The Korean Journal of Applied Statistics (2016) DOI: http://dx.doi.org/10.5351/KJAS.2016.29.6.1007
29(6), 1007–1023

Conditions and potentials of Korean history research based on 'big data' analysis: the beginning of 'digital history'

Sangkuk Lee[a,1]

[a]Department of History, Ajou University

(Received September 19, 2016; Revised October 10, 2016; Accepted October 10, 2016)

Abstract

This paper explores the conditions and potential of newly designed and tried methodology of big data analysis that apply to Korean history subject matter. In order to advance them, we need to pay more attention to quantitative analysis methodologies over pre-existing qualitative analysis. To obtain our new challenge, I propose 'digital history' methods along with associated disciplines such as linguistics and computer science, data science and statistics, and visualization techniques. As one example, I apply interdisciplinary convergence approaches to the principle and mechanism of elite reproduction during the Korean medieval age. I propose how to compensate for a lack of historical material by applying a semi-supervised learning method, how to create a database that utilizes text-mining techniques, how to analyze quantitative data with statistical methods, and how to indicate analytical outcomes with intuitive visualization.

Keywords: big data, Korean history, digital history, interdisciplinary convergence, quantitative analysis

1. 서론

최근 전세계적으로 '빅데이터'에 대한 관심이 고조되고 있다. 빅데이터는 "신종 천연자원" 혹은 "21세기 원유"로 지칭되며 새로운 성장동력으로 각광받기 시작한 것이다. 각 학문분야에서도 빅데이터를 적극적으로 활용하기 위한 다양한 방안들이 모색되고 있으며, 인문학, 특히 역사학 분야에서도 빅데이터의 중요성과 활용방안에 대한 관심이 큰 폭으로 증가하고 있다. 역사학 분야에서 역사적 자료(사료 등)의 중요성은 역사학이 시작된 이래 꾸준히 강조되어 왔지만, 근래 불어 닥친 '역사학 빅데이터'에 대한 관심은 그 차원을 달리하는 것이다.

역사학 자료가 '역사학 빅데이터'로 일컫게 되는 중요한 계기는 2000년대 전후의 시점이라고 할 수 있다. 다양한 학문분야 연구에 대한 정부 차원의 재정적 지원이 이루어졌고, 역사학 분야에서도 역사학 자료에 대한 전산화가 본격화되었다. 「삼국사기」나 「삼국유사」, 「고려사」, 그리고 「조선왕조실록」 등이 전산화되었을 뿐만 아니라 금석문, 문집류, 고문서, 호적, 족보 등의 다양한 사료에 대한 전산화도 진행

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2015S1A5B6037107).
[1]Department of History, Ajou University, 206 Worldcup-ro, Yeongtong-gu, Suwon 16499, Korea.
E-mail: okllsskh@ajou.ac.kr

되고 있다. 활자 매체를 통해서 역사학 자료를 접하던 연구자들에게 '디지털 데이터'의 창출과 활용이라는 새로운 연구 환경이 조성된 것이다. 활자로 기록된 역사 자료를 자구 하나하나 그리고 그 사이 행간의 의미를 살피는 전통적인 질적분석(qualitative analysis)의 역사 연구방법론으로는 변화하는 연구 환경에 대처하는데 분명한 한계가 드러나기 시작하였다. 전산화된 방대한 역사학 빅데이터를 대량으로 효율적으로 처리하기 위한 양적분석(quantitative analysis)을 역사학 연구에 적용하는 방법론적 모색이 절실히 요청되는 것이다.

본 글은 역사학, 그 중에서 한국사 연구에서 디지털 기술을 활용한 빅데이터 분석 방법론을 모색하고, 이를 활용한 '디지털 역사학'의 가능성에 대해 검토하는 것을 목적으로 한다. 우선 전산화된 한국사 빅데이터의 현황을 소개한다. 다음으로 전산화된 데이터를 활용한 연구를 검토하고, 이들 연구의 문제점을 살펴본다. 이 과정에서 한국사 관련 비정형데이터를 정형데이터로 전환하여 분석하는데 필수적인 다양한 학문분야들 간 융합연구의 필요성을 제기할 것이다. 기술 데이터의 텍스트를 의미 분석하여 마이닝하고 계량화 데이터로 전환하는 언어학과 전산학 연구방법론, 전환된 데이터베이스를 통계적으로 처리하여 시각화 기법에 적용하는 통계학과 시각화 연구방법론 등을 역사학 연구에 적용하는 학제 간 융합연구를 모색하는 것이다. 마지막 장에서는 한국사 빅데이터 기반 학제 간 융합연구의 방법론을 제안하고, 이를 적용하여 한국사의 연구주제를 탐구하는 연구의 일례를 소개한다. 이를 통해 역사학 기반 학제 간 융합연구의 가능성을 모색해 보고자 한다.

2. 한국사 '빅데이터'의 현황과 관련 연구

한국사 관련 자료의 전산화에 있어 [국역 조선왕조실록] CD롬의 간행(1995)은 역사자료 정보화 작업의 시초로 평가받을 만큼 중요한 사건이었다 (Lee, 2003). 이후 [삼국사기], [고려사], [경국대전] 등 한국사 연구의 기초 사료들의 전산화와 더불어 기존의 역사연구 방식에 일대 전환을 촉진했기 때문이다. 책의 형태로 활자화된 자료를 하나하나 읽어가면서 자신의 연구를 뒷받침하는 근거를 찾아내는 방식에서, 관심을 갖고 있는 키워드를 전산화된 CD롬에 입력하여 대규모로 다양한 관련 사료를 검색하고, 이를 취사선택하는 방식으로 전환된 것이다. 연구자가 원하는 사료적 근거를 보다 쉽게 찾아낼 수 있다는 점에서, 한국사 연구에 큰 진전이 있으리라는 기대는 자연스러운 것이었다. 사료 읽기에 들였던 많은 시간을 절약할 수 있고, 장기간에 걸친 자료의 수집으로 확대된 시야를 갖고 역사적인 사건의 장기추세를 살펴볼 수 있는 여건이 마련된 것이다.

이러한 기대는 더 많은 자료에 대한 전산화의 요구로 이어졌고, 1999년을 전후하여 정부차원에서 더 많은 역사자료가 전산화되기에 이르렀다. 한국사 자료의 전산화는 2000년 9월 정보통신부에서 수립된 「지식정보자원관리 기본계획(안)」에 따라 "한국역사정보통합시스템 구축사업"을 계기로 국사편찬위원회, 서울대 규장각, 민족문화추진회(현 한국고전번역원), 한국정신문화연구원(현 한국학중앙연구원) 등이 참여하는 통합 검색시스템인 "한국역사정보통합시스템(http://www.koreanhistory.or.kr)"로 일원화되기에 이르렀다 (Joo, 2008). 또한 한국연구재단의 기초연구지원사업 등으로 인해 한국사 관련 사료의 전산화는 그 규모가 더욱 확대되었다. 호적과 족보 데이터, 지방지, 문집류 등도 전산화되어 '역사학 빅데이터'라고 불릴만한 방대한 한국사 관련 데이터가 축적되기에 이르렀던 것이다.

자료 접근성의 용이성은 한국사 연구의 진전에 대한 기대를 더욱 높였지만, 그만한 정도로 한국사 연구의 위기를 지적하는 목소리도 높아진 것도 사실이다. 전산화된 사료가 구축된 통합시스템을 통해 연구자의 목적에 맞는 사료를 키워드를 통해 쉽게 검색할 수 있게 되었지만, 그것이 사료에 대한 깊은 성찰로는 이어지지 못하는 것에 대한 반성이 대두된 것이다. 사실 한국 역사학계가 거의 전적으로 의지하고 있는 연구방법론은 실증적 연구 방법론으로 (Cho 등, 1994), 이것은 역사 자료에 기록되어 있는 사실들을 연구자의 연구 목적에 따라 자구 하나하나 그리고 그 사이 행간의 의미까지도 세세하게 해석하고 서

술하는 역사 연구방법론이다. 역사 자료의 행간 읽기를 통해 특정 시기 횡단면적 역사적 사실에 대한 다양하고 풍부한 해석을 도출하였고, 한정된 문헌의 분석에 기반 한 특정시기의 횡단면적 연구에 많은 도움을 주었다. 이에 따라 구체적이고 세밀한 자구 해석을 기반으로 한 의미 파악이 역사 연구에 있어 무엇보다 중요하게 인식되었고, 사료 하나하나 혹은 행간에 대한 깊은 성찰은 역사 연구자의 중요한 미덕으로 재확인 되었던 것이다.

그럼에도 기술자료(qualitative data)를 활용한 역사적 사실의 서술(narrative)에 의존하는 한국사 연구의 문제점을 지적하는 목소리는 잦아들지 않고 있다. 미리 재단된 한정된 문헌들의 틀에서 벗어나지 못하며, 다양하고 방대한 자료를 통해 나타나는 다양한 역사적 현상을 파악하는 데에는 한계를 노정하고 있기 때문이다. 이제 전통적인 연구방법론은 방대한 '역사 빅데이터'의 축적으로 인해 촉발된 연구 여건의 변화에 능동적으로 대처할 수 없으며, 다양하고 복잡한 현대사회에서 역사학에 바라는 요구에 부응하지 못하는 결과를 낳고 있다는 것이다. 와그너와 송준호가 「안동권씨성화보」, 「문화유씨가정보」 등 족보의 자료적 가치를 높이 평가하고, 「문과방목」 등을 선도적으로 데이터베이스화하면서 이를 활용할 새로운 연구방법론의 출현을 고대했지만 (Wagner, 2007), 그들의 바람은 여전히 현재 진행형이다 (Kim과 Lee, 2014). 확장된 데이터를 효과적으로 다룰 수 있는 새로운 연구 방법론의 필요성이 역사학에 대두되고 있는 것이다.

하지만 최근까지 역사학계보다 다른 학문분야에서 더 적극적으로 '한국사 빅데이터'를 활용한 연구를 진행하고 있는 것이 사실이다. 한국사 관련 자료가 디지털화 되어 누구나 용이하게 접근할 수 있게 되어 타 학문분야의 전공자들도 그들의 연구방법론으로 '역사학 빅데이터'를 가공할 수 있는 기반이 마련된 것이다. 사회과학이나 데이터사이언스, 심지어 자연과학 분야의 전공자들이 '한국사 빅데이터'를 그들의 방법론으로 가공해 역사콘텐츠 생산을 모색하고 있다. 족보 등 가계기록의 데이터베이스 표준화 시도 (Lee, 2006), 한국 중세 고고학 기초자료의 데이터베이스 구축을 위한 기초조사 실시 (Hong, 2013), 그리고 웹페이지 상에서 사용자들에 의해 생산 및 소비되는 웹 서비스 기반 데이터를 활용한 역사 온톨로지 생성 시도 (Kang 등, 2015) 등이 그것이다. 또한 역사적 사실에 대한 실질적 분석도 이루어지고 있다. 삼국시대 왕을 중심으로 인물 네트워크를 구축하고 구축된 인물을 국내정치, 국제관계, 가족관계, 왕이 아닌 인물들과의 관계로 분류하고, 링크수를 반영하여 각 왕의 국가 통치나 국가경영 방식을 도출한 연구 (Chung와 Kim, 2011), 그리고 족보에서 제공하는 본관의 정보를 통해 혼인의 지리적 이동성을 추적하는 연구 (Lee 등, 2014) 등이 그것이다.

이들 연구는 학제간 융합연구의 필요성이 더욱 대두되는 현 상황에서 학문 간 경계를 허문 시도로 평가받기에 충분하다고 생각한다. 하지만 '한국사 빅데이터'를 데이터베이스화하기 위한 표준화 방법론의 논의에 머물러 있어, '한국사 빅데이터'에 디지털 기술을 적용하여 역사학적 주제에 대한 해답을 찾는 데까지는 나아가지 못하고 있다. 또한 사료가 갖는 역사성에 대한 이해 없이 자의적으로 데이터를 가공함으로써 역사적 사실관계에 대한 오류를 범하거나 과대 해석을 도출하는 등의 문제점을 드러내고 있다. 즉, 삼국시대 인물 네트워크 연구에서 분석대상 텍스트가 1차 사료인 「삼국사기」나 「삼국유사」가 아닌 현대 연구자가 펴낸 가공된 역사서였다는 점에서 역사학적인 비판을 피해갈 수 없으며, 따라서 그 분석 결과를 역사적 해석으로 수용하기 곤란한 측면이 있다. 또한 족보를 통한 이동성을 연구한 연구에서는 족보에 기재된 본관의 의미에 대한 역사학적 접근 없이 데이터에 나타나는 정보 그대로를 수량화하여 분석하였다는 문제점을 노정하였다.

이처럼 역사학 이외의 학문분야에서 시도한 연구들 중 일부는 '한국사 빅데이터'의 자료적 성격을 역사학적으로 검증하지 않은 채 수량 데이터로 전환한 후 분석 결과를 도출하였기 때문에, 이로 인해 나타나는 오류를 피하기 어려웠다. 전산화된 빅데이터를 양적분석하기 위해서는, 그러므로 한국사 자료 자체에 대한 깊이 있는 연구가 필수적인 것이다. 이점과 관련하여 2000년 이후 현재까지 진행되고 있는 조

선시대 '호적대장'의 전산화와 연구는 좋은 규범이 될 만하다고 생각한다. 이중 '단성호적대장'의 전산화는 18–19세기 경상도 단성현(현재 산청군)의 호적대장에 기록된 개인 정보를 엑셀로 전환하는 방식으로 이루어졌다. '단성호적대장'에 기록된 연 인원 20여 만 명의 모든 정보를 한글과 한문으로 전산화하여 대중에 보급한 것이다 (성균관대학교 대동문화연구원 http://daedong.skku.ac.kr/). 이를 바탕으로 한국사 연구자들은 호적대장의 자료적 성격을 수년에 걸쳐 검토하고, 이를 바탕으로 조선후기 사회사 연구를 심화하고 있다 (Household Registers Research Team, 2003). '단성호적대장'의 전산화는 기술자료 분석에 의지하던 한국사 연구에 획기적인 전환을 마련하였다. '단성호적대장'에 기록된 개인에 대한 정보는 역사인구학이나 가족사 연구의 주요한 자료였으며 (Miyazima, 2004), 국내 한국사 연구자들에게 역사인구학이라는 학문 분야를 개척하는 계기를 제공하였다 (Academy of East Asian Studies at Sungkyunkwan University eds.). 또한 '단성호적대장'은 해외 학계에도 알려져 해외의 연구자들이 한국의 '호적대장' 자료를 활용하여 동아시아의 역사인구학적 현상을 비교사적으로 검토하는 계기를 제공하기도 하였다 (Dong 등, 2015).

그럼에도 '한국사 빅데이터'를 활용한 한국사 연구에는 해결해야할 과제가 산적해있다. 우선, 전산화된 한국사 관련 데이터에 대한 역사학적 검증 및 분석이 필요하다. 한국사 연구자 이외에 다양한 학문분야에서 '한국사 빅데이터'에 용이하게 접근할 수 있게 되었지만, 해당 데이터의 사료적 성격에 대한 이해, 그리고 이를 바탕으로 한 분석 대상 데이터의 선정이 적정하게 이루어지고 있지 않는 것이다. 둘째, 한국사 연구자들이 인접 학문에 대한 이해와 관심이 요청된다. 사실, '한국사 빅데이터'를 활용한 양적분석이 한국사 연구자가 아닌 데이터 사이언스 연구자들에 의해 시도된 것은, 한국사 연구자들이 통계학적인 데이터 분석방법론에 익숙하지 않기 때문이기도 하다. 한국사 연구자들은 통계적 방법론을 활용한 양적분석보다는 기술자료(qualitative data)를 활용한 질적분석 방법론에 전적으로 의지해왔던 것이다. 정리하면, 한국사 연구자들에게는 통계학 등을 기반으로 한 양적분석 방법론, 그 외 연구자들에게는 '한국사 빅데이터'에 대한 역사학적 이해가 동시에 필요하다. 이제 역사학에는 다방면의 학문분야의 방법론을 '역사학 빅데이터'에 적용하는데 필요한 기준을 제시하고, 사료의 역사성에 바탕을 둔 해석의 방향성을 정립할 소명이 주어졌다. 역사학이 타 학문분야와 적극적으로 대화해야할 시점에 이른 것이다.

역사학과 타 학문분야와의 학제 간 융합의 필요성은 일찍부터 제기되었다. 역사학은 "인접 학문인 사회과학의 관심"을 이끌어내어야 서로 다른 리듬의 시간들과 구조들이 중층적으로 상호작용하는 가운데 형성·지속되어온, '분리될 수 없는 전체'에 적확히 접근할 수 있다는 것이다 (Braudel, 1980). 여기에 현재 '한국사 빅데이터'를 제대로 활용하기 위해서는 사회과학만이 아닌 데이터사이언스를 포함한 자연과학과의 학문적 교류도 필요하다. 하지만 현재까지도 학제적 융합을 통한 역사학 연구는 실질적으로 이루어지지 않고 있으며, 다른 학문분야에서 역사학 데이터를 활용하여 시도한 몇몇 연구는 역설적이게도 역사학과 타 학문분야와의 학제 간 융합의 필요성을 촉구하고 있다.

그러므로 '한국사 빅데이터'를 기반으로 한국사의 다양한 주제들을 연구하기 위해서는 역사학적인 관점에서 데이터를 이해하고 해석하는 역사학 연구자, 기술 데이터를 마이닝하여 의미를 분석하고 계량화 데이터로 전환하는 언어학과 전산학 연구자, 전환된 데이터베이스를 통계적으로 처리하여 시각화 기법에 적용하는 통계학과 시각화 연구자 등과의 긴밀한 학제 간 융합 연구가 필수적이다. 즉, 디지털 기술을 역사학에 적용해서 역사학적 명제를 분석하고 해결하는 '디지털 역사학'이 필요한 시점인 것이다.

3. 한국사 '빅데이터' 기반 '디지털 역사학' 연구 방법론의 모색

3.1. 연구 방법론

본 논문에서 제안하는 '한국사 빅데이터' 기반 학제 간 융합연구는 역사학, 언어/전산학, 통계학, 정보

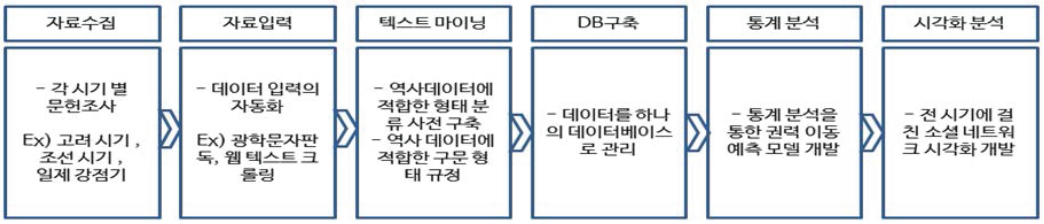


Figure 3.1. A roadmap of the interdisciplinary convergence research based on big data of Korean history.

Figure 3.2. A capture from the webpage of the Annals of Joseon dynasty.

시각화 분야의 전문가들이 각 분야의 방법론을 '한국사 빅데이터'에 적용하는데서 시작된다. 각 학문 분야 별 또는 각 학문 분야 간 다양한 융합 연구를 바탕으로 '한국사 빅데이터'에 최적화된 방법론을 찾아가는 것이다. 학제 간 융합연구를 통해 우선적으로 해결해야 하는 점은 다음과 같다. '한국사 빅데이터'에서 관련 주제를 분석하기 위한 데이터베이스화는 많은 시간과 노동력이 투여된다는 점, 하나의 변수를 생성하기 위해서는 문서를 정성적으로 읽어야 한다는 점, 역사적 인물들 간의 혈연, 학맥, 혼인 등 다양한 관계의 추출 등이 그것이다. Figure 3.1은 이러한 문제점을 극복하기 위한 '한국사 빅데이터' 기반 학제 간 융합연구의 체계도이다.

3.1.1. 데이터 수집 '한국사 빅데이터' 중 가장 기본적인 사료는 고려시대와 조선시대 관찬사서인 「고려사」와 「조선왕조실록」 (http://db.history.go.kr/KOREA/item/level.do?itemId=kr&types=r; http://sillok.history.go.kr/main/main.jsp)이다 (Figure 3.2).

Figure 3.3. Data crawling(left) and data refining(right).

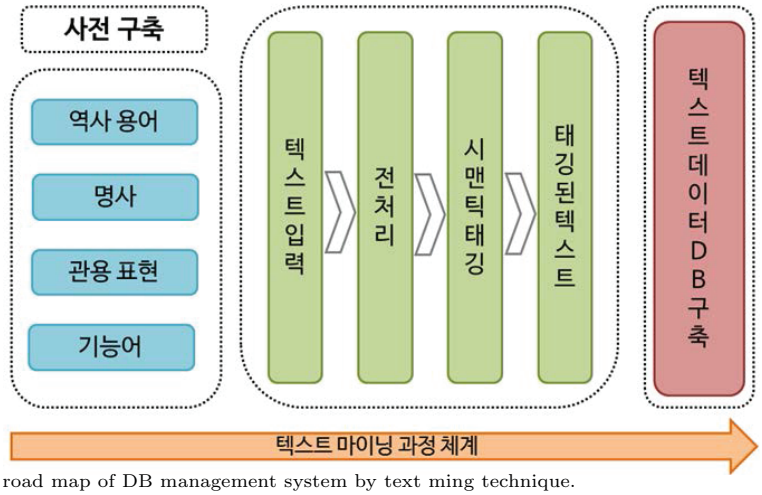


Figure 3.4. A road map of DB management system by text ming technique.

조선왕조실록 웹페이지에는 왕들의 재위 순서에 따라 기록된 정보를 한글로 풀어쓴 문자 형태의 방대한 데이터들이 제공되고 있다. 웹페이지의 형태를 분석하고 JAVA를 활용하여 웹페이지에서 나타내는 모든 정보를 가져온다. Figure 3.3은 조선왕조실록의 웹페이지에서 크롤링한 데이터와 데이터를 정제한 일부를 보여준다.

데이터의 형태는 여러 종류가 있으나 일반적으로 JSON형태와 CSV형태로 데이터를 가공하여 사용한다. 1차적으로 JAVA를 활용하여 웹상에서 가져올 수 있는 모든 정보를 JSON형태의 데이터로 크롤링한다. 또한 2차적으로는 분석 목적에 따라 데이터를 정제하여 사용한다.

3.1.2. 데이터베이스와 언어학/전산학적 방법론 '역사 빅데이터'는 그 양이 방대하여 텍스트 마이닝에 의한 분석이 필수적이다. 이를 위해 언어학/전산학적 방법론을 활용하여 다음과 같은 시스템을 구축할 필요가 있다.

Figure 3.4에서 나타나는 것처럼, 우선 입력 텍스트에 대해서 형태소분석 등의 전처리를 시행한 뒤 사전을 참조하여 의미 분석한 결과를 태깅한다. 그런 다음 이렇게 태깅된 텍스트를 데이터베이스의 형태로 전환하여 저장한다. 그런데 현재 개발되어 있거나 개발 중인 텍스트마이닝 시스템은 여러 가지 문제

를 안고 있다. 그 가운데 가장 큰 것은 키워드 방식에 의존하여 출현빈도 만을 통계 처리함으로써 정보의 내부구조를 파악하지 못하고 있다는 점이다. 이로 인해, 예컨대 권력 재생산 구조의 메커니즘을 고찰하기 위해 친족관계나 인척관계를 알아보려 할 때 기존의 마이닝 시스템은 '아버지', '아들', '딸', '혼인' 등의 키워드를 검색하여 처리하고 있으나, 이로부터 정보의 내부구조, 즉 누가 누구의 아들인지, 누가 누구와 혼인을 한 것인지를 분석해 낼 수 없다. 그러나 본 논문에서 제안하는 언어학/전산학적 방법론은 키워드 분석이 아닌 문장의 구조의 의미를 분석해 내는 기법이다. 그 과정은 다음과 같다. 먼저 문장의 의미를 분석하는 시스템이 인간관계를 나타내는 술어들을 형태소 분석이 이루어진 텍스트에서 검색한다. 여기에는 '아들', '딸', '배위(配位)'와 같은 관계명사뿐 아니라 용언, 즉 '낳다', '장가들이다'와 같은 동사와 '이다', '있다' 같은 형용사들도 검색된다. 텍스트에서 이 같은 술어들이 발견되면 시스템은 이들을 사전에서 찾아 이들이 구성하는 문장의 구성요소들(주어, 보어 등)을 해당 키워드의 전후에서 찾는다. 문장의 구성요소들은, '세종전자사전'에서 보듯이, 키워드를 표제어로 하는 어휘내항의 격틀(<frame>)에 나와 있다 (Park 등, 2016).

한편 한국어는 모든 문장성분이 생략될 수 있는 언어여서 자동 처리에 어려움을 야기한다. 그러나 문장성분의 생략이 빈번한 현대 구어 자료와 달리, 역사문헌의 경우에는 문장의 거의 모든 구성성분이 실현되어 분석이 어렵지 않다. 다만 주어가 생략되어 있는 경우들이 많은데, 이때는 주어를 선행 절의 마지막 명사('이다' 앞의 명사)로 추정한다.

3.1.3. 통계적 계량분석 및 검증 언어학/전산학적 텍스트 마이닝 방법론을 통해 '한국사 빅데이터'에서 역사적 인물들 간의 다양한 관계를 데이터베이스로 구축한다. 구축된 역사 데이터베이스에는 각 개인, 그들의 부계·모계·처계 등 혈연관계, 혼인관계, 학맥관계 등이 기록되어 있다. 각 개인은 자신의 고유한 ID를 갖도록 설계되어 있는데, 이 ID를 통해 그의 부-조 등 혈연관계 및 혼인관계를 추적할 수 있다 (Lee, 2006; Lee와 Park, 2008). 이러한 설계를 바탕으로 기초 통계량 분석을 시도한다. 우선 관직의 높고 낮음(정·종1품-정·종9품의 관품, 지방관, 왕실 등)을 변수로 개인의 사회적 지위에 미친 부와 조의 영향력을 분석하고 (Lee, 2013), 친족의 사회적 지위가 서로 어떤 긴밀한 연결성을 갖고 있는지를 살펴보고, 시대 별로 직계 가족 간의 계급 관계가 어떻게 변화해 왔는지 추적한다. 또한 혼인을 통한 사회적 지위 및 계급의 확장과 세습에 관한 유기적 연결 관계를 분석한다. 개인과 가문이 혼인 네트워크를 활용해 어떻게 그들의 사회적 지위를 유지하는 가를 살펴볼 수 있다 (Lee와 Lee, 2016).

그런데 전반적인 권력 분포도는 흔히 가정하는 정규분포(normal distribution)를 따르지 않는다. 따라서 데이터의 기대 분포 값을 바탕으로 갖고 있는 데이터의 성질을 반영한 시뮬레이션 데이터를 가공하여 데이터의 분포 속성을 파악하고 이 속성에 적용할 수 있는 비모수추정(Non-parametric Estimation)법을 사용하여 세대와 세대 사이의 권력 생성, 확장 그리고 승계 구도를 밝혀낼 Spearman나 Pearson 상관분석 검증치를 추출한다. 또한 부와 조의 영향력을 다항로짓 모형(multinomial logit model)이나 다항로짓 모형의 설명변수들에 대한 한계효과로 추정하여 살펴볼 수 있다 (Lee, 2013).

또한 다양한 응용 계량 기법을 적용하여 그 결과값의 안정성 및 신뢰성을 측정할 수 있다.첫째, 그랜저 인과검정(Granger causality)은 이러한 가계 혹은 개인 자료를 수집한 데이터 안에서 각 변수들 간의 인과관계를 통계적으로 파악하는 데에 응용될 수 있다. 둘째, 이 비대칭적인 관계성 구도를 파악하기 위하여 아들과 아버지의 등급에 공통적으로 영향을 미칠 수 있고(예컨대, 할아버지의 등급), 통제변수와 아버지의 등급이 설명하지 못하는 오차 항에 포함될 수 있는 변수들(예컨대, 혼인 후 장인의 등급)에 초점을 맞춘다. 셋째, 통계학적인 분석기법 중에서도 패널 데이터를 다룰 때에 가장 주의해야 할 변수 간 내생성(endogeneity) 문제를 면밀히 검토하고 필요에 따라 "데이터를 정리한다. 마지막으로 비모수 추정법과 모수 추정법의 비교도 실행하여 그 추정치에 관한 신뢰성을 높이고, 선형 관

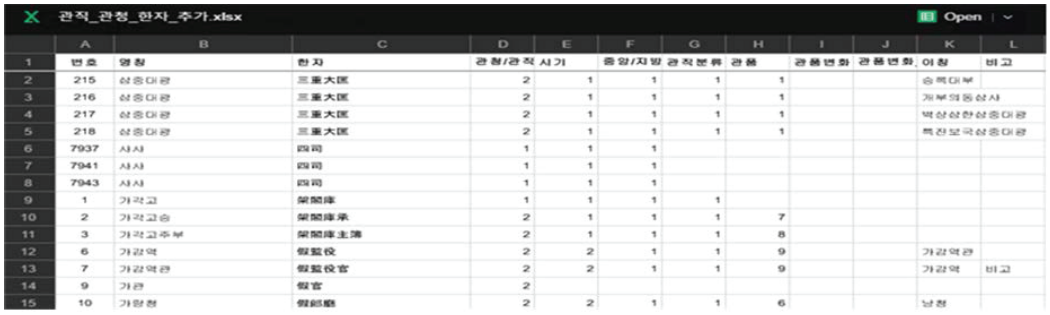
계에서 통계적으로 유의한 관계성 모수를 찾아내지 못할 경우, 비선형 관계가 적용 가능한 최우 추정법(maximum likelihood estimator)를 적용하여 통계적으로나 역사적으로 유의한 역사상과 그 의미를 밝혀낸다 (Lee와 Yoo, 2016).

'역사학 빅데이터' 분석 후, 그 결과 값의 안정성 혹은 신뢰성을 검증하기 위하여 몇몇 통계적 기법을 추가 도입한다. 변수의 패턴을 파악하는 데에는 주로 다른 변수들과의 관계성에서 그 증거를 찾을 수 있는데, 이 관계성을 함축하는 모수는 시간에 따라 혹은 가족의 큰 분류 등급에 따라 비대칭적으로 나타날 수 있다. 이 비대칭 관계성 구도를 파악하기 위하여 비대칭(asymmetry)요인을 도입한 응용 계량 모형을 사용한다. 예컨대 횡단면적인 성격을 분류하는 더미(dummy)변수를 생성하여 계량 분석에 포함시킬 수 있다. 개인별 네트워크 데이터를 주축으로 하여 유난히 깊은 세습 연결 고리를 가지는 가족의 성향은 나머지 가족들의 내부 세습 구도보다 큰 연결 고리를 보이면서 평균적인 연결의 특이성을 왜곡할 수 있다. 이런 가족은 주의 깊게 들여다보아야 할 필요성이 있으므로 그러한 가족 혈연관계에 추가 변수를 부여하여 시대 전체의 평균적인 연결성을 정리할 수 있다. 이 더미변수의 통계적 유의성이 검증된다면, 큰 데이터 안에서 전반적인 관계성 고리를 만드는 동시에, 비대칭적인 사회구조 통찰이 가능할 것이다.

구조적 비대칭성은 횡단면 분석에서만 발견되는 것이 아니라 시대의 흐름에 따른 동적 사회 구조 변화에서도 감지된다. 이러한 시계열적인 구조변화 파악을 위해서는 구조변화 검정(structural break)을 시행한다. 구조변화 검정 역시 사회 구조에 충격을 준 사건을 기점으로 더미변수를 생성하여 그 유의성을 입증할 수 있다. 다만 통계학 모형을 전공한 계량전공학자들에게는 어떤 사건이 유의한 사건인지 가늠하기가 어려우므로 구조 변화의 시점을 정리할 때에 큰 사건을 기준으로 먼저 테스트를 시행하고, 역사학적 사료 분석 자료를 통해 얻은 유의한 사건을 뽑아 구조 변화 유의성을 테스트 한다.

3.1.4. 네트워크 시각화 모델 및 구현 방법 시각화를 통한 데이터 분석은 데이터를 시각적으로 표현하여 수많은 정보를 요약함으로써 통시적이고 직관적으로 파악할 수 있다는 장점이 있다. 하지만 데이터의 양이 많아질 경우 모든 데이터의 특징을 한 눈에 보기 어려워진다. 예를 들어 데이터의 개수가 일정 수를 넘어갈 경우 하나의 데이터에 하나의 화소를 부여해도 모든 데이터를 각기 다르게 나타내기 어렵다. 이 때문에 일반적으로 방대한 정보는 군집화(clustering)를 통해 분류하고, 그에 따른 특성을 시각화하여 데이터의 이해를 돕는 연구들이 수행되어 왔다. 그런데 데이터를 분류하여 군집화 할 경우, 개별 노드들의 속성을 파악하는데 한계가 있을 수밖에 없다. 이러한 한계를 극복하기 위해 세대 간 혈연의 계승관계를 보여줌과 동시에 각 개인이 가지는 영향력이나 다양한 특성을 효과적으로 분석할 수 있도록, 데이터마이닝과 다양한 그래프 시각화 연구를 병행하여 비교적 정보의 양이 많은 족보와 같은 역사 데이터를 분석하는 방안을 모색한다. 도입할 수 있는 기존 그래프 시각화로는, 노드의 위치를 바꾸어서 표현하는 방법, 3차원 공간을 이용하는 방법, Edge Bundling 알고리즘을 사용하여 인접한 링크를 묶어 간략히 표현하는 방법 등이다 (Huang 등, 2009; Bezerianos 등, 2010; Beck 등, 2014).

또한 장기 추세를 보여주는 동적 데이터가 가지는 특징을 고려한 시각화 방법을 적용한다. 동적 데이터는 시간의 변화에 따라 그룹 내의 생성과 결합, 그리고 소멸을 포함하고 있으므로, 이를 분석하여 시각화함으로써 해당 그룹의 진화 과정을 알 수 있다. 시계열 시각화는 연속적인 시간 간격에 의해 측정된 데이터를 보여주는 시각화 기법으로 시간의 흐름에 따른 데이터의 변화 측정 및 이전 데이터를 기반으로 향후의 데이터를 예측하는데 사용된다. 정적인 데이터의 경우 시간의 흐름에 영향을 받지 않으면서 데이터의 단면적인 표현이 가능한데 비해 동적인 데이터는 시간의 흐름에 영향을 받으면서 시시각각 변화하는 모습을 표현하는 것이 중요하다. 이점을 고려하여 역사 데이터의 서로 다른 시간에 존재하는 네트워크 그룹 간의 유사도를 계산하여 각 그룹을 추적하고 식별함으로써 그룹들이 시간의 흐름에 따라 어떻게 변화했는지 파악하고, 그래프 이론과 레이아웃 알고리즘에 관한 선행연구를 고찰하여 역사적 인물의



관직_관청_한자_추가.xlsx

	A	B	C	D	E	F	G	H	I	J	K	L
1	번호	명칭	한자	관청/관직 시기		중앙/지방	관직분류	관품	관품변화	관품변화	이칭	비고
2	215	삼중대광	三重大匡	2	1	1	1	1			승록대부	
3	216	삼중대광	三重大匡	2	1	1	1	1			개부의동삼사	
4	217	삼중대광	三重大匡	2	1	1	1	1			벽상삼한삼중대광	
5	218	삼중대광	三重大匡	2	1	1	1	1			특진보국삼중대광	
6	7937	사사	四司	1	1	1						
7	7941	사사	四司	1	1	1						
8	7943	사사	四司	1	1	1						
9	1	가각고	架閣庫	1	1	1	1					
10	2	가각고승	架閣庫承	2	1	1	1	7				
11	3	가각고주부	架閣庫主簿	2	1	1	1	8				
12	6	가감역	假監役	2	2	1	1	9			가감역관	
13	7	가감역관	假監役官	2	2	1	1	9			가감역	비고
14	9	가관	假官	2								
15	10	가랑청	假郎廳	2	2	1	1	6			낭청	

Figure 3.5. Office and Official position titles of Joseon dynasty.

네트워크 분석에 적합한 시계열 시각화 방법을 적용한다 (Moon 등, 2016).

마지막으로 역사적 인물들 간의 상호 관계에 적합한 시각화 방법론을 적용할 수 있다. 인물들 간의 다양한 네트워크의 시각화를 위해서는 'Force-Directed' 방식을 이용한 소셜 그래프인 네트워크 그래프 시각화가 적합하며, 영향력의 계승관계를 확인하고자 할 때는 노드 트레이싱을 이용하여 그 관계를 선을 따라가며 추적하는 방식의 시각화가 적합하다. 그리고 네트워크의 노드 간의 관계를 알아보는 가장 일반적인 방법인 네트워크 그래프 시각화는 노드들의 전체적인 분포를 확인하고 유사한 노드들의 클러스터를 확인하는 면에서는 쉽지만, 한 집단을 중심으로 다른 집단과의 관계를 파악하는 데 한계가 있다. 이러한 문제점을 보완하는 Proximity based Circular 시각화 방법론 (Choi 등, 2015)은 방사형 그래프와 라인 그래프를 함께 사용함으로써 특정 시점과 연속적인 흐름을 동시에 비교할 수 있고 이를 통해 네트워크 그래프 시각화로 확인하기 어려운 노드 간의 유사도 추이를 확인할 있다.

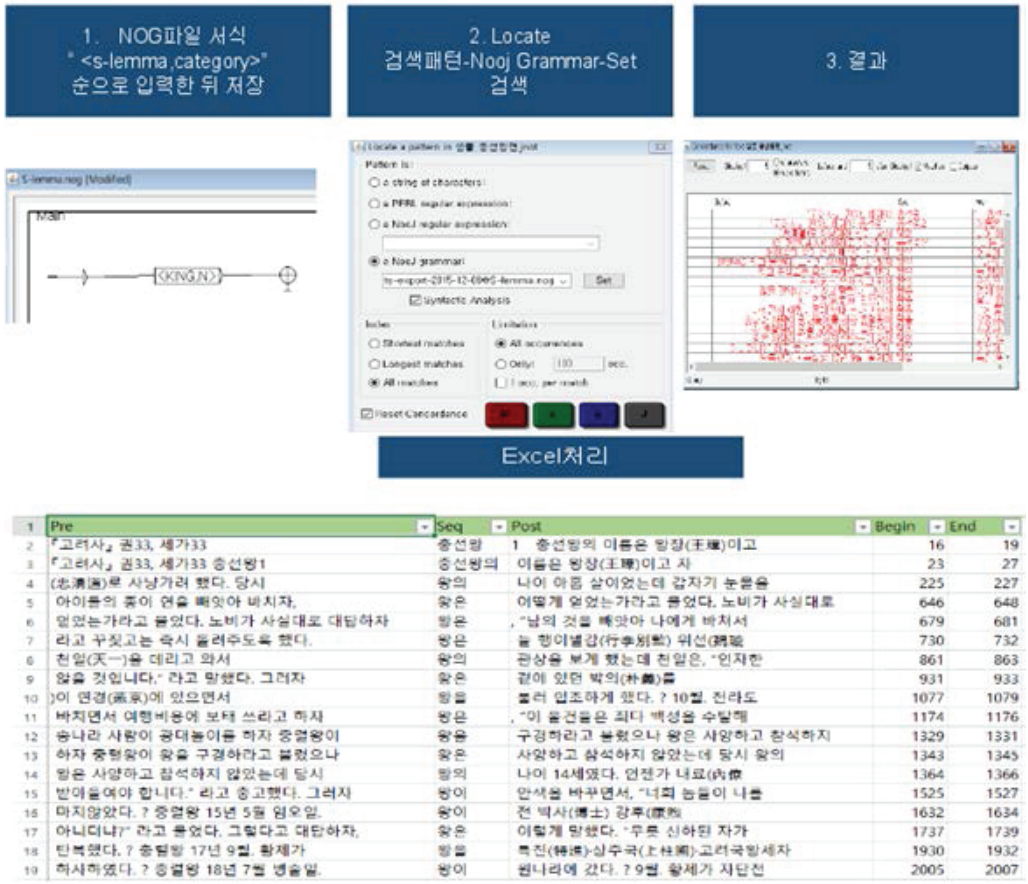
3.2. 학제 간 융합연구의 일례; 역사적 주요 인물들의 권력 관계 추정

이 절에서는 언어/전산학, 데이터마이닝/통계학, 정보시각화 분야의 방법론을 실제 역사적 주제를 탐구하는데 적용해보고자 한다. 연구주제는 조선 전기 역사적 주요 인물들의 권력 관계를 추정하는 것이다. 본 연구를 위한 기준 인물들[Time Markers]은 [안동권씨성화보]에 기록된 조선 전기의 인물들을 바탕으로 선정되었다. 역사팀이 선정한 타임 마커들에 대한 정보를 바탕으로 언어/전산팀, 데이터마이닝/통계팀, 정보시각화팀 등 각 학문 분야 별 또는 각 학문 분야 간 융합 연구의 과정을 소개한다. 이를 통해 역사학 기반 학제 간 융합연구 방법론과 그 가능성을 구체적으로 모색할 수 있을 것이다.

3.2.1. 역사팀 권력 관계를 추정하기 위해서는 각 인물들의 정보와 인물들 간의 다양한 관계를 파악하는 것이 중요하다. 역사팀이 수집한 인물 정보 데이터는 언어/전산팀에 제공되어 [고려사]와 [조선왕조실록] 등에서 권력 관련 사료의 형태소 분석과 온톨로지(사전)를 구축하는데 주요한 자료로 활용되며, 데이터마이닝/통계팀과 시각화팀에게는 분석 대상 시기 타임 마커로 활용되고 있다. 역사팀의 연구는 우선, 전근대 시기 관직과 관청과 [문과방목] 등의 정보를 엑셀로 전환하는 작업을 시행하였다. Figure 3.5와 Figure 3.6은 역사팀이 구축한 관직과 관청과 [문과방목]의 일부를 캡처한 것이다.

3.2.2. 언어/전산팀 언어/전산팀은 [고려사]와 [조선왕조실록] 등 '한국사 빅데이터'의 텍스트마이닝을 위한 형태소 분석 및 코퍼스 사전을 구축한다. 이를 위해 언어개발도구인 NooJ를 활용하여 '역사 빅데이터'의 데이터베이스화 툴을 구축한다. NooJ는 Corpora를 단어별, 형태별, 통사별, 의미별 레벨에서 분석 가능한 프랑스에서 개발한 텍스트분석 프로그램이다. NooJ를 통해 사용자 정의의 문법 요

Figure 3.6. Information of historial figures from Gukjo mungwa bangmok.



	Pre	Seq	Post	Begin	End
2	『고려사』 권33, 세가33	충선왕	1 충선왕의 이름은 왕장(王璋)이고	16	19
3	『고려사』 권33, 세가33 충선왕1	충선왕의	이름은 왕장(王璋)이고 자	23	27
4	(忠淸道)로 사냥가려 했다. 당시	왕의	나이 아홉 살이었는데 갑자기 눈물을	225	227
5	아이들의 종이 연을 빼앗아 버치자,	왕은	어떻게 얻었는가라고 물었다. 노비가 사실대로	646	648
6	얻었는가라고 물었다. 노비가 사실대로 대답하자	왕은	, "남의 것을 빼앗아 나에게 바치서	679	681
7	라고 꾸짖고는 즉시 돌려주도록 했다.	왕은	늘 행이별감(行李別監) 뒤선(隨從	730	732
8	천일(天一)을 데리고 와서	왕의	관상을 보게 했는데 천일은, "인자한	861	863
9	왕을 것입니다." 라고 말했다. 그러자	왕은	같이 있던 박의(朴義)를	931	933
10	)이 연경(燕京)에 있으면서	왕을	불러 업조하게 했다. ? 10월, 진라도	1077	1079
11	바치면서 여행비용에 보태 쓰라고 하자	왕은	, "이 물건들은 죄다 백성을 수탈해	1174	1176
12	송나라 사람이 공대놀이를 하자 충렬왕이	왕을	구경하라고 불렀으나 왕은 사양하고 참석하지	1329	1331
13	하자 충렬왕이 왕을 구경하라고 불렀으나	왕은	사양하고 참석하지 않았는데 당시 왕의	1343	1345
14	왕은 사양하고 참석하지 않았는데 당시	왕의	나이 14세였다. 언전가 내료(內僚	1364	1366
15	받아들여야 합니다." 라고 충고했다. 그러자	왕이	안색을 바꾸면서, "너희 놈들이 나를	1525	1527
16	마지않았다. ? 충렬왕 15년 5월 원오일,	왕이	전 박사(博士) 강후(康侯	1632	1634
17	아니더냐?" 라고 물었다. 그렇다고 대답하자,	왕은	이렇게 말했다. "우릇 신하된 자가	1737	1739
18	탄복했다. ? 충렬왕 17년 9월, 황제가	왕을	특진(特進)상주국(上柱國)고려국왕세자	1930	1932
19	하사하였다. ? 충렬왕 18년 7월 병술일,	왕이	원나라에 갔다. ? 9월, 황제가 자단전	2005	2007

Figure 3.7. Data mining process by using NooJ program.

소(예를 들면 본 연구에 필요한 명사, 관직계급 등)를 [고려사]나 [조선왕조실록]에서 추출할 수 있다. 이를 바탕으로 역사적 주요 인물들의 권력 관계에 대한 정보를 '역사 빅데이터'에서 추출, 수집한다.

Figure 3.8. Database system of historical figures and families in pre-industrial Korea.

F_C		505	506	507	508	509	544	545	546	547	548	549	552	595	596	597	598	599	600	601	604	607
	Name	이영부	권숙희	이장	최급	서절	권림	여의보	권우	권빈	이의번	권류	채수	노철강	김영견	유경	최진	김구	김구	김영서	김영정	윤계겸
505	이영부	0	2	2	2	2	6	6	6	6	6	6	6	10	10	11	11	11	11	10	10	10
506	권숙희	2	0	2	2	2	6	6	6	6	6	6	6	10	10	11	11	11	11	10	10	10
507	이장	2	2	0	2	2	6	6	6	6	6	6	6	10	10	11	11	11	11	10	10	10
508	최급	2	2	2	0	2	6	6	6	6	6	6	6	10	10	11	11	11	11	10	10	10
509	서절	2	2	2	2	0	6	6	6	6	6	6	6	10	10	11	11	11	11	10	10	10
544	권림	6	6	6	6	6	0	2	2	2	2	2	4	10	10	11	11	11	11	10	10	10
545	여의보	6	6	6	6	6	2	0	2	2	2	2	4	10	10	11	11	11	11	10	10	10
546	권우	6	6	6	6	6	2	2	0	2	2	2	4	10	10	11	11	11	11	10	10	10
547	권빈	6	6	6	6	6	2	2	2	0	2	2	4	10	10	11	11	11	11	10	10	10
548	이의번	6	6	6	6	6	2	2	2	2	0	2	4	10	10	11	11	11	11	10	10	10
549	권류	6	6	6	6	6	2	2	2	2	2	0	4	10	10	11	11	11	11	10	10	10
552	채수	6	6	6	6	6	4	4	4	4	4	4	0	10	10	11	11	11	11	10	10	10
595	노철강	10	10	10	10	10	10	10	10	10	10	10	10	0	2	3	3	3	3	2	2	2
596	김영견	10	10	10	10	10	10	10	10	10	10	10	10	2	0	1	1	1	1	2	2	2
597	유경	11	11	11	11	11	11	11	11	11	11	11	11	3	1	0	2	2	2	3	3	3
598	최진	11	11	11	11	11	11	11	11	11	11	11	11	3	1	2	0	2	2	3	3	3
599	김구	11	11	11	11	11	11	11	11	11	11	11	11	3	1	2	2	0	2	3	3	3
600	김구	11	11	11	11	11	11	11	11	11	11	11	11	3	1	2	2	2	0	3	3	3
601	김영서	10	10	10	10	10	10	10	10	10	10	10	10	2	2	3	3	3	3	0	2	2
604	김영정	10	10	10	10	10	10	10	10	10	10	10	10	2	2	3	3	3	3	2	0	2
607	윤계겸	10	10	10	10	10	10	10	10	10	10	10	10	2	2	3	3	3	3	2	2	0
608	김영순	10	10	10	10	10	10	10	10	10	10	10	10	2	2	3	3	3	3	2	2	2
611	박원정	9	9	9	9	9	9	9	9	9	9	9	9	5	5	6	6	6	6	5	5	5
612	정지산	9	9	9	9	9	9	9	9	9	9	9	9	5	5	6	6	6	6	5	5	5

Figure 3.9. Kin network of Seo, Geojeong in Andong Gwon-ssi suhwa-po genealogy.

Figure 3.7은 NooJ를 통해 1차 가공하여 얻은 데이터를 캡처한 것이다 (Park 등, 2016).

NooJ를 통해 추출한 정보를 데이터베이스로 전환하는 과정이 필요하다. Figure 3.8은 아주대 학제간 융합연구팀의 언어/전산팀이 족보와 문과방목, [고려사], [조선왕조실록] 등 '한국사 빅데이터'로부터 추출한 관직과 인명을 데이터베이스화하기 위한 툴을 캡처한 것이다.

3.2.3. 데이터마이닝/통계팀 데이터마이닝/통계팀은 비정형 '역사 빅데이터'에 대한 다양한 통계 분석 및 기계학습 기법 적용하여 추출된 정보를 분석하는 알고리즘 개발하고, 통계적 처리 및 네트워크 분석에 적용한다. 역사팀과 언어/전산팀에서 구축한 [안동권씨성화보] 데이터의 인물정보와 해당 인물들이 [조선왕조실록]에서 보이는 정치행위에 대한 데이터를 바탕으로 권력 집단의 세력분포를 분석한다. Figure 3.9와 Figure 3.10은 [안동권씨성화보]를 작성한 서거정을 중심인물로 하여 그와 혈연관계를 맺고 있는 인물들의 촌수를 계산한 부분과 족보 네트워크의 개념도이다 (Lee 등, 2016).

이를 바탕으로 족보 네트워크와 역사 기록물을 바탕으로 정의된 권력 집단을 준지도 학습(semi-supervised learning) 방법론에 적용할 수 있다. Figure 3.11에서 보이는 연구 설계를 바탕으로 족

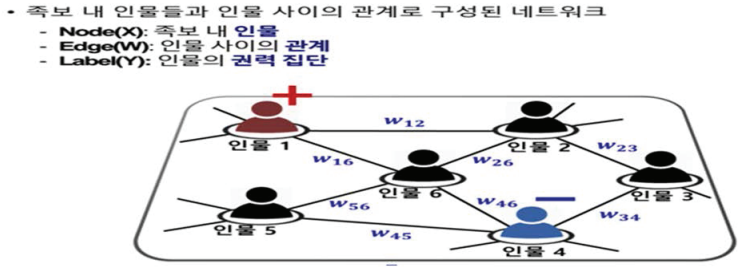


Figure 3.10. Conceptual diagram of networks of historical figures in a genealogy.

Figure 3.11. Conceptual diagram of Semi-Supervised Learning system applied to networks of historical figures in a genealogy.

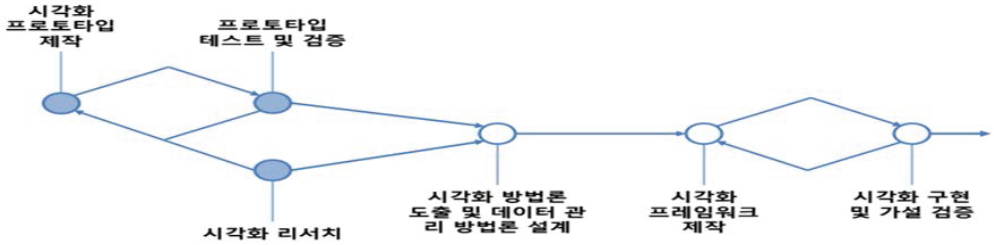


Figure 3.12. A road map of network visualization optimized to big data of Korea history.

보 네트워크 내의 인물들의 정치적 지향을 조사하고, 기계학습을 통해 정치 집단을 구분할 수 있다. 이러한 방법론을 적용하여 임의로 선택한 인물이 속한 권력 집단을 추론하였는데, 68%의 정확도가 있음을 확인하였다. 이는 역사 사료에서 확인할 수 없어 세력 분포가 불분명한 역사적 인물들도 준지도 학습 방법을 통해 그들의 정치적 성향을 추정할 수 있음을 의미한다. 이 방법론은 사료가 충분하지 않은 역사적 인물과 사건들을 분석하는데 유용하게 활용될 것으로 기대된다 (Lee 등, 2016).

3.2.4. 시각화팀 시각화팀은 역사팀과 언어/전산팀에서 구축한 데이터를 바탕으로 '역사 빅데이터'에 최적화된 네트워크 시각화를 모색한다 (Figure 3.12).

시각화팀은 구축된 데이터베이스와 통계학적 표준화계수를 바탕으로 시각화를 구현한다 (Lee과 Lee, 2014). 우선 Figure 3.13에서 보이는 것처럼「성화보」전체를 조망할 수 있는 시각화를 구현하였다. 데이터베이스에 포함된 모든 변수를 시각화 프로그램에 내장하여 각 변수마다의 필터링을 활용해 원하는 정보를 얻을 수 있게 고안되었다. 각 개인을 선택하면 본인의 관직뿐만 아니라 부, 조의 관직도 한 눈에 알아볼 수 있으며, 그의 자, 손자 등 후손들의 관직도 살펴볼 수 있다. 족보데이터의 구조와 정보를 직관

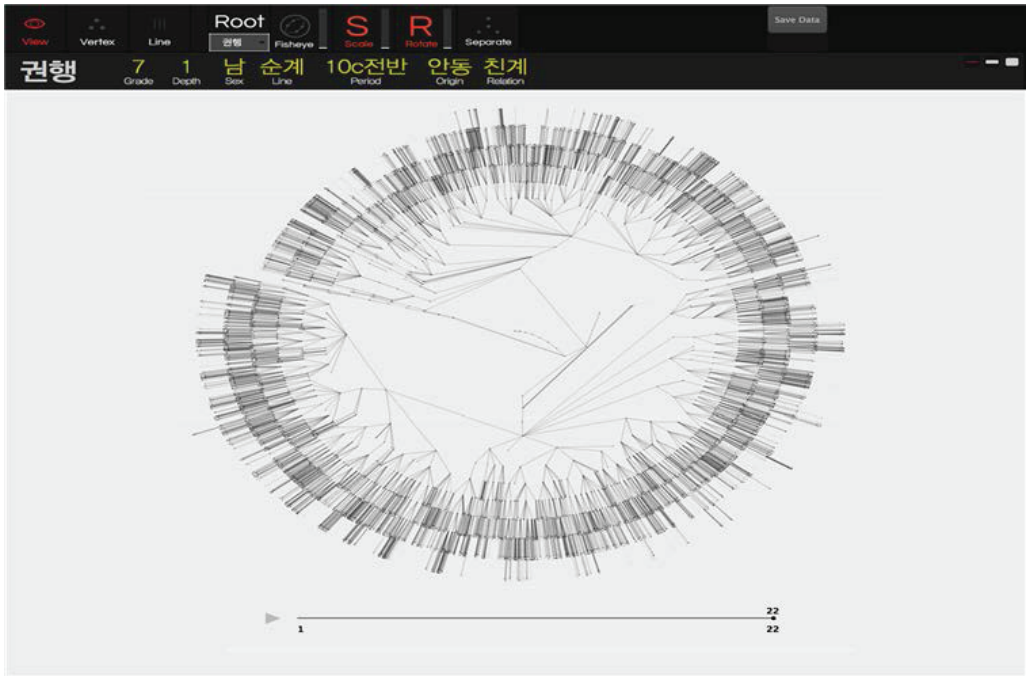


Figure 3.13. A visualization program for kin networks of Andong Gwon-ssi suhwa-po genealogy.

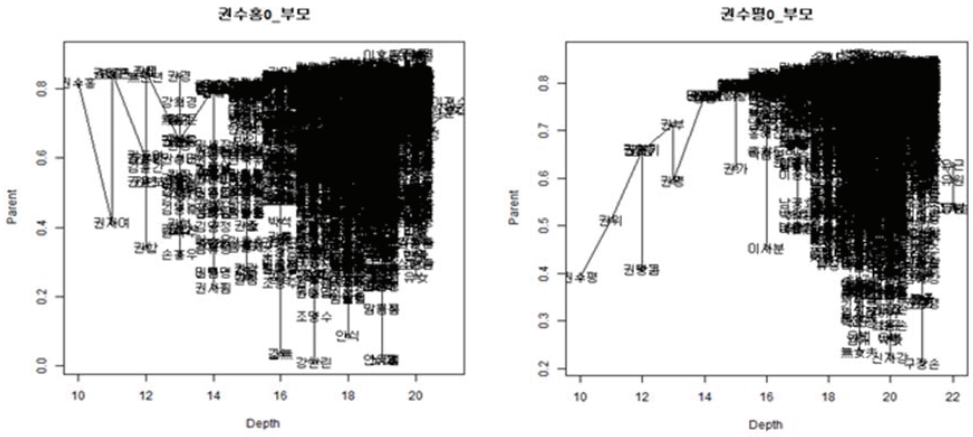


Figure 3.14. A visualization of the impact of a father on individual's social achievement based on Andong Gwon-ssi suhwa-po genealogy.

적이고 통시적으로 이해하는데 유용한 시각화이다.

시각화팀은 Figure 3.13의 시각화를 바탕으로 역사적 인물들의 각 집단 간 사회적 지위가 유지되는 주요한 전술이 어떻게 달라지는 지를 비교하였다. Figure 3.14는 [안동권씨성화보]에 나타난 권수홍계열과 권수평계열의 개인의 사회적 성취에 미친 아버지의 영향력을 시각화한 것이다.

Figure 3.14는 개인의 사회적 성취에 미친 부의 영향력에 대한 매우 흥미로운 사실을 직관적으로 보여준

다. 각 계열에 따라 개인의 성취도에 영향을 미친 요인이 다르게 나타난다. 추밀공파(권수평계열)의 경우, 부의 영향력을 보여주는 그림이 대부분 상단에 집중되어 있는 반면 복야공파(권수홍계열)는 추밀공파와 다른 양상을 보이고 있다. 즉, 추밀공파의 경우, 개인의 사회적 지위는 부, 즉 직계혈연에 큰 영향을 받는 반면에 복야공파(권수홍계열)의 경우 직계혈연의 영향력이 제한적이다. 그러므로 개인의 사회적 성취에 미치는 영향력의 요소들은 각 가문마다 다르게 나타나고 있다는 것을 알 수 있다. 이상과 같이 시각화 방법론을 '한국사 빅데이터'에 적용함으로써 역사적 주제에 대한 입체적이고 장기적 추세를 읽어낼 수 있다.

4. 결론

방대한 '한국사 빅데이터'를 활용한 한국사 연구를 위해서는 기존의 질적분석 방법론뿐만 아니라 양적분석 방법론이 모색되어야 한다. 이를 위해서는 다양한 학문 분야와의 학제 간 융합연구가 요청된다. 본 글에서는 '한국사 빅데이터'를 활용한 다양한 융합연구의 출현을 고대하면서, 학제 간 융합연구의 연구방법론을 제안하고, 이를 적용한 연구의 한 사례를 소개하였다. 이는 '한국사 빅데이터'의 출현으로 야기된 역사 연구의 패러다임의 전환을 촉구하는 것이기도 하다. 역사학이 언어학/전산학, 데이터마이닝/통계학, 시각화 등 각 학문 분야 별 또는 각 학문 분야 간 융합연구에 적극적으로 대화해야 할 여건이 성숙된 것이다.

그럼에도 불구하고 역사학 분야에서 학제 간 융합을 통한 양적분석 방법론을 선뜻 받아들이지 못하는 것은 다음과 같은 문제점 때문이라고 생각한다. 즉, 데이터화하여 계량화할 때 각 개인은 세대별 그룹의 일원으로 동질성이 강요되어, 개인의 역사성은 세대별 그룹 내의 다른 구성원과 동일한 성격으로 규정되어 경시될 수밖에 없다. 둘째, 이론이나 통계학적 방법론을 자의적으로 적용할 수 있다. 측정된 결과를 자료 자체 내의 역사성을 무시한 채 일반화하거나 특정 역사적 사건만을 설명하는데 사용할 수 있다는 것이다 (Isaac과 Griffin, 1989). '한국사 빅데이터'를 계량화하여 양적분석을 하고자 할 경우에도 유사한 문제가 제기될 수 있는 것이다. 자구 하나하나의 의미를 분석하고 해석하는 질적분석을 모범으로 삼는 역사학 분야에서 이와 같은 문제점을 갖는 양적분석을 받아들이기 쉽지 않다.

이에 대해서는 계량화를 통해 경시된 개인의 역사성은 그 개인이 속한 동일 세대라는 집단을 통해 복원될 수 있다는 주장이 제기되었다. 즉, 각 세대 구성원들은 동일한 시대를 살아가며 유사한 역사적 경험을 했다고 할 수 있으므로 (Hollingsworth, 1957), 이들을 하나의 집단으로 설정할 경우 이들을 개별 단위로 한 세대별 혹은 시대별 역사적 과정을 살펴볼 수 있다는 것이다 (Lee, 2013). 하지만 대량의 '빅데이터'를 기반으로 한 연구 과정에서 간과할 수밖에 없는 '행간의 의미읽기의 부재'는 여전히 문제점으로 남을 수밖에 없다. 본 글에서 제안한 언어학/전산학적 방법론을 활용한 문장의 의미를 분석하는 방법론의 구현이 시급히 요청되는 이유이다. 문장의 의미를 분석하는 텍스트 분석방법으로 '한국사 빅데이터'에서 원하는 정보를 추출한다면, 양적분석 방법론의 단점으로 지적되는 '행간의 의미읽기의 부재'를 점차 보완해 갈 수 있을 것이다. 그리고 이러한 방법론으로 구축한 데이터베이스를 바탕으로 준지도 학습 방법론을 적용할 경우, 사료가 충분하지 않은 전근대 한국사의 역사적 인물과 사건들을 분석하는데 유용하게 활용될 것으로 기대된다. 분석 결과를 직관적으로 보여주는 시각화를 통해서도 평면적 연구에서 찾아내지 못한 역사적 사실들을 밝혀낼 수 있을 것이다.

이상과 같이 역사학을 기반으로 다양한 학문 분야의 방법론을 '한국사 빅데이터'에 다채롭게 적용하여 양적분석 방법론이 갖는 문제점을 보완해 갈 때, '한국사 빅데이터'로 초래된 새로운 역사 연구의 환경에 걸맞은 패러다임의 전환이 이루어질 것으로 기대한다. 이제 '디지털 역사학'의 서막이 오른 것이다.

References

Academy of East Asian Studies at Sungkyunkwan University eds. (forthcoming). The Attempt of Historical Demography Research in Korea, Sungkyunkwan Univeristy Press.

Beck, F., Burch, M., Munz, T., Di Silvestro, L., and Weiskopf, D. (2014). Generalized Pythagoras trees for visualizing hierarchies. In *Information Visualization Theory and Applications (IVAPP), 2014 International Conference on* (pp. 17–28), IEEE.

Bezerianos, A., Dragicevic, P., Fekete, J. D., Bae, J., and Watson, B. (2010). Geneaquilts: A system for exploring large genealogies, *IEEE Transactions on Visualization and Computer Graphics*, **16**, 1073–1081.

Braudel, F. (1980). *On History* (trans. by Sarah Matthews), University of Chicago, Chicago.

Cho, D., Han, Y., and Park, C. (1994). *Korean Historian and Historical Awareness*, Yeoksabipyungsa Press.

Choi, H., Moon, S., Ha, H., and Lee, K. (2015). Proximity based Circular Visualization for similarity analysis of voting patterns between nations in UN General Assembly,*Design Convergence Study*, **53**, 133–150.

Chung, J. S. and Kim, H. Y. (2011). Analysis of people networks in Goguryeo, Baekje, and Silla Dynasty Silloks, *Journal of Korea Contents Association*, 11-5.

Dong, H., Campbell, C., Kurosu, S., Yang, W., and Lee, J. Z. (2015). New sources for comparative social science: Historical population panel data from East Asia, *Demography*, **52**, 1061–1088.

Hong, Y. (2013). Database Construction Process of Basic Archaeological Resources in the Medieval Period in Korea – Promoting the utilization archaeological resources of the medieval history in Korea, *Korean Medieval Research*, **36**, 71–106.

Hollingsworth, T. H. (1957). A demographic study of the British ducal families, *Population Studies*, **11**, 4–26.

Household Registers Research Team (2003). *he Study of Danseong Household Register, Daedong Institute for Korean Studies*, Sungkyunkwan University Press.

Huang, M. L., Huang, T. H., and Zhang, J. (2009). TreemapBar: Visualizing additional dimensions of data in bar chart. In *2009 13th International Conference Information Visualisation* (pp. 98–103), IEEE.

Issac, L. W. and Griffin, L. J. (1989). Ahistoricism in time-series analyses of historical process: Critique, redirection, and illustrations from US labor history, *American Sociological Review*, 873–890.

Joo, S. (2008). A study on the Bulding of digitized history materials and standardization, *The Study of Historical Folklore*, **26**, 209–246.

Kang, D., Lee, J., and Kim, T. (2015). A study of history a generation method of history ontology using meta-data. In *Presented at 2015 KICS(Korea Information and Communications Society) Winter Conference*.

Kim, K. and Lee, B. (2014). History as the core source of cultural contents, *The Journal of Korean Historical Association*, **224**, 424–448.

Lee, D., Lee, S., and Shin, H. (2016). The inference of power mechanism of early Joseon Dynasty. In *Presented at 2016 'Interdisciplinary Convergence Research Methods for Analysis of 'History Bigdata': The Beginning of 'Digital History'' Conference*.

Lee, K. (2006). Research on a new data model for analysis of pedigree records, *Jangseogak*, **16**, 195–232.

Lee, N. (2003). Humanity Sciences and Knowledge Information - Forcusing on the Law Concerned with Knowledge Information Resources and the Korean History Information Unification System -, Humanities Contents 1, 117-130.

Lee, S. (2013). The Impact of Family Background on Bureaucratic Reproduction in the thirteenth-to-fifteenth century Korea: A case study on the Andong Kwon-ssi Sunghwabo, *Daedong Institute for Korean Studies*, **81**, 42–67.

Lee, S. and Lee, K. (2014). Visualization approach to a Korean genealogy data. Presented at 2014 Digital Humanities International Conference at Ajou University.

Lee, S. and Lee, W. (2016). *Strategizing Marriage: An Analysis of Marriage Networks in the Andong Gwon-ssi Genealogy* (working paper).

Lee, S. and Park, H. (2008). Marriage, social status, and family succession in Medieval Korea (Thirteenth-Fifteenth centuries), *Journal of Family History*, **33**, 123–138.

Lee, S. and Yoo, J. (2016). *Microstudy on the Length of Life in the Late Joseon Korea* (working paper).

Lee, S. H., Francon, R., Abrams, D. M., Kim, B. J., and Porter, M. A. (2014). Matchmaker, matchmaker, make me a match: migration of populations via marriages in the past, *Physical Review X*, **4**, 041009.

Miyazima, Hiroshi (2004). The present situation and problems in the study of Korean population history, *Daedong Institute for Korean Studies*, **46**, 61–78.

Moon, S., Choi, K., Han, H., Lee, K., and Kim, J. (2016). VoteStream Vis: Visual Analisys of Congressional Votes in the Annals of the Joseon Dynasty. In *Presented at 2016 'Interdisciplinary Convergence Research Methods for Analysis of 'History Bigdata': The Beginning of 'Digital History'' Conference*.

Park, M., Ye, H., and Kwon, K. (2016). Construction of Ontology and Thesaurus for expression searching for government office titles and appointment and dismissal·reward and punishment. In *Presented at 2016 'Interdisciplinary Convergence Research Methods for Analysis of 'History Bigdata': The Beginning of 'Digital History'' Conference*.

Wagner, E. W. (2007). *Achievement and Ascription in Joseon Dynasty* (translated by Lee, Hoonsang and Sookkyung Son), The Iljogak Press.

'빅데이터' 분석 기반 한국사 연구의 현황과 가능성: 디지털 역사학의 시작

이상국[a,1]

[a]아주대 사학과

(2016년 9월 19일 접수, 2016년 10월 10일 수정, 2016년 10월 10일 채택)

요 약

본 글은 역사학, 그 중에서 한국사 연구에서 활용 가능한 빅데이터 분석 방법론을 모색하고, 이를 활용한 '디지털 역사학'의 가능성을 검토하는 것을 목적으로 한다. 방대한 '한국사 빅데이터'를 활용한 한국사 연구를 위해서는 기존의 질적분석 방법론뿐만 아니라 양적분석 방법론이 모색되어야 한다. 이를 위해서는 다양한 학문 분야와의 학제 간 융합연구가 요청된다. 본 글에서는 '한국사 빅데이터'를 활용한 다양한 융합연구의 출현을 고대하면서, 학제 간 융합연구의 연구방법론을 제안하고, 이를 적용한 연구의 한 사례를 소개하였다. 즉, 문장의 의미를 분석하는 텍스트 분석방법으로 '한국사 빅데이터'에서 원하는 정보를 추출한다면, 양적분석 방법론의 단점으로 지적되는 '행간의 의미읽기의 부재'를 점차 보완해 갈 수 있을 것이다. 그리고 이러한 방법론으로 구축한 데이터베이스를 바탕으로 준지도 학습(Semi-Supervised Learning) 방법론을 적용할 경우, 사료가 충분하지 않은 전근대 한국사의 역사적 인물과 사건들을 분석하는데 유용하게 활용될 것으로 기대된다. 분석 결과를 직관적으로 보여주는 시각화를 통해서도 평면적 연구에서 찾아내지 못한 역사적 사실들을 밝혀낼 수 있을 것이다. 이제 '디지털 역사학'의 서막이 오른 것이다.

주요용어: 빅데이터, 한국사, 디지털 역사학, 학제간 융합, 양적분석

이 논문 또는 저서는 2015년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2015S1A5B6037107).

본 글은 한국연구재단의 2015년 선정 학제간융합연구사업에 선정된 ('빅데이터' 분석기반 한국사권력메커니즘) 연구를 소개하여, 빅데이터를 활용한 역사학 기반 융합연구를 촉진하고자 작성되었다. 본 글에서는 필자가 아주대 학제간융합연구팀의 공동연구원들과 수행하고 있는 공동작업의 일부를 소개한다. 하지만, 본 논문에서 제기되는 모든 오류에 대한 책임은 오롯이 필자의 몫이다.

[1](16499) 경기도 수원시 영통구 월드컵로 206, 아주대학교 인문대학 사학과. E-mail: okllsskh@ajou.ac.kr