권2에는 왕집중 자신의 독창적인 견해가 논문

형태로 수록되어 있다. 권3~7까지 총 5권에 걸쳐서는 약 4천여 개의 침구치료법이 병증에 따라 달려 있다. 양적으로 방대할 뿐만 아니라 병증에 따라 일목요연하게 분류되어 있어 텍스트의 반구조화도 비교적 용이하다.

침구자생경의 대표판본으로는 원(元) 천력(天曆) 광근서당인본(廣勤書堂印本), 명(明) 정통(正統) 광근서당신간본(廣勤書堂新刊本), 청(淸) 사고전서본(四庫全書本)이 있으며(Fig. 1), 현대 교점(校點)본으로는 침구명저집성본(鍼灸名著集成本, 1997)이 있다⁹⁾. 본 연구에서는 현대 교점본을 기준으로 원문을 전산화하여 연구를 진행하였다.

2) 디지털 텍스트 구축: 한자 텍스트의 전산화 작업은 문자 인코딩(character encoding)으로 인해 상당히 복잡한 문제를 야기한다.



Fig. 1. The Classic of Nourishing Life with Acupuncture and Moxibustion, Old Book Image.

현재 가장 합리적인 방법은 국제적으로 통용되고 있는 유니코드 (UTF-8, UTF-16)로 한자 텍스트를 구축하는 것이다. 이 때 주의할 것은 유니코드 상에 존재하는 ‘다중 유니코드 문자(이하 다중코드자)’와 ‘이체자’ 문제이다.

(1) 다중코드자 처리: 코드 포인트(code point)는 컴퓨터가 어떤 문자를 인식하는 고유번호라고 할 수 있다. UTF-8의 경우 16진수 4자리로 이루어져 있다. 유니코드에는 한자의 모양이 같으나 코드 포인트가 다른 복수의 글자가 존재한다. 주로 두음법칙 혹은 다수의 훈음(訓音) 등 음이 2개 이상인 한자 때문에 야기된 문제이다. ‘更’의 경우 ‘깡’과 ‘경’ 두 가지 음으로 읽하는데, 각각의 음에 따라 유니코드에도 ‘更(F901)’와 ‘更(66F4)’ 두 가지 코드가 존재한다(괄호 안은 코드 포인트). 육안으로 보았을 때에는 똑같지만 컴퓨터에서는 전혀 다른 글자로 인식되기 때문에 분석 대상 텍스트에 이러한 글자들을 하나로 합치지 않으면 서로 다른 용어로 인식된다. 한자 텍스트 검색을 지원하는 DB들에서도 데이터 입출력시 다중코드자 통합이 이루어지지 않아 검색 결과가 다르게 나타나는 경우가 드물지 않다¹⁰⁾.

The Unicode Consortium(이하 콘소시움)에서 이미 460여개
다중코드자에 대한 매핑 정보를 제공하고 있기 때문에 이를 통해
디지털 텍스트의 코드를 통합할 수 있다^[1]. 본 연구에서도 Table 1과
같이 The Unicode Consortium이 제공하고 있는 매핑 데이터를

Table 1. Multi-code Characters(example)

豈(F900) / 豈(8C48), 更(F901) / 更(66F4), 車(F902) / 車(8ECA),
賈(F903) / 賈(8CC8), 滑(F904) / 滑(6ED1), 串(F905) / 串(4E32),
匂(F906) /匂(53E5), 龜(F907) / 龜(9F9C), 契(F909) / 契(5951),

Table 2 Variants Characters

隸(隸), 窓(窗), 屛(屏), 韻(斷), 絶(絕), 蓋(蓋), 豊(豐), 脇(脅), 郎(郎), 郎(郎), 面(面), 胸(脣), 舉(舉), 兩(両), 繩(繩), 雙(雙), 窮(窮), 兩(両), 脍(脉), 關(闔), 麥(麦), 清(清), 青(青), 邊(邊), 翳(翳), 宮(宫), 壘(垒), 強(強), 往(往), 葱(葱), 葱(葱), 檳(浜), 蟲(虫), 蟲(虫), 鹽(塩), 條(条), 蘇(蘚), 屬(属), 解(解), 鹹(鹹), 蒜(蒜), 杯(盃), 怪(怪), 賓(賓), 靈(灵), 逃(逃), 體(體), 冒(冒), 竊(竊), 賣(宗), 腸(腸), 寶(冠), 嘴(岬), 寶(宝), 驗(驗), 繼(繼), 裏(裡), 膽(胆), 曜(晒), 異(异), 斷(断), 來(来), 亂(乱), 站(並), 濕(湿), 鬱(鬱), 鉛(鉛), 虎(虎), 略(畧), 蟹(蟹), 微(微), 濁(濁), 飲(飲), 疎(疎), 婚(婚), 敝(敝), 凉(凉), 脚(腳), 值(值), 群(羣), 僞(偽), 鑑(鑑), 啓(啟), 嘗(尝), 飢(饑), 甜(甜), 雞(鷄), 奇(奇), 概(槩), 頗(頗), 呂(頗), 壇(填), 爲(為), 尚(尚), 喻(喻), 愉(俞), 苟(苟), 水(冰), 眞(真), 卍(廿), 濕(溼), 承(乘), 慎(慎), 既(既), 卽(即), 昂(昂), 產(產), 痢(痹), 鈎(鈎), 雞(雞), 迴(迴), 併(併), 霽(皋), 沈(沉), 巍(巔), 驅(駆), 衆(眾), 禿(廩), 抵(抵), 隘(鄰), 爭(鬥), 證(証), 教(教), 晚(晚), 姮(姪), 臟(肚), 脣(唇), 奮(奮), 漑(溉), 黃(黃), 鄉(鄉), 猪(豬), 概(概), 頤(願), 羸(羸), 溫(溫), 沈(沉), 寶(寶), 鐵(鐵), 頤(凶), 兑(兌), 臥(卧), 兔(兔), 癢(癩), 跖(蹠), 肓(厘), 啞(唾), 淩(惱), 最(最), 麵(麵), 陷(陷), 體(体), 瘡(瘡), 筋(筋), 喳(喫), 脈(脈), 軟(軟), 遍(徧), 麻(麻), 宿(宿), 却(卻), 邑(郤), 咳(嗽), 虛(虛), 悅(悅), 說(說), 蓋(蓋), 瓶(瓶), 素(素), 穗(穂), 研(研), 幷(并), 別(另), 瘦(痩), 喰(嗜), 彥(彥), 卷(卷), 卷(券), 經(經), 瘡(暗), 瘤(啞), 頸(頸), 翼(翼), 鍼(針), 巍(岩), 商(商), 脣(謁), 歸(帰), 髮(口), 祀(祕), 脣(口), 覺(蹠), 蝶(蝶), 謬(廊), 非(菲), 蒼(莽), 瘋(口), 騙(口), 嘔(口), 譯(禪), 駭(駭), 瘋(口), 呪(口), 聰(口), 呪(口), 聰(口), 蟲(蟲), 蟲(蟲), 驚(蟲), 驚(蟲), 凡(凡), 黑(黑), 瞳(睭), 眇(睭), 焰(焰), 香(脣), 瀉(瀉), 鮎(鰐), 鮎(鰐), 滿(滿), 點(點), 脿(臚), 瘡(瘡), 遊(遊), 嘴(齶), 臟(臟), 烹(煮), 鵝(鴛), 兒(兒), 雁(鴈), 檉(樹), 桑(桑), 陰(陰), 隱(隱), 搖(搖), 吞(吞), 髮(髮), 慈(慈), 穩(穗), 深(深), 更(更), 忘(忘), 選(選), 荒(荒), 旨(吉), 處(處), 处(處), 等(等), 郭(郭), 濟(濟), 槩(槩), 槩(槩), 游(游), 游(遊), 幼(幼), 幻(幻), 疎(疎), 疎(疎), 臺(臺), 哉(哉), 觀(觀), 析(析), 修(脩), 國(國), 國(國), 第(第), 虎(虎), 變(變), 變(變), 顯(顯), 顯(顯), 富(富), 覺(覺), 覺(覺), 從(從), 曾(曾), 協(協), 翁(翁), 楚(楚), 飭(餳), 𠂔(𠂔), 滔(滔), 宜(宜), 鬱(鬱), 養(養), 氣(氣), 藥(藥), 惠(惠), 藏(藏), 濟(濟), 佛(仏), 繢(續), 繢(續), 觀(觀), 縣(県), 傳(伝), 歷(歷), 廣(廣), 錢(錢), 縱(縱), 聰(聰), 增(增), 舉(举), 閱(閱), 營(營), 搜(搜), 歲(歲), 廢(廢), 峻(峻), 拜(拜), 怯(怯), 圖(图), 權(權), 團(团), 戸(戸), 輪(軸), 勸(勸), 二(二), 鄉(鄉), 寶(美), 錄(錄), 曉(曉), 每(每), 擇(択), 驗(驗), 劑(剂), 梔(梶), 驅(駆), 總(總), 變(變), 擴(扩), 榮(榮), 澤(沢), 賀(賀), 獸(獸), *溪(谿), *渝(腋), *願(凶)

*경혈에만 특수하게 적용되는 이체자.

이용해 다중코드자를 전부 통합하여 사용하였다.

(2) 이체자 처리: 이체자는 모양은 다르지만 유사하거나 동일한 의미를 가지는 글자로서, 絶과 絶, 盖와 蓋, 臀과 胸 등이 그것이다. 콘소시움에서도 이 점을 인식하여 한중일 한자 일반에 대해 포괄적인 이체자 정보를 꾸준히 업데이트 하여 제공하고 있다. 그러나 이 정보는 상당히 포괄적이어서 분석하고자 하는 텍스트의 성질에 따라 커스터마이징이 불가피하다.

일반적인 이체자 외에 한의학 문헌에만 등장하는 이체자도 존재 한다. 예를 들어 豊(豐) 같은 경우에는 콘소시움 이체자 정보에 들어 있지만 面(面)과 같은 경우에는 포함되어 있지 않다. 또 溪(谿), 臌(臉), 頤(囟) 등과 같이 전통의학 중에서도 경혈에서만 통용되는 특수한 경우도 있는데, 이들을 따로 처리하지 않을 경우 ‘太谿’와 ‘太溪’는 전혀 다른 2개의 경혈을 의미하는 것이 되어 잘못된 결과를 도출하게 된다.

본 연구에서는 원문에 사용된 정경혈(正經穴) 명칭들을 전수조사하여 통일하는 방식으로 이 문제를 해결하였다. 이 때 통합된 이체자들은 Table 2와 같다.

2. 디지털 텍스트 전처리 및 용어 사전 구축

1) 텍스트 반구조화: 텍스트를 기반으로 한 데이터 분석에는 용어의 빈도가 기초 자료가 된다. 따라서 데이터를 완전히 행(row)과 열(column)로 구조화할 수 없다하더라도 원하는 단위별로 나누어 반구조화 할 필요가 있다. 『침구자생경』의 경우, 194개의 병증에 따라 치료방법이 편재되어 있고, 각 편에는 다시 여러 개의 치료방법들이 모여 있다. 따라서 각 치료방법들을 하나하나 구분할 필요가 있다. 『침구자생경』의 「중풍(中風)」 앞 부분을 예로 들면, 줄글로 나열된 글들을 내용에 따라 구분하고 정렬하여 반구조화 데이터(semi-structured data)를 만들 수 있다. 용어사전이 구축되면 용어사전 가운데 어떤 용어가 포함되어 있는지 검색하여 이를 다시 구조화 데이터(structured data)로 정제한다. Table 3은 중풍(中風)의 치료방법 일부를 반구조화 한 예이다.

2) 용어사전 구축: 반구조화된 데이터를 구조화 데이터로 만들기 위해서는 용어 빈도 측정에 기준으로서 용어사전이 필수적이다. 대상 텍스트의 양이 적을 때에는 수작업으로 의미 있는 용어를 정리해 나갈 수 있지만, 텍스트의 양이 많아지면 오류의 가능성이 커지고 필요한 자원도 많아져 수작업은 사실상 불가능하다. 용어사전을

Table 3. Semi-structured Data Example for Stroke(中風) Treatment

Original Text	Semi-structured Data	Structured Data		
		Major Symptom	Minor Symptoms	Acupuncture Points
小兒但是風病，諸般醫治不瘥，灸率谷(『明下』). 黃帝療中風，眼戴上，及不能語者，灸第二椎・第五椎上各十壯，齊下火，炷如半棗核大，立瘥。黃帝問岐伯曰：中風半身不遂，如何灸？答曰：凡人未中風一兩月前，或三五月前，非時足經上忽酸重頑痺，良久方解。此將中風之候，急灸三里・絕骨四處三壯，後用葱・薄荷・桃柳葉煎湯淋洗，驅逐風氣于瘡口出。凡人不信此法，飲食不節，酒色過度，忽中此風，言語謇澁，半身不遂。宜七處齊下火，各三壯。風在左灸右，右灸左。百會・耳前髮際・肩井・風市・三里・絕骨・曲池七穴，神效不能具錄，依法灸無不愈。百會・曲鬢・肩髃・曲池・風市・足三里・絕骨共十三穴，灸風中臟，氣塞涎上，不語，極危者，下火立效。其狀覺心中憤亂，神思不怡，或手足麻，此將中臟之候，不問風與氣，但依次自上及下，各灸五壯，日別灸隨年壯。凡遇春秋，常灸以泄風氣。素有風人，可保無虞。此能灸暴卒，百會・風池・大椎・肩井・曲池・間使・足三里，共十二穴。范子默自壬午五月間口眼喎斜，灸聽會等三穴卽正。右手足麻無力，灸百會・髮際等七穴得愈。未年八月間，氣塞涎上，不能語，金虎丹加膩粉服至四丸半。氣不通，涎不下，藥從鼻中出，魂魄飛揚，如墜江湖中，頃欲絕，灸百會・風池等左右共十二穴。	01 : 小兒但是風病，諸般醫治不瘥，灸率谷(『明下』).	中風	小兒；風病	率谷
	02 : 黃帝療中風，眼戴上，及不能語者，灸第二椎・第五椎上各十壯，齊下火，炷如半棗核大，立瘥。	中風	眼戴上；不能語	第二椎；第五椎上
	03 : 黃帝問岐伯曰：中風半身不遂，如何灸？答曰：凡人未中風一兩月前，或三五月前，非時足經上忽酸重頑痺，良久方解。此將中風之候，急灸三里・絕骨四處三壯，後用葱・薄荷・桃柳葉煎湯淋洗，驅逐風氣于瘡口出。	中風	半身不遂；頑痺；淋；風氣；瘡口	三里；絕骨
	04 : 凡人不信此法，飲食不節，酒色過度，忽中此風，言語謇澁，半身不遂。宜七處齊下火，各三壯。風在左灸右，右灸左。百會・耳前髮際・肩井・風市・三里・絕骨・曲池七穴，神效不能具錄，依法灸無不愈。	中風	言語謇澁；半身不遂	百會；耳前髮際；肩井；風市；三里；絕骨；曲池
	05 : 百會・曲鬢・肩髃・曲池・風市・足三里・絕骨共十三穴，灸風中臟，氣塞涎上，不語，極危者，下火立效。其狀覺心中憤亂，神思不怡，或手足麻，此將中臟之候，不問風與氣，但依次自上及下，各灸五壯，日別灸隨年壯。	中風	曲鬢；中臟；氣塞；涎上；不語；心中憤亂；神思不怡；手足麻	百會；曲鬢；肩髃；風市；足三里；絕骨
	06 : 凡遇春秋，常灸以泄風氣。素有風人，可保無虞。此能灸暴卒，百會・風池・大椎・肩井・曲池・間使・足三里，共十二穴。	中風	泄風；風氣	百會；風池；大椎；肩井；曲池；間使；足三里
	07 : 范子默自壬午五月間口眼喎斜，灸聽會等三穴卽正。	中風	口眼喎斜	聽會
	右手足麻無力，灸百會・髮際等七穴得愈。	中風	手足麻；無力	百會；髮際
	09 : 未年八月間，氣塞涎上，不能語，金虎丹加膩粉服至四丸半。	中風	氣塞；涎上；不能語	-
	10 : 氣不通，涎不下，藥從鼻中出，魂魄飛揚，如墜江湖中，頃欲絕，灸百會・風池等左右共十二穴。	中風	氣不通；涎不下；藥從鼻中出；魂魄飛揚	百會；風池

Table 4. Terminology Dictionary Example of Acupuncture Points

대표어	관련 동의·유의어	대표어	관련 동의·유의어	대표어	관련 동의·유의어	대표어	관련 동의·유의어
巨陽	足巨陽	陽交	足陽	伏兔	外勾	地機	太陰鄧
肩髃	扁骨, 扁肩	陽池	手少陽	跗陽	付陽	攢竹	始光, 脊柱
肩井	鵠井	溫溜	逆注, 蛇頭, 溫留	臂臑	頭衝	脊中	脊俞
京門	氣俞	腰陽關	陽關	臂臑	頸衝	天突	玉戶
膏肓	膏肓俞	腰陽關	背陽關	飛揚	厥陽	天窗	窗籠
曲骨	屈骨, 屈骨端	涌泉	湧泉	絲竹空	目髎	天樞	大腸募
鳩尾	鶡軒·軒鶡	委中	郄中, 血郄	三間	少谷	聽會	聽訶
口禾髎	禾髎	陰交	少關	上腕	上管	衝陽	足陽明
歸來	溪穴	陰郄	手少陰郄	石門	丹田	太谿	足少陰
極泉	腋下	陰都	食宮	素髎	面王, 面正	太白	足太陰
氣海	下育, 脘腴	二間	間谷	水溝	人中	太淵	手太陰, 太泉
膩會	膩髎	耳和髎	和髎	水分	分水	通里	通理
腦戶	匝風, 合顱	長強	概骨, 氣之陰郄, 胸之陰郄	手五里	五里, 臂五里	風府	舌本
大巨	腋門	章門	脇髎, 季肋	水泉	水原泉	下廉	巨虛下廉
大杼	百勞	膻中	胸堂	承筋	脇腸	下髎	腰尻交
督俞	高蓋	睛明	淚空	承扶	皮部, 肉郄, 皮郄, 扶承	下腕	下管
瞳子髎	前關	照海	陰蹻	承泣	面髎	合谷	虎口
勞宮	手心, 掌中, 勞官, 五里	足竅陰	竅陰	神闕	臍中	血海	血郄
目窗	至營	足臨泣	臨泣	申脈	陽蹻	環跳	骹骨
百會	天滿, 巍上	足三里	三里	神門	兌衝, 手少陰	會陽	利機
步廊	步郎	足通谷	通谷	齒門	舌橫, 暗門		
腹結	腹屈, 腸窟	中都	中郄	液門	掖門		
復溜	伏白, 昌陽	中腕	中管, 胃管, 太倉	支溝	飛虎		

Table 5. Contingency Table of Observed Frequency

빈도	(병증)포함	미포함	합계
(경혈)포함	o_1	o_3	$o_1 + o_3$
미포함	o_2	o_4	$o_2 + o_4$
합계	$o_1 + o_2$	$o_3 + o_4$	n

구축해 두면 짧은 시간 내에 용어의 빈도를 측정할 수 있고 분석 대상이 달라지는 경우에도 재활용할 수 있어 효율적이다.

『침구자생경』에는 대표병증별로 치료법이 구분되어 있어 병증에 대한 용어사전은 따로 필요하지 않았다. 그러나 경혈의 경우에는 약 600여개가 나타나 경혈에 대한 용어사전을 구축하였다. 본 연구에서는 분석과 해석의 용이성을 위해 별혈(別穴)을 제외한 12 경맥과 임독맥의 정경혈만을 변수로 삼았다. 별혈은 종류에 비해 사용 빈도가 낮고 명칭 또한 단일화하기 어렵기 때문이다. 그러나 정경혈의 경우에도 동의어 혹은 유의어가 존재하기 때문에 용례에 따라 동의어·유의어를 통합해야 했다. 이를 위해 경혈들을 전수 조사하여 용어사전을 구축하였다. 연구에 사용된 동의어·유의어는 Table 4와 같다.

3. 분석 및 시각화

1) 연관성 측정 방법 설정: 텍스트 분석은 분석 대상이 되는 범주

Table 6. Contingency Table of Expected Frequency

빈도	(병증)있음	없음	합계
(경혈)있음	e_1	e_3	$e_1 + e_3$
없음	e_2	e_4	$e_2 + e_4$
합계	$e_1 + e_2$	$e_3 + e_4$	n

$$e_1 = (o_1 + o_3) \times (o_1 + o_2) / n$$

$$e_2 = (o_2 + o_4) \times (o_1 + o_2) / n$$

$$e_3 = (o_1 + o_3) \times (o_3 + o_4) / n$$

$$e_4 = (o_2 + o_4) \times (o_3 + o_4) / n$$

형 자료(categorical data)로서의 용어가 나타나는 빈도(frequency)를 관측하여 수치형 자료(numerical data)를 생성하는 것에서 출발한다. 본 연구에서는 치료대상과 치료방법, 즉 병증과 경혈 사이의 연관성을 측정하는데 목적을 두었기 때문에 용어의 빈도를 통해 병증-경혈 사이에 관찰빈도의 분할표(contingency table)를 Table 5와 같이 만들 수 있었다. 만약 중풍(中風)과 백회(百會)를 예로 든다면, 중풍을 치료 대상으로 하는 치료법 전체의 개수는 $o_1 + o_2$, 백회가 사용된 치료법 전체의 개수는 $o_1 + o_3$ 가 된다. 여기서 o_1 는 중풍 치료에 백회가 사용된 치료법 개수, 즉 함께 일어난 빈도수(co-occurrence frequency)가 된다.

연관성을 측정하는 데에는 여러 가지 방법들이 존재하지만, 전통의학 텍스트에 적합하다고 확인된 방법은 없다. 따라서 본 연구

에서는 벡터공간모델(vector space model)에 기반한 ‘거리측정법’과 확률모델(probabilistic model)에 기반한 ‘확률측정법’을 적용하였다.

벡터공간모델에서 거리 측정은 두 벡터 사이의 거리를 측정한다. 따라서 관찰빈도로부터 기대빈도 분할표를 Table 6과 같은 형태로 추가로 도출하였고, 이 두 분할표의 인자 값으로부터 각각 \vec{o} 와 \vec{e} 두 벡터를 추상화 하였다.

$$\vec{o} = (o_1, o_2, o_3, o_4), \vec{e} = (e_1, e_2, e_3, e_4)$$

벡터공간모델에서의 거리측정에는 매우 다양한 방식이 사용되지만, 기본적으로는 유클리디언 거리(Euclidean distance, 이하 D_e)를 기본으로 적당한 가중치를 부여하거나 공식에 변형을 꾀한 것들이다. 본 연구에서는 D_e 이외에 가중유클리디언거리(Weighted Euclidean distance, 이하 D_{we})¹²⁾, 크레머의 V(Cramér's V, 이하 D_{cv}), 캔베라거리(Canberra distance, 이하 D_{cb})를 적용하여 병증과 경혈의 연관성을 측정하였다. 각각의 공식은 Table 7과 같다.

Table 7. Measure Methods of Association with Vector Space Model

방법	적용공식
Euclidean distance	$D_e(o, e) = \sqrt{\sum_{i=0}^N (o_i - e_i)^2}$
Weighted Euclidean distance	$D_{we}(o, e) = \sqrt{\sum_{i=0}^N (o_i - e_i)^2 \frac{1}{o_i + e_i}}$
Cramér's V	$D_{cv}(o, e) = \sqrt{\sum_{i=0}^N \frac{(o_i - e_i)^2}{e_i} \frac{1}{n}}$
Canberra distance	$D_{cb}(o, e) = \sum_{i=0}^N \frac{ o_i - e_i }{ o_i + e_i }$

한편, 본 연구에서는 확률모델을 적용하여 관측된 관찰빈도가 나타날 확률을 계산하는 방법으로 병증과 경혈의 상관성을 측정하였다. 연관강도(Association strength, 이하 P_{as})는 두 사건이 동시에 일어날 확률에 대해 독립이라고 가정했을 때와 실제로 관측했을 때의 비를 측정하는 방법으로, 연관성분석(association rules)에서의 lift, 정보검색(information retrieval) 용어 기중치(term weighting) 방법으로서의 tf-idf와 동일한 값이다. P_{as} 는 지수적인 분포를 보이기 때문에 본 연구에서는 log scale을 적용하여 보정된 값을 사용하였다.

피셔정확검정(Fisher's exact test, 이하 P_{ft})은 본래 표본수가 적은 범주형 자료의 연관성 검증에 사용되는 방법으로, 해당 관찰빈도가 나타날 확률을 추정하는 방법이다. 본 연구에서는 통계적 유의성이 아니라 관찰빈도가 나타날 확률을 측정한다는 관점에서 이 방법을 사용하였다. P_{ft} 는 본래 값이 작을수록 우연히 나오기 어려운 경우가 되므로, 다른 거리 공식과의 비교를 위해 1에서 빼낸 값으로 보정하여 연구에 활용하였다.

자카드거리(Jaccard distance, 이하 D_{jc})는 범주형 자료 사이의 거리를 측정하는 공식으로 실제로 동시에 일어난 경우(곱사건)의 수를 사건이 일어난 전체 경우(합사건)의 수로 나누어 준 값으로, 곱사건의 확률을 의미하는 수치이다. 각각의 공식은 Table 8과 같다.

2) 시각화 방법 선정: 측정 결과는 수치형태로 나타나므로 같은 병증에 사용된 경혈들을 서로 비교하거나, 하나의 경혈이 사용된 병증들을 비교하기 위해서는 이에 적합한 시각화 방법이 필요하다. 병증과 경혈이라는 범주형 자료, 텍스트 상 용어의 빈도와 연관성 측정값이라는 수치형 자료를 모두 포괄하기 위해서는 다양한 변수들을 하나의 차트에 담을 수 있는 방법이어야 한다. 이러한 기준을 모두 만족하는 것이 트리맵(tree map)이다. 트리맵은 영역의 구분으로 범주형 자료를 표현할 수 있고, 영역의 크기와 색깔로 2가지 이상의 수치형 데이터를 나타낼 수 있다. 또 범주형 자료들을 크게

Table 8. Measure Methods of Association with Probabilistic Model

방법	적용공식
Association strength	일반공식 $p_{as}(o) = \frac{o_1 / n}{(o_1 + o_2) / n \cdot (o_1 + o_3) / n} = \frac{o_1 n}{(o_1 + o_2)(o_1 + o_3)}$ 보정공식 $P_{as}(o) = \ln(p_{as}(o))$
Fisher's exact test	일반공식 $p_{ft}(o) = \frac{(o_1 + o_2)! (o_3 + o_4)! (o_1 + o_3)! (o_2 + o_4)!}{o_1! o_2! o_3! o_4! n!}$ 보정공식 $P_{ft}(o) = 1 - p_{ft}(o)$
Jaccard distance	$D_{jc}(o) = \frac{ x_1 }{ x_1 + x_2 + x_3 }$

묶어 표현할 수 있다는 장점도 있다.

4. 분석 방법의 적합성 검토

마지막으로 병증-경혈 연관성 측정에 사용된 측정법들 가운데 가장 적합한 방법이 무엇인지 확인하였다. 분석 결과의 검정에는

Table 9. Observed or Expected Frequency of Acupuncture Points for Stroke

경혈	관측빈도				기대빈도			
	o_1	o_2	o_3	o_4	e_1	e_2	e_3	e_4
百會	13	106	49	5684	1.261	117.739	60.739	5728.657
肝俞	6	113	48	5685	1.098	117.902	52.902	5744.498
曲池	6	113	41	5692	0.956	118.044	46.044	5758.377
足三里	5	114	65	5668	1.423	117.577	68.577	5712.837
風池	5	114	35	5698	0.813	118.187	39.187	5772.273
心俞	5	114	35	5698	0.813	118.187	39.187	5772.273
大椎	4	115	21	5712	0.508	118.492	24.492	5802.107
肩髃	4	115	9	5724	0.264	118.736	12.736	5826.029
腎俞	3	116	48	5685	1.037	117.963	49.963	5750.444
照海	3	116	43	5690	0.935	118.065	45.065	5760.362
間使	3	116	37	5696	0.813	118.187	39.187	5772.273
肺俞	3	116	37	5696	0.813	118.187	39.187	5772.273
風市	3	116	11	5722	0.285	118.715	13.715	5824.033
崑崙	2	117	50	5683	1.057	117.943	50.943	5748.462
足臨泣	2	117	43	5690	0.915	118.085	44.085	5762.346
解谿	2	117	28	5705	0.610	118.390	29.390	5792.154
液門	2	117	23	5710	0.508	118.492	24.492	5802.107
大腸俞	2	117	18	5715	0.407	118.593	19.593	5812.068
肩井	2	117	13	5720	0.305	118.695	14.695	5822.038
頸車	2	117	12	5721	0.285	118.715	13.715	5824.033

Table 10. Measure of Association Values between Acupuncture Points and Stroke

경혈	벡터공간모델				확률모델		
	D_e	D_{we}	D_{cv}	D_{cb}	P_{as}	P_{ft}	D_{fc}
百會	49.068	3.422	0.139	0.987	2.333	1.000	0.077
肝俞	60.101	2.009	0.063	0.766	1.698	0.999	0.036
曲池	66.950	2.108	0.069	0.811	1.837	1.000	0.038
足三里	45.263	1.523	0.041	0.603	1.256	0.986	0.027
風池	74.627	1.951	0.063	0.801	1.816	0.999	0.032
心俞	74.627	1.951	0.063	0.801	1.816	0.999	0.032
大椎	90.310	1.931	0.067	0.874	2.063	0.999	0.029
肩髃	102.234	2.208	0.098	1.073	2.717	1.000	0.031
腎俞	65.533	1.177	0.028	0.520	1.062	0.915	0.018
照海	70.452	1.258	0.031	0.563	1.165	0.934	0.019
間使	76.367	1.358	0.035	0.618	1.305	0.953	0.019
肺俞	76.367	1.358	0.035	0.618	1.305	0.953	0.019
風市	102.142	1.864	0.070	0.957	2.355	0.997	0.023
崑崙	65.482	0.823	0.017	0.327	0.637	0.714	0.012
足臨泣	72.370	0.938	0.020	0.396	0.782	0.767	0.012
解谿	87.187	1.201	0.028	0.570	1.187	0.876	0.014
液門	92.143	1.296	0.032	0.640	1.370	0.909	0.014
大腸俞	97.108	1.397	0.037	0.720	1.593	0.939	0.015
肩井	102.081	1.505	0.044	0.813	1.881	0.964	0.015
頸車	103.076	1.528	0.046	0.834	1.950	0.968	0.015

다양하고 복잡한 방법들이 있지만, 보통 전체 데이터를 학습데이터와 테스트데이터로 나누고 전자로 모델을 만든 다음은 후자로 적합성을 테스트하는 것이 일반적이다. 하지만 본 연구에 사용된 데이

터 수는 이런 절차를 밟기에 충분하지 않을 뿐 아니라 연구의 목적이나 모델링에 있지 않기 때문에 이 방법을 적용할 수 없었다. 따라서 본 연구에서는 텍스트에 등장하는 침구 치료법들에 측정값들을 반

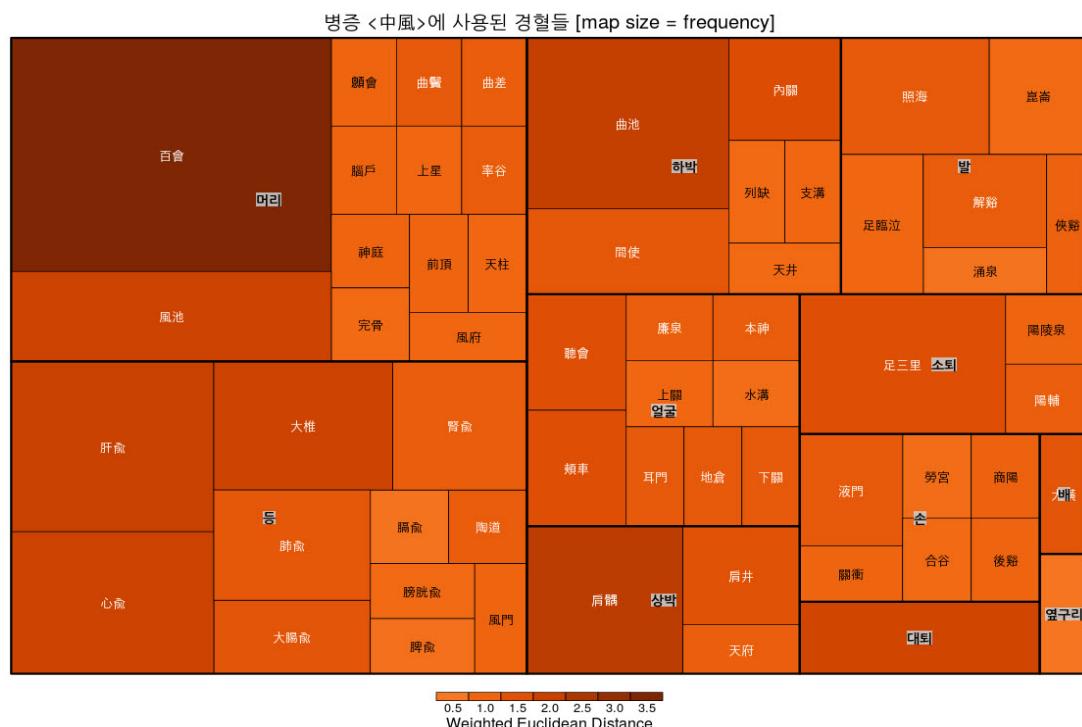
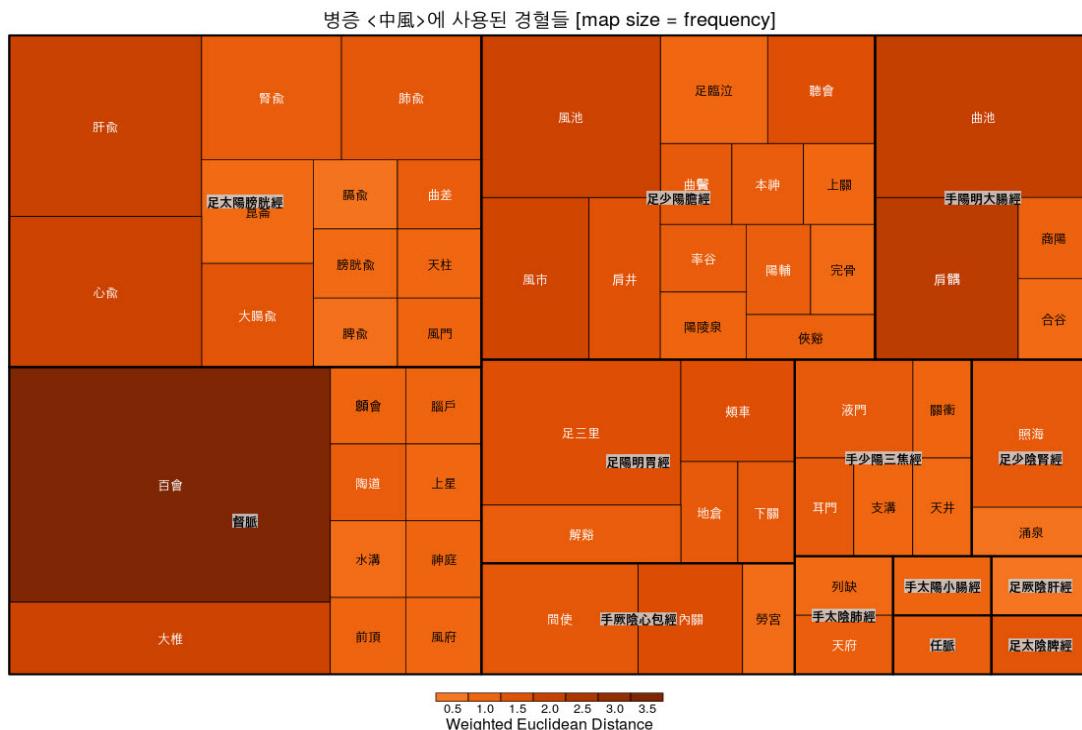


Fig. 2. Tree Map of Acupuncture Treatment for Stroke Using Weighted Euclidean Distance.

대로 대입하여 해당 치료법이 쓰였을 병증을 추론한 뒤에, 이를 실제 텍스트 병증과 비교하여 단순히 어떤 방법이 가장 예측을 잘 하였는지 살펴보았다.

결과

『침구자생경』을 디지털 텍스트로 입력하고 다중코드자와 이체

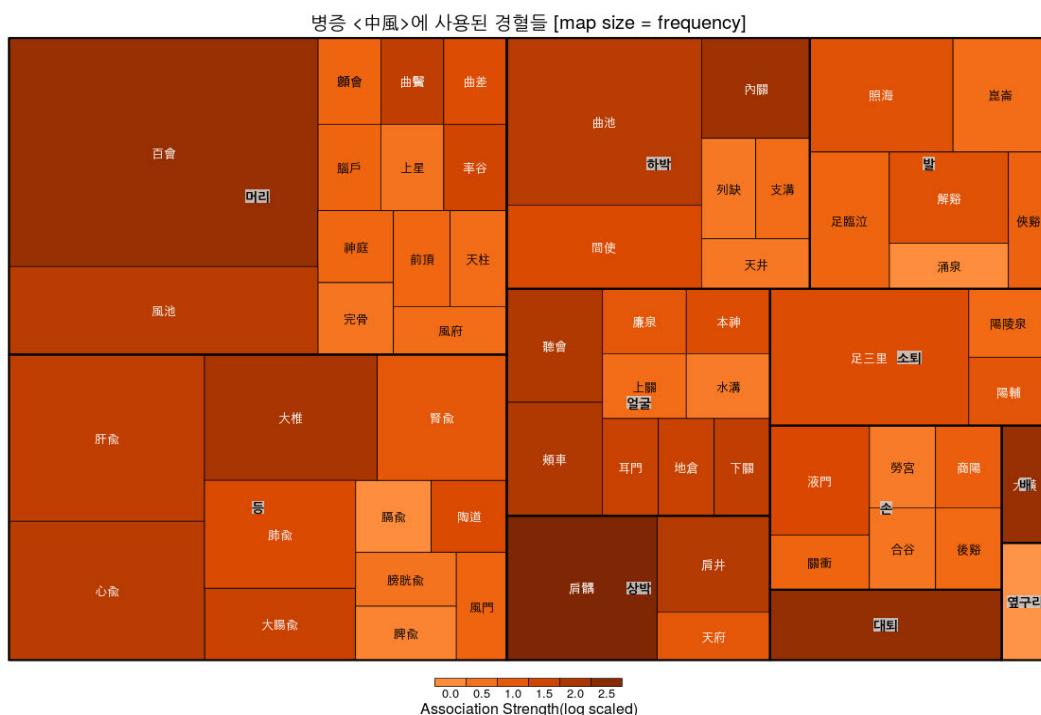
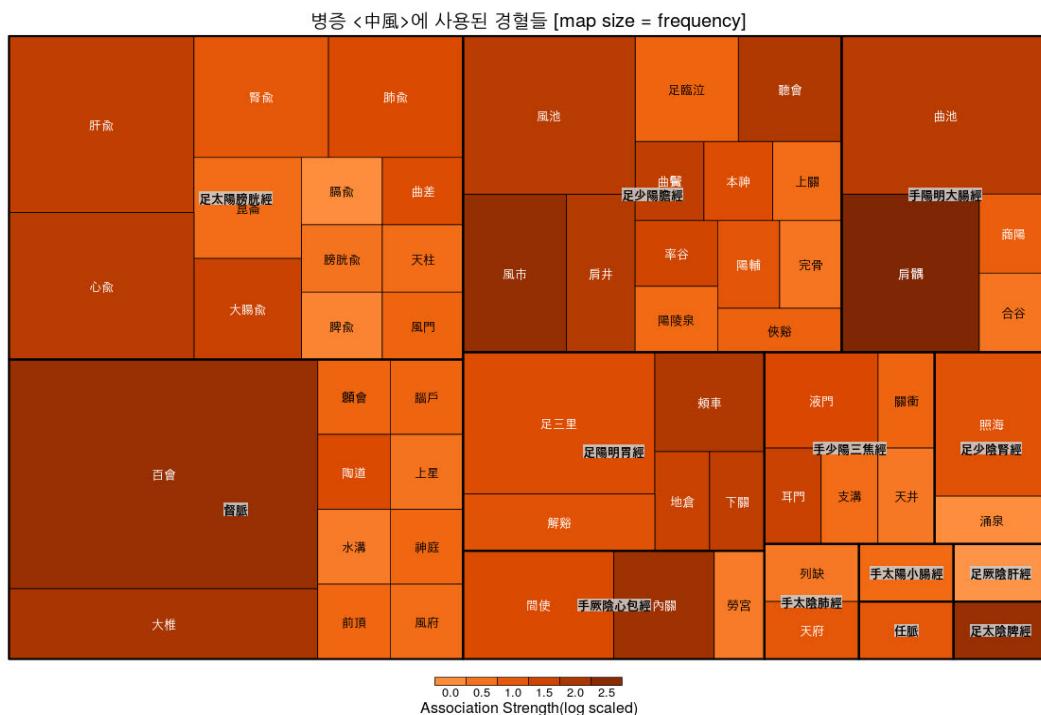


Fig. 3. Tree Map of Acupuncture Treatment for Stroke Using Association Strength.

Table 11. Example of Predictive Simulation Result for Stroke(中風) Treatment

Test data		Major Symptom	D_e	D_{we}	D_{av}	D_{cb}	P_{as}	P_{ft}	D_{ic}
Treatment									
率谷	中風	嘔吐	嘔吐	腦痛	腦痛	腦痛	中風*	中風*	
第二椎 ; 第五椎上	中風	-	-	-	-	-	-	-	
三里 ; 絶骨	中風	脚氣	脚氣	脚氣	脚氣	脚氣	癲狂	癲狂	
百會 ; 耳前髮際 ; 肩井 ; 風市 ; 三里 ; 絶骨 ; 曲池	中風	中風*	中風*	脚氣	脚氣	脚氣	癲狂	中風*	
百會 ; 曲鬢 ; 肩髃 ; 曲池 ; 風市 ; 足三里 ; 絶骨	中風	中風*	中風*	脚氣	中風*	中風*	癲狂	中風*	
百會 ; 風池 ; 大椎 ; 肩井 ; 曲池 ; 間使 ; 足三里	中風	中風*	中風*	中風*	中風*	中風*	癲狂	中風*	
聽會	中風	耳鳴	耳鳴	耳鳴	耳鳴	耳鳴	癲狂	癲狂	
百會 ; 髮際	中風	中風*	中風*	中風*	中風*	中風*	心驚恐	中風*	
(치료 경혈 없음)	中風	-	-	-	-	-	-	-	
百會 ; 風池	中風	中風*	中風*	中風*	中風*	中風*	目眩	目眩	

*Major Symptom과 부합하는 시뮬레이션 결과들

Table 12. Predictive Simulation Matching Frequency

Group	측정 방법	일치 빈도	일치율
벡터공간모델	Euclidean distance	865	20.78%
	Weighted Euclidean distance	1012	24.32%
	Cramér's V	826	19.85%
	Canberra distance	963	23.14%
확률모델	Association strength	886	21.29%
	Fisher exact test	609	14.63%
	Jaccard similarity	283	6.80%

자를 통합한 뒤 전처리를 거친 결과 전체 4162개 침구 치료법을 얻을 수 있었다. 이 치료법에는 중복을 포함하여 정경혈이 모두 5852회(n) 사용되었다. 즉, 하나의 치료법에 약 1.4개의 정경혈이 사용된 셈이다. 이 치료법에는 모두 342개의 정경혈이 사용되었으며, 허손(虛損) · 노채(勞瘵) · 소갈(消渴) · 중풍(中風) 등 모두 194개 대표병증(major symptom)을 포함하고 있었다. 설명의 편의를 위해 병증을 중풍(中風)으로 한정해 결과를 서술하고자 한다. 지면의 한계로 인해 전체 결과는 온라인 공개 페이지(<http://goo.gl/UqPPZT>)로 대신한다.

1. 연관성 측정 결과

『침구자생경』 전체 4162개 침구 치료법에서 정경혈은 모두 5,852회(중복포함; n) 나타나는데, 이 가운데 중풍(中風)에 사용된 경혈들의 빈도는 모두 119회(o_1+o_2), 백회가 사용된 경혈 빈도는 62회(o_1+o_3)였다. 또 중풍 치료에 백회가 사용된 치료법은 13회(o_1)였다. 이와 같은 방식으로 중풍에 사용된 경혈에 대한 관찰빈도 및 기대빈도 분할표(contingency table) 수치를 얻을 수 있었다(Table 9 참조, 빈용된 상위 20개 경혈만 표로 정리함. 소수점 이하 3째자리까지 표시, 이하 같음).

이렇게 병증과 경혈의 관계를 빈도를 기준으로 수량화 시키면,

병증-경혈 쌍에 대해 하나의 관찰빈도 벡터와 다른 하나의 기대빈도 벡터가 도출된다. 이 두 벡터의 거리를 앞의 방법으로 연산하여 거리측정치를 얻을 수 있고, 관찰빈도를 통해 해당 관찰빈도가 나타날 가능성을 측정하여 확률측정치를 얻을 수 있었다(Table 10 참조).

2. 시각화 결과

이 결과에는 ‘병증’ 및 ‘경혈’이라는 2가지 범주형 데이터와, ‘병증-경혈 동시 관찰 빈도(o_1)’ 및 ‘병증-경혈 연관성 측정값’이라는 2가지 수치형 데이터가 있다. 또 경혈의 경우 소속 경락이나 위치한 인체 부위로 크게 분류할 수 있기 때문에 이러한 모습을 종합적으로 표현해 줄 수 있어야 한다. 이러한 점을 고려하여 중풍에 사용된 경혈 전체 상황을 트리맵으로 나타낼 수 있었다.

트리맵에서의 면적은 병증-경혈이 동시에 관찰된 빈도(o_1)이며, 색의 진하기는 연관성 측정치를 나타낸다. Fig. 2는 D_{we} 로 도출된 연관성 측정치를, Fig. 3은 P_{as} 로 도출된 연관성 측정치를 기준으로 그린 트리맵이다. 각각의 첫 번째 그림은 경혈이 속한 경락을 중심으로 표현한 것이고, 두 번째 그림은 경혈이 위치한 인체 부위를 중심으로 표현한 것이다. 경락을 중심으로 보면, 음경맥에 소속된 경혈은 비교적 적게 사용된 반면, 독맥(督脈), 족삼양경(足三陽經), 수양명대장경(手陽明大腸經) 등 양경맥에 소속된 경혈이 많이 사용되었다는 사실을 알 수 있다. 인체 부위별로 본 트리맵에서도 머리와 등의 경혈들이 자주 사용되었다는 것을 다시 한 번 확인할 수 있다.

D_{we} 로 측정한 연관성 결과를 보면, 백회와 견우의 연관성이 가장 뚜렷하며, 이어 곡지, 간수, 풍지가 비슷한 정도로 나타난다. P_{as} 측정치를 보면 견우와의 연관성이 가장 높으며, 풍지와 백회가 뒤를 잇는다. 특이한 것은 D_{we} 에서 두각을 보이지 않았던 대획이나 내관과 같은 경혈들의 연관성이 높게 나타났다는 점이다. 측

정 결과에 대해서는 다음 절에서 더 자세히 고찰하겠다.

3. 분석 방법의 적합성 검토 결과

마지막으로 간단한 시뮬레이션을 통해 분석 방법들의 적합성을 확인해 보았다. 중풍의 예를 보면, 중풍을 치료하는 첫 번째 치료법에는 솔곡(率谷) 1개 경혈이 사용되었다. ‘솔곡’을 사용하는 치료법이 대표병증 가운데 두통(頭痛)을 치료하는 처방이라고 가정했을 때, D_{we} 기준으로 두통과 솔곡의 연관성 측정치로 1.065를 얻을 수 있다. 하지만 구토(嘔吐)를 치료하는 처방이라고 가정하면 D_{we} 기준으로 구토와 솔곡의 연관성 측정치로 1.460를 얻을 수 있다. 따라서 D_{we} 기준으로 보면, 솔곡은 두통과 구토 가운데 구토를 치료하는 처방이라고 추론할 수 있다. 솔곡이 194개 대표병증에 치료법이라고 각각 가정하고 솔곡과 해당 대표병증과의 연관성 측정치를 모은 뒤, 이 가운데 가장 높은 결과치를 나타내는 병증을 뽑아내

면 솔곡 1개 경혈을 사용한 치료법이 치료할 것으로 보이는 병증을 도출해 낼 수 있다. 이 과정에서 경혈과 병증 사이의 연관성 측정치가 변수가 되기 때문에 연관성 측정 방법에 따라 도출되는 병증들은 서로 달라진다. 솔곡의 경우 D_e 와 D_{we} 는 구토, D_{cv} , D_{cb} , P_{as} 는 뇌통(腦痛), P_{ft} 와 D_{jc} 는 중풍(中風)을 치료할 가능성이 가장 높다는 결론을 보여주었다.

치료법에 사용된 경혈이 2개 이상일 때에는 병증과 각 경혈과의 연관성 측정값의 평균을 가지고 이 값이 가장 높은 병증을 해당 치료법이 치료할 병증으로 보았다. 백회(百會)와 풍지(風池)가 쓰인 10번재 치료법을 D_{we} 를 기준으로 살펴보면, 중풍(中風) 치료법이라고 가정했을 때 2.727, 심기(心氣)라면 1.004, 목루출(目淚出)이라면 0.973, 면종(面腫)이라면 0.888 값을 얻게 된다. 따라서 D_{we} 를 기준으로 보면, 해당 침구법은 중풍을 치료하는 처방일 가능성이 가장 높다고 할 수 있다. 백지와 풍지는 P_{ft} 와 D_{jc} 가 목현(目

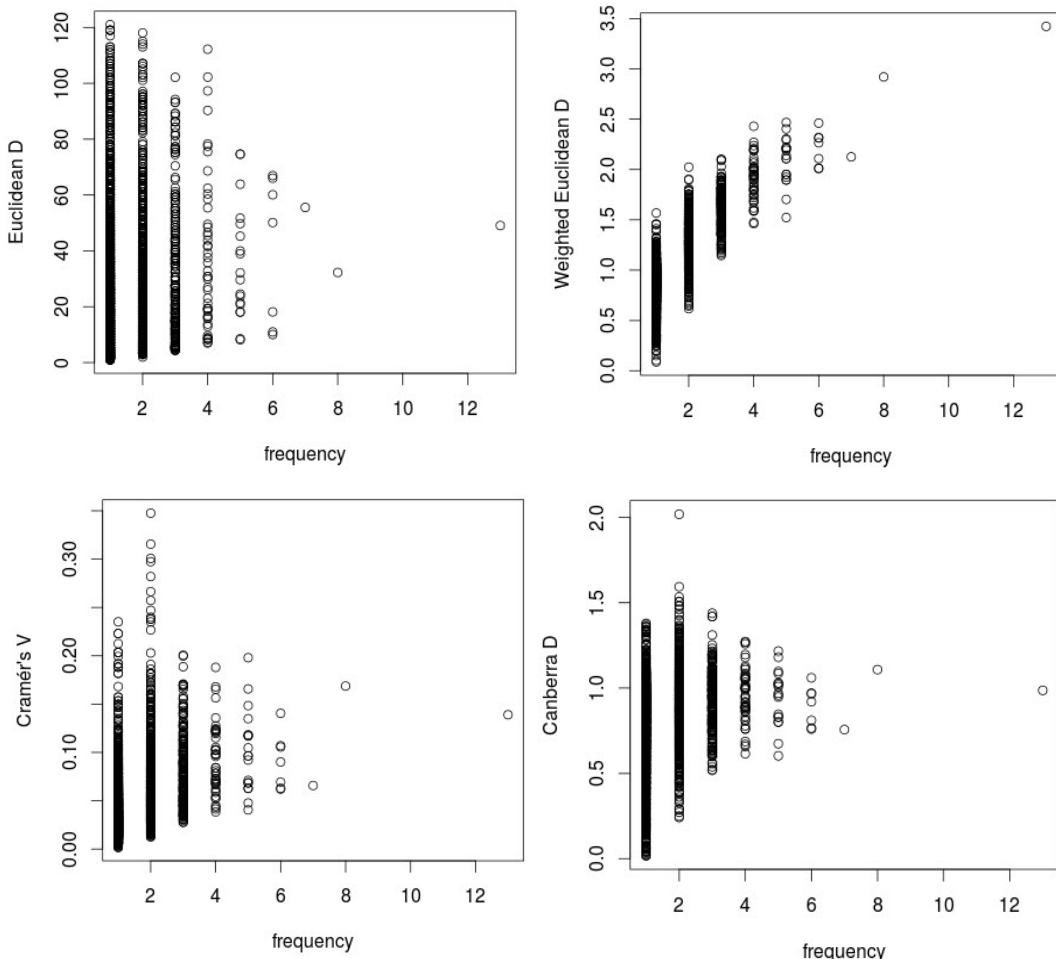


Fig. 4. Relations between Measure Method of Association Based on Vector Space Model and Co-occurrence Frequency.

眩)이라고 결론지은 것 이외에는 모든 측정법에서 중풍(中風)을 치료하는 처방일 것이라는 결론을 보였다. 앞에서 예로 든 중풍 치료 침구법에 대한 측정 방법 별 예측 결과값은 Table 11과 같다.

이런 방식으로 4162개 치료법 전체를 예측하고 그 결과가 원래 텍스트와 일치하였는지를 살펴 Table 12와 같은 결과를 얻을 수 있었다. D_{we} 이 추측 결과와 실제 텍스트 결과가 1012개 일치하여 가장 높은 일치율을 보였으며, 이어 D_{cb} , P_{as} 등이 높게 평가 되었다. 시뮬레이션 결과는 복수의 값을 선택할 수 없지만 실제 텍스트에서는 동일한 경혈 조합이 다수의 병증에 사용되는 경우가 많기 때문에 일치율이 저조하다고 할 수는 없다.

고 찰

병증에 사용된 경혈의 빈도와 연관성 측정치의 관계는 빈도와 분명한 상관성을 보이는 경우와 그렇지 않은 경우로 나눌 수 있었

다. 전자의 대표적인 경우는 $D_{we}(0.806)$ 이다. D_{we} 의 경우 빈도가 높아질수록 연관성이 함께 늘어나는 모습을 보인다. $P_{ft}(0.409)$ 의 경우에도 빈도가 높아질수록 값이 높아지는 경향을 보이지만, 측정치가 극단적으로 치닫고 있어 결과를 신뢰하기 어렵다는 것을 알 수 있었다(Fig. 4 및 Fig. 5 참조).

이에 반해 후자에 속하는 $D_{jc}(0.467)$, $D_{cv}(0.425)$, $D_{cb}(0.231)$, $P_{as}(0.154)$ 의 경우에는 모두 빈도와 낮은 상관성을 보였으며, 빈도가 높을수록 측정치의 분포 범위가 좁아지는 공통점을 보였다. 특히 P_{as} 의 경우 빈도에 민감하지 않도록 의도적으로 고안된 방법인데^[13], 본 연구에서도 매우 낮게 측정되었다. 빈도와 가장 상관관계가 낮은 $D_e(0.107)$ 의 경우에는 빈도와 거의 무관하게 측정치가 고르게 나타났다(괄호 안의 숫자는 빈도와의 상관계수 correlation coefficient, 전체는 Table 13 참조).

병증-경혈의 실제 사용 빈도와 비교적 적게 관련되어 있는 D_{cv} , D_{cb} , P_{as} , D_{jc} 4가지 방법 사이에는 Fig. 6과 같이 양의 상관관계가 나타났다. 즉, P_{as} 와의 상관계수는 D_{cb} 가 가장 높은 0.954,

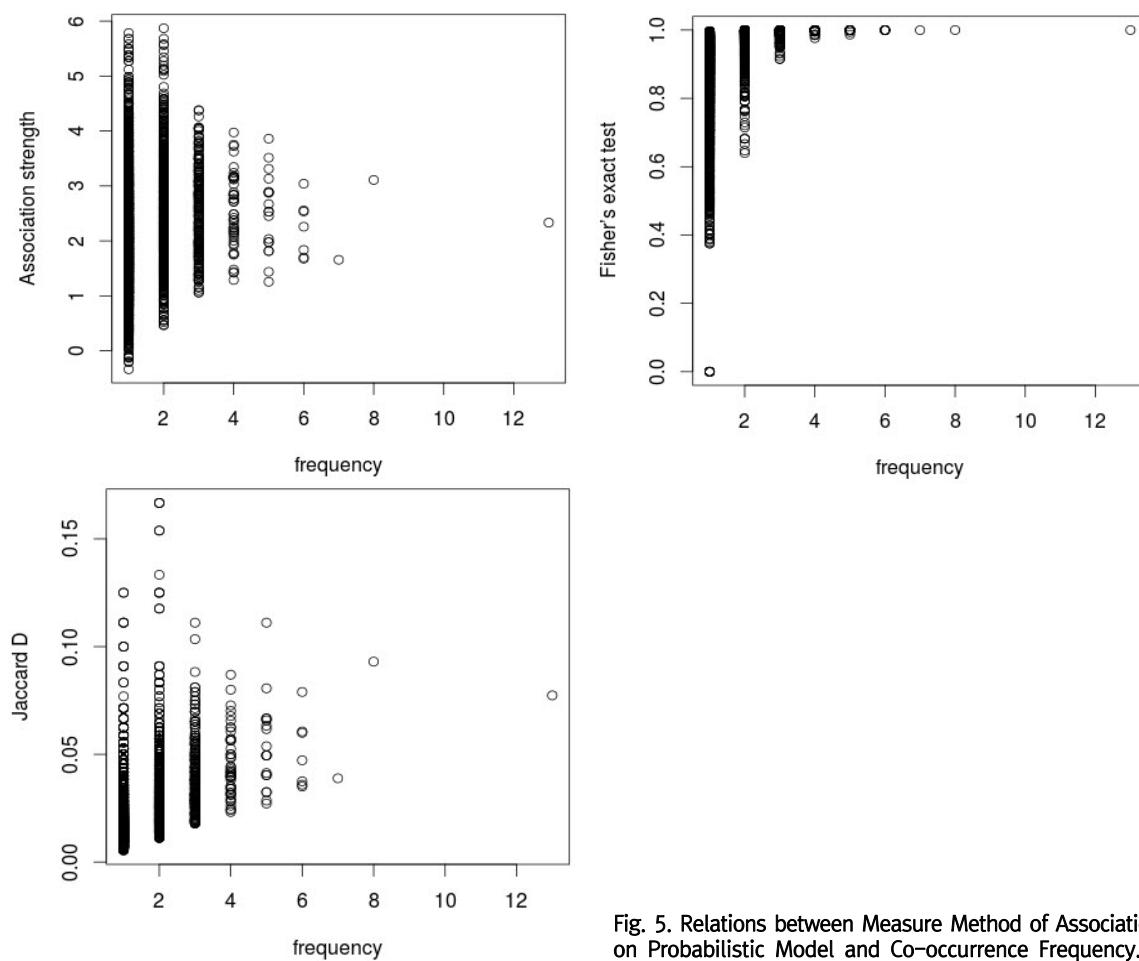


Fig. 5. Relations between Measure Method of Association Based on Probabilistic Model and Co-occurrence Frequency.

D_{cv} 가 0.882, D_{jc} 가 0.794로 나타났다. 따라서 이들은 비록 연산 방식이 다르지만 대체로 유사한 결과를 가리킨다고 할 수 있다. 특히 P_{as} 와 D_{cb} 는 적합성 검토 결과에서도 상대적으로 높은 일치율을 나타냈고 상관계수도 분석 방법 전체에서 가장 높은 수치를 나타냈다.

D_{jc} 와 P_{ft} 는 적합성 측정 결과가 낮은 방법들로서, 분석 대상과 방법이 서로 맞지 않는 경우이다. 텍스트에서 특정 용어의 존재 유무는 이진형 자료(binary data)에 해당한다. 하지만 해당 경혈이

사용되지 않았다고 해서 써서는 안된다고 볼 수 없으므로 고의서에서 병증과 치료방법을 통해 도출된 이진형 자료는 성별과 같은 이진형 자료와 달리 대칭적이라고 할 수 없다. D_{jc} 는 대칭성이 확보된 이진형 자료에 적합한 방법이기 때문에 이런 경우에는 사용하기 어렵다. D_{jc} 의 예측률이 가장 낮게 나온 것은 이와 무관하지 않다.

P_{ft} 의 경우에는 해당 관찰빈도가 나타날 확률을 연산하는 측정 방법이다. 따라서 어떤 병증에 특정 경혈이 많이 사용된 경우 뿐만 아니라 이례적으로 적게 사용된 경우에도 그와 같은 현상이 일어날

Table 13. Correlation Coefficient Value between Measure Methods of Association

상관계수	Frequency	D_e	D_{we}	D_{cv}	D_{cb}	P_{as}	P_{ft}	D_{jc}
Frequency	1.000	0.108	0.806	0.425	0.232	0.155	0.410	0.467
D_e	0.108	1.000	0.234	-0.131	-0.166	-0.240	-0.164	-0.329
D_{we}	0.806	0.234	1.000	0.728	0.676	0.585	0.722	0.649
D_{cv}	0.425	-0.131	0.728	1.000	0.812	0.883	0.646	0.937
D_{cb}	0.232	-0.166	0.676	0.812	1.000	0.955	0.893	0.696
P_{as}	0.155	-0.240	0.585	0.883	0.955	1.000	0.793	0.794
P_{ft}	0.410	-0.164	0.722	0.646	0.893	0.793	1.000	0.599
D_{jc}	0.467	-0.329	0.649	0.937	0.696	0.794	0.599	1.000

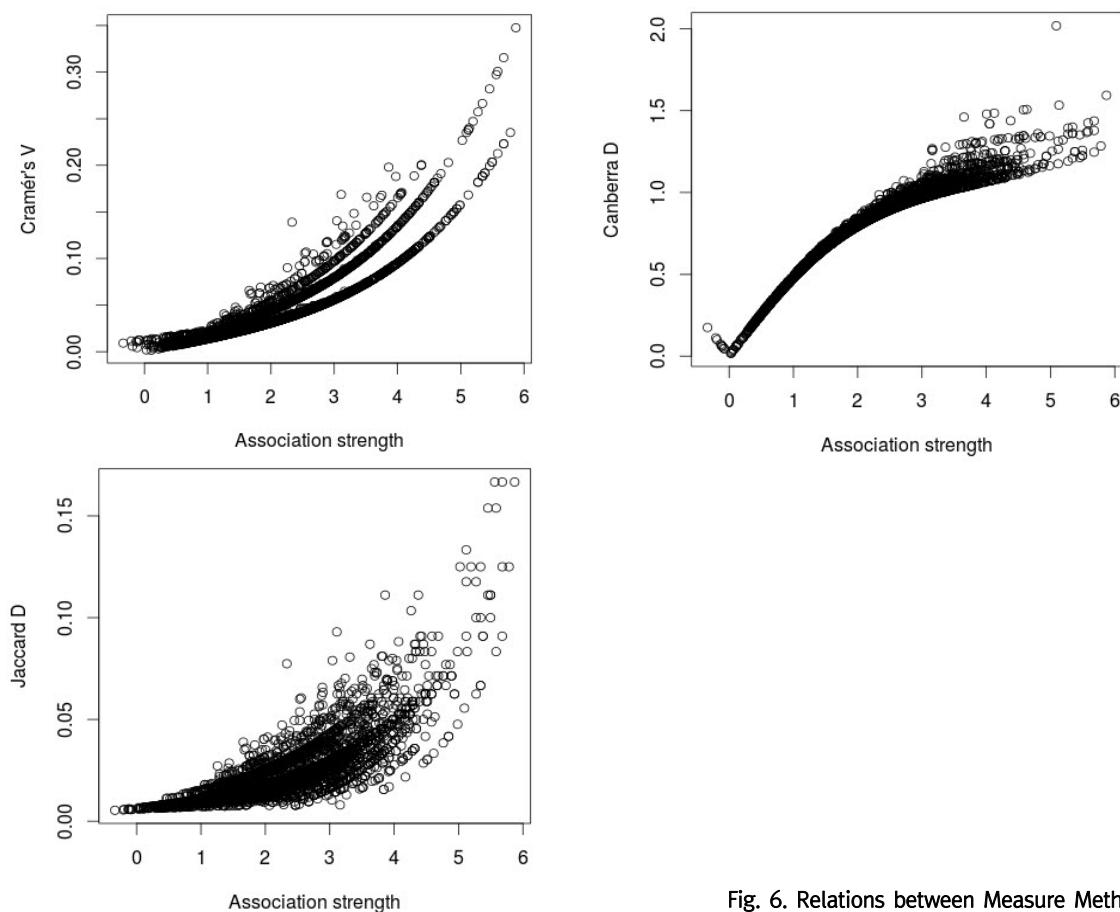


Fig. 6. Relations between Measure Methods of Association.

Table 14. Characteristic of Measure Method of Associations

연관성 측정법 구분	벡터공간모델	확률모델	비고
빈도와 유관	Weighted Euclidean distance	-	이미 알려진 사실을 요약할 때
빈도와 무관	Canberra distance	Association strength	알려지지 않은 새로운 의미를 포착할 때

확률은 낮아지므로 P_{ft} 측정치는 양자 모두에서 높게 나타날 수 있다. 따라서 빈도를 기반으로 한 연관성 측정에는 적합하다고 할 수 없다.

적합성 결과 20% 이상의 일치율을 보인 D_{we} , D_e , D_{cb} , P_{as} 사이에도 차이점은 존재한다. D_{we} 는 D_e 에 가중치를 부여한 방식으로 D_e 가 각 인자들 사이에 격차가 크게 벌어질수록 결과가 왜곡된다는 단점을 보완한 방법이다. 따라서 본 연구와 같이 빈도 측정치 사이에 격차가 큰 경우에는 D_e 보다 D_{we} 가 더 적합하다고 할 수 있다.

분석 목적에 따라서도 다른 측정법을 생각할 수 있다. 의학 지식의 습득 과정에서 보통 자주 반복되어 사용되는 치료법을 해당 질환과 관련이 높다고 인식하기 마련이다. 따라서 연관성 측정치가 빈도와 상관성을 보이는 D_{we} 는 치료에 대한 일반적인 통념을 잘 반영해 준다. 『침구자생경』 분석결과에서도 전반적으로 질환에 연관성이 높을 것이라고 예상되는 경혈에 대해 D_{we} 는 높은 측정치를 보여주었으며, 트리맵을 이용한 시각화 결과에서도 이를 확인할 수 있다. 반면 P_{as} 나 D_{cb} 는 사용 빈도가 높지 않은 의외의 결과들을 보여주는 경우가 많았다. 이것은 사용 빈도가 높지 않은 경우에도 다른 질환보다 더 많이 사용되었기 때문이다. 따라서 이러한 방법들은 의료인의 일반적인 통념으로 파악하지 못한 새로운 의미들을 발견해내거나, 특수한 조건 하에서 중점적으로 사용된 치료법을 추출하는 목적으로 유리할 수 있다. 본 연구에 사용된 측정법의 성격을 요약하면 Table 14와 같다.

결 론

동아시아 전통의학 고문헌에는 과거의 치료경험들이 집적되어 있다. 전통사회에서 고의서는 의학을 배우고 치료하기 위해 숙고하고 깨우쳐야 할 대상이었다. 하지만 전통사회에 만들어진 고의서의 양이 방대하여 오늘날 이를 효과적으로 이해하고 활용하는 데에 적지 않은 어려움이 따른다.

고의서를 방대한 치료경험이 담긴 데이터로 바라보았을 때, 치료대상이 되는 질병과 치료방법이 되는 약재 혹은 경혈의 연관성을 측정할 수 있다면 방대한 분량의 고의서의 내용을 효과적으로 이해

할 수 있을 것이다. 이를 위해서는 먼저 고의서의 전산화, 치료대상과 치료방법의 연관성 측정 방법의 개발, 그리고 그 결과를 표현하는 적합한 시각화 방법의 고안이 필수적이다.

본 연구에서는 침구 치료경험이 4천여건 수록된 『침구자생경』이라는 텍스트를 기반으로 치료대상과 치료방법, 즉 병증과 경혈의 상관성을 측정해 보았다. 이 과정은 앞의 3가지 선행 과제들에 대한 한 가지 예시에 해당한다. 이 과정에서 얻은 지견들을 정리하면 다음과 같다.

첫째, 전통의학 고의서 텍스트 분석에서, 해당 텍스트의 전산화는 필수적이다. 한자 텍스트의 경우 다중코드자 및 이체자를 통합해야 한다.

둘째, 텍스트 분석에서 수치형 데이터는 특정 용어의 빈도 측정 값이 되기 때문에 분석에 앞서 자연어로 구성된 텍스트를 반구조화 혹은 구조화해야 하며, 측정하고자 하는 용어에 대해 동의어 및 유의어를 아우르는 용어집이 구축되어야 한다.

셋째, 다양한 연관성 측정 방법 가운데, 분석 목적에 맞는 측정법을 사용해야 한다. 본 연구에서 7가지 측정법을 적용한 결과, 용어 출현 빈도에 민감하여 이미 알려진 사실들을 요약해 내는 목적에는 가중유클리디언거리(Weighted Euclidean distance)가 유용하였으며, 용어 출현 빈도에 민감하지 않아 새로운 아이디어를 얻어내는 목적에는 캔베라거리(Canberra distance), 연관강도(Association strength) 등의 방법이 유용했다.

넷째, 분석 결과의 시각화 방법은 분석 대상이 되는 변수의 종류와 성격을 고려하여 선택해야 한다. 본 연구에서 수행한 병증-경혈 연관성의 경우, 카테고리형 변수와 수치형 변수가 적절히 섞여 있어 트리맵을 활용하여 이를 표현할 수 있었다.

감사의 글

This Research was supported by the “KIOM Customization of TKM Knowledge for R&D(K16670)” funded by Korea Institute of Oriental Medicine.

References

1. Taki-mototsugu. isekiko. beijing:xueyuanchubanshe. 2007 ; i
2. Tang-shenwei. chongxiuzhenhejinshizhengleibeiyongcaoyao. beijing:zhonggouzongyiyaochubanshe. 2012 ; 1.
3. Harananyou. keyikeciyikayi. beijing:xueyuanchubanshe. 2008 ; vii
4. Oh JH. The Achievements and Problems of Computerization of Korean Medical Classics. *J Korean Medical History*. 2015 ; 28(1) : 111-9.
5. Park IS, Jung WM, Lee YS, Hahm DH, Park HJ, Chae YB. Characterization of Five Shu Acupoint Pattern in Saam Acupuncture Using Text Mininig. *Korean J Acupunct* 2015 ; 32 : 66-74.
6. Lee T, Jung WM, Lee IS, Lee YS, Lee H, Park HJ, Kim N, Chae Y. Data Mining of Acupoint Characteristics from the Classical Medical Text: DongUiBoGam of Korean Medicine. *Evid Based Complement Alternat Med*. 2014;2014:329563. doi: 10.1155/2014/329563.
7. Jung WM, Lee T, Lee IS, Kim S, Jang H, Kim SY, Park HJ, Chae Y. Spatial Patterns of the Indications of Acupoints Using Data Mining in Classic Medical Text: A Possible Visualization of the Meridian System. *Evid Based Complement Alternat Med*. 2015 ; 2015 : 457071. doi: 10.1155/2015/457071.
8. Oh JH. Deduction of Acupoints Selecting Elements on Zhenjiuzishengjing using hierarchical clustering. *J Daejeon Univ Oriental Medical Research Center*. 2014 ; 23(1) : 115-24.
9. Longxiang Huang. The Classic of Nourishing Life with Acupuncture and Moxibustion. huaxiachubanshe. 1996 ; 233-364.
10. Lee Jeong-hyeon. Problems with Chinese Ideographs Search in Unicode and Solutions to Them. *Informatization policy*. 2012 ; 19(3) : 50-63.
11. The Unicode Consortium[homepage on the Internet]. The Unicode Consortium; unknown [cited 22 Feb 2016]. Available from: <http://www.unicode.org/Public/8.0.0/ucd/DerivedNormalizationProps.txt/>
12. Ming Zhe Jin, Myung Hoe Huh. Author Identification of Korean Texts by Minimum Distance and Machine Learning. *Survey research*. 2012 ; 13(3) : 175-90.
13. Nees Jan Van Eck, Ludo Waltman. How to Normalize Co-Occurrence Data? An Analysis of Some Well-Known Similarity Measures. *JASIST*. 2009 ; 60(8) : 1635-51.