텍스트마이닝을 이용한 윤동주 연구의 개체계량학적 분석*

Entitymetrics Analysis of the Research Works of Dong-ju Yun using Textmining

박 진 균 (Jinkyeun Park)**
김 택 윤 (Taekyoun Kim)***
송 민 (Min Song)****

초 록

이 연구는 텍스트마이닝 기술을 이용한 개체계량학적 분석을 인문학 분야 인물 연구에 적용하기 위해 수행하였다. 연구 대상으로 한 인물은 작품뿐만 아니라 종교, 생애에 대해 많은 연구가 이루어진 윤동주를 선정하였다. 본 논문에서는 윤동주 관련 연구 1,076건을 수집하여 이중에서 초록 정보를 가지고 있었던 220건의 논문을 대상으로 LDA(Latent Dirichlet Allocation) 방식의 토픽모델링 분석을 수행하였으며, 참고문헌 정보를 추출할 수 있었던 121건의 논문을 대상으로 저자동시인용 분석을 통해 연구의 동향을 살펴보았다. 또한 초록에서 인명, 작품명의 개체를 추출하여 이들의 관계를 살펴보았다. 이 연구를 통해 윤동주에 관련한 연구 동향은 생애, 시, 실존의식, 비교문학, 번역문학, 종교적 신념에 대한 연구로 다양한 분야에 걸쳐 이루어졌다는 것을 데이터를 기반으로 보다 객관적으로 분석해 볼 수 있었으며, 윤동주와 함께 연구되는 다른 인물이 어떤 작품을 매개로 하여 연구되어 왔는지에 대해서도 알 수 있었다. 이러한 결과는 인문학 분야의 지적구조를 밝히는데 개체계량학적 방법이 유용함을 증명하는 한편 인문학 연구의 새로운 시각적 접근을 제안했다는 데에 의의가 있다.

ABSTRACT

This paper employs entitymetrics analysis on the research works of Dong-ju Yun. He was a Korean poet who was studied by many researchers on his works, religion and life. We collected 1,076 papers about Dong-ju Yun and conducted various approaches including co-author citation analysis, topic modeling analysis to identify the topic trend in the study of Dong-ju Yun. Also we extracted entities like person's name and literature's title from abstract to examine the relationship among them. The result of this paper enables us to objectively identify the topic trend and infer implicit relationships between key concept associated with Dong-ju Yun based on text data. Moreover, we observed sub-research topics such as life, poem, aesthetic existence, comparative literature, literary translation, and religious beliefs. This paper shows how entitymetrics can be utilized to study intellectual structures in the humanities.

키워드: 텍스트마이닝, 토픽모델링, 개체계량학, 윤동주, 저자동시인용, LDA
Textmining, Topic Modeling, Entitymetrics, Dong-ju Yun, Co-author Citation Analysis, LDA

---

* 이 논문은 2015년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2015S1A3A2046711).
** 연세대학교 문헌정보학과 박사과정(libpark@yonsei.ac.kr) (제1저자)
*** 연세대학교 문헌정보학과 석사과정(taekyoun_kim@naver.com) (공동저자)
**** 연세대학교 문헌정보학과 교수(min.song@yonsei.ac.kr) (교신저자)
논문접수일자 : 2017년 2월 20일 논문심사일자 : 2017년 3월 8일 게재확정일자 : 2017년 3월 14일
한국비블리아학회지, 28(1): 191-207, 2017. [http://dx.doi.org/10.14699/kbiblia.2017.28.1.191]

1. 서 론

온라인 데이터 베이스의 발전은 학문 분야를 막론하고 축적된 연구들에 대해 쉽고 빠른 접근을 제공하고 있으며 인문학 분야 데이터베이스 확장에도 기여하였다. 인문학 분야의 연구는 연구자 개개인의 주관과 경험, 가치관에 따라 차이를 보일 수 있는 질적연구 중심이다. 특히 문학작품을 연구하는 분야의 경우 실제로 연구자가 많은 문학작품을 접해보고 오랜 시간 연구를 해야만 작품 안에 함축되어 있는 의미를 이끌어 낼 수 있었다. 하지만 이러한 연구의 흐름속에서도 하나의 작품이 다른 작품이나 작가에 어떠한 영향력을 미쳤는지를 보려고 하는 연구가 있었으며(Tonussi 2013), 문학작품 연구를 대상으로 계량서지학적 분석을 통해 연구동향과 작품의 영향력을 살펴보려는 연구가 있었다(윤순근 1992; Hammarfelt 2011). 또한 하나의 문학작품 내에서 등장인물이라는 개체를 추출하여 이들의 동적 네트워크를 보는 연구도 수행되는 등(김학용 2012) 계량정보학적 연구가 인문학 분야에도 충분히 적용될 수 있음을 알 수 있었다. 그럼에도 불구하고 이러한 선행연구들은 해외 유명 작가의 작품만 연구대상으로 했거나, 인용관계 분석을 통한 연구의 동향 파악에만 중점이 맞추어져 있었으며, 단일 작품 내에 존재하는 개체들의 관계만을 파악했다는 한계가 있었다. 따라서 이번 연구에서는 국내 문학작품을 대상으로 텍스트마이닝을 이용한 개체계량학적 방법론을 적용하여 대량의 데이터를 빠른 시간에 분석하여 연구의 동향을 알아보는 한편 다양한 연구자의 연구논문 속에 나타난 인물명과 작품명이라는 개체를 추출하여 이들의 관계까지도 추론해 보고자 한다.

국내 문학작품에 대한 연구를 함에 있어서 시험적으로 윤동주를 연구대상으로 설정하였다. 윤동주는 문학적 연구뿐만 아니라 정치적, 사회적으로도 우리나라에서 많이 연구된 작가이기 때문이다. 또한 최근에는 윤동주의 삶이 영화로 제작되기도 하였으며, 국내 인물 중에서 다양한 방면으로 연구결과가 나와 있는 인물 중 한 명이다.

본 연구의 목적은 텍스트마이닝 기법을 이용한 LDA방식의 토픽모델링과 저자동시인용분석 방법을 적용하여 윤동주 관련 연구의 동향을 파악하고, 새로운 방법론인 개체계량학적 분석을 통하여 윤동주 관련한 연구 안에서 인물명과 작품명 개체를 추출하여 이들의 숨겨진 관계를 추론해 내는 것이다. 이러한 연구방법의 적용은 기존 인문학자들이 수행했던 선행연구 분석 시간을 단축시켜 주고, 새로운 시각으로 인문학에 접근할 수 있는 기회를 제공할 수 있을 것이다. 또한 과학, 의학 분야를 중심으로 진행되어온 텍스트마이닝 및 개체계량학적 분석을 인문학에서도 인물의 생애, 작품연구 분야에 적용한 최초의 연구로서 의미를 가진다.

본 논문에서는 이와 관련된 선행연구를 살펴보고, 세부적인 데이터 수집과 분석방법을 설명하며, 토픽모델링, 저자동시인용, 개체계량학적 방법을 통해 윤동주에 관련한 연구를 분석한 결과를 제시함으로써 인문학 분야에 개체계량학적 방법론 적용 가능성을 살펴보고자 한다.

2. 선행연구

2.1 토픽모델링

토픽모델링은 문헌의 집합에 숨겨져 있는 토픽을 발견하기 위한 통계 모델의 유형이다. 특히 Latent Dirichlet allocation(이하 LDA)은 토픽 모델링 중 가장 대표적인 방법으로 각광 받고 있다. Blei, Ng, and Jordan(2003)은 각 문헌에 어떤 주제들이 분포되어 있는지에 대한 확률모형으로 LDA를 제안하였다. Wang과 McCallum (2006)은 LDA 토픽 모델링 중 시간에 따라 토픽이 어떻게 변화하는지 살펴볼 수 있는 Topics Over Time 모델을 제시하였다. 또한 Blei(2012)는 Science 저널에 수록된 17,000편의 논문과 Yale Law 저널을 대상으로 LDA 모델을 활용하였으며, 계량서지 분석에도 활용될 수 있음을 언급하였다.

국내 사례로 Song과 Kim(2013)은 생물정보학 분야의 지적구조 분석을 위하여 LDA 기반의 토픽모델링을 활용하였으며, 이를 통해 10개의 토픽을 추출하였다. 추출 결과를 통해 주요 토픽은 계산 관점의 생물정보학보다는 생물학적 관점이라는 것을 밝혔다. 또한 박자현과 송민(2013)은 LDA 기반의 토픽모델링을 활용하여 1970년부터 2012년까지 국내 문헌정보학 주요 5개 학술지에 발표된 논문의 초록에 대하여 실험을 수행하였다. 또한 토픽모델링을 통해 도출된 연구주제를 문헌정보학 주제분류표와 비교를 통하여 서로 연결하였다. 실험 결과 연구주제가 가장 많이 발견된 학문 영역은 정보학과 도서관서비스 분야인 것으로 나타났다. 이처럼 토픽모델링을 통한 분석은 많은 분량의 데이터를 소화할 수 있으며, 연구자의 주관적 가치와 개인적인 의견을 최대한 배제한 객관적인 연구결과를 얻을 수 있다는 장점이 있다.

2.2 저자동시인용 분석

동시인용 분석은 Small(1973)에 의해 제시된 분석방법으로서 두개의 문헌이 나중에 발표된 문헌에서 동시에 인용되는 것을 근간으로 학문분야의 지적구조를 분석하는 것이다. 이재윤(2005)은 국내 문헌정보학의 연구 전선 파악을 위하여 문헌동시인용 분석을 수행하였다. 이 논문에서 문헌정보학 분야 중요 논문에서 추출한 인용정보를 사용하여 군집 분석 및 네트워크 분석을 수행하였으며, 그 결과 27개의 복수 논문 군집과 8개의 단일 논문 군집을 도출하였다. 저자동시인용 분석은 인용 분석과는 다르게 문헌 보다는 연구자를 중심에 두는 지적구조 분석 방법이다. 따라서 문헌 단위가 아닌 저자 단위의 동시인용을 분석하는 기법이다. White와 Griffith(1981)는 두 명 이상의 저자가 동시에 같은 문헌에서 인용되는 것은 그들 사이에 더 밀접한 연관성이 있으며, 문헌동시인용 분석뿐만 아니라 저자동시인용 분석도 주제분야 전문성을 분석하는데 좋은 도구라는 것을 주장하였다. 저자동시인용 분석은 문헌이 아닌 저자명을 분석 대상으로 하므로 저자가 특정 주제를 포괄적으로 나타내는 특성을 이용하기 때문에 연구자가 그 분야의 상세한 서지 정보와 뚜렷한 전문 지식을 갖고 있지 않아도 분석이 가능하다는 장점이 있다(김희전, 조현양 2010; McCain 1983; White 1990). 저자동시인용 분

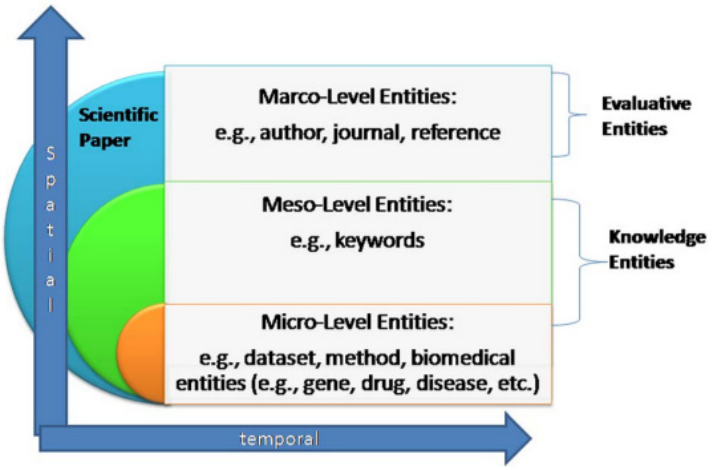
석이 가지는 단점으로는 인용 지체 현상이 있다. 인용 지체 현상이란 최근 연구 동향을 분석하거나 동시에 활발하게 연구하고 있는 저자를 파악하기에는 한계를 가지고 있다는 것을 의미한다. 따라서 저자동시인용 분석을 통해 파악한 지적 구조에 나타난 연구자와 현재 해당 분야를 연구하고 있는 연구자는 인용 지체 현상으로 인하여 일치하지 않을 수도 있다(이재윤 2008).

2.3 개체계량학

개체계량학이란 기존의 계량서지학이나 계량정보학이 분석대상으로 했던 개체가 저자, 초록, 키워드, 저널, 참고문헌 등 이라면, 이러한 개체를 중간단계의 개체로 보고 이보다 더 심층영역(micro level)인 주요 용어나, 텍스트 마이닝 기법으로 추출한 함축적인 정보를 개체로 하여 통계적 기법을 활용해 분석하는 방법론으로서 논문이나 연구자의 영향력을 좀 더 질적으로 분석할 수 있고, 기존 연구에서 발견할 수 없었던 새로운 지식을 추론할 수 있는 최신의 방법론이다(Ding et al. 2013). 이러한 개체계량학적 접근은 주요 용어(개체명)들이 잘 정리되어 있는 과학이나 의학, 생명공학 등의 분야에 효과적으로 적용될 수 있으며, 현재 활발한 연구가 진행중이다. Ding et al.(2013)은 지식의 영향력을 측정하기 위하여 개체계량학을 제안하였으며 과학 논문의 개체를 〈그림 1〉처럼 Macro-Level, Meso-Level, Mircro-Level로 정의하였다.

이 논문은 당뇨병 치료 물질인 메타포민과 관련 논문을 대상으로 인용정보를 활용하여 관련 개체들의 네트워크를 구성하였으며 또한 이 결과를 CTD(Comparative Toxicogenomics Database)와 비교하였다. 비교 결과 수작업으로 만들어진 CTD의 상호작용을 해당 연구결과가 거의 모두 탐지하는 것으로 나타났으며 개체계량학의 유용성을 확인 하였다.

함정은과 송민(2015)은 개체계량학 이론에 근거해 Swanson의 ABC 모델에 인용정보를



〈그림 1〉 개체계량학(Entitymetrics: Ding et al. 2013)

적용하여 논문 제목 및 초록에서 관련 있는 개체를 찾아내는 연구를 수행하였다. 수집된 논문의 참고문헌에서 인용정보를 추출하고 이를 활용하여 논문의 표제와 초록을 대상으로 텍스트마이닝 기법을 적용하여 주요 단어를 추출하였다. 이 연구를 통해 인용정보 기반 네트워크에서 발견되는 단어들이 동시출현 기반 네트워크를 통해 발견되는 단어들보다 관련 있는 개체들이 더 강하게 연결되어 있으며 암묵적인 관계도 더 많이 잡아내고 있다는 것을 밝혔다. 하지만 이러한 개체계량학이 비단 과학과 의학 분야에만 적용될 수 있는 것은 아니며 작품과 인명이 중요 개념으로 사용되는 인문학 분야에도 적용 가능하다.

2.4 윤동주 연구

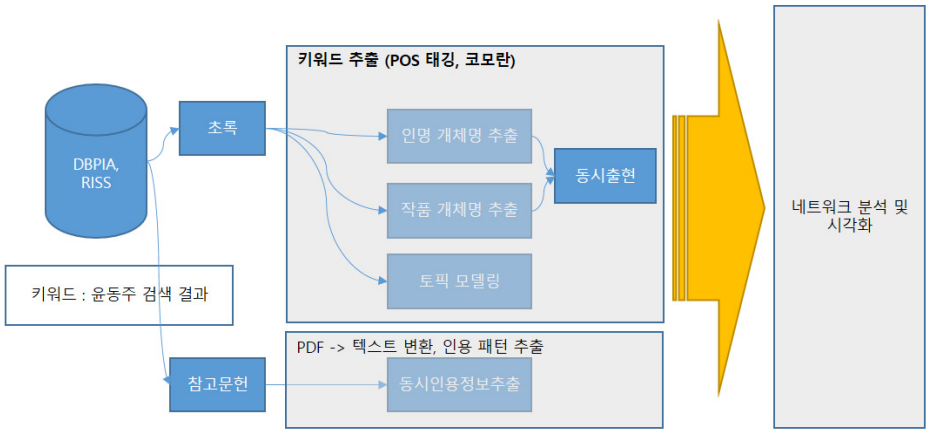
윤동주에 대한 인물 및 작품 연구는 다각적인 측면에서 이루어져 왔다. 윤동주 연구에 관한 선행연구를 알아보기 위해 우선 가장 최근에 발표된 논문을 기준으로 윤동주에 관한 연구를 찾아보고, 해당 연구들의 선행연구를 확인해 보았다.

김형태(2015)는 윤동주 시의 실존의식 연구라는 박사과정 논문에서 윤동주와 관련된 선행연구를 시기별로 정리하였다. 우선 윤동주에 대한 최초의 언급은 1947년 정지용에 의해 소개글 형식으로 이루어졌고, 1948년 2월 16일 정음사에서 유고시 31편을 묶어 『하늘과 바람과 별과 시』를 간행하였다. 그리고 1954년에 윤동주 시에 대한 최초의 비평으로 평가받는 고석규의 「윤동주의 정신적 소묘」가 발표되었고, 1960년대에 이상비, 이유식, 김열규, 최홍규 등에 의해서 윤동주 시에 관한 연구가 본격화되었다. 또한 1968년도판 『하늘과 바람과 별과 시』에 백철, 박두진, 문익환, 장덕순, 정병욱 등의 글이 더해졌는데, 이것도 윤동주 연구에 있어서 소중한 자료라고 할 수 있다. 1970년대는 대체로 저항성을 중심으로 윤동주의 문학사적 위치에 대한 논의가 이루어졌으며, 1980년대에는 윤동주의 개별 작품 해석에 중점을 두는 연구가 많이 수행되었다. 1990년대에 들어서면서 윤동주 시에 대한 어느 정도의 합의가 이루어지면서 학회지나 학위논문을 중심으로 새롭고 다양한 연구들이 시도되었다. 2000년대 이후로도 다양한 방법론을 통한 연구가 활발히 진행되었으며, 결론적으로 윤동주 연구는 정신사적 측면(주체성, 저항성, 종교성)에서의 연구, 전기적 사실에 대한 연구, 문학사적 위치에 관한 연구, 비교문학적 연구, 원전 확정에 관한 연구, 형식적 측면의 연구 등으로 나눌 수 있다고 분석하고 있다.

3. 연구설계

3.1 연구 개요

본 연구의 연구 개요는 〈그림 2〉와 같다. 우선 연구 대상이 되는 윤동주 관련 연구 논문을 수집하기 위해 '윤동주'라는 키워드를 기반으로 논문을 수집하였다. 검색된 결과를 대상으로 분석에 활용할 논문명, 초록, 발행년도 및 참고문헌 정보 추출을 위한 원문 PDF를 수집하였다. 수집된 논문을 대상으로 동시인용정보, 작품 및 작가의 개체명을 추출한 후 네트워크 분석 및 시각화를 수행하였다.



〈그림 2〉 연구 개요

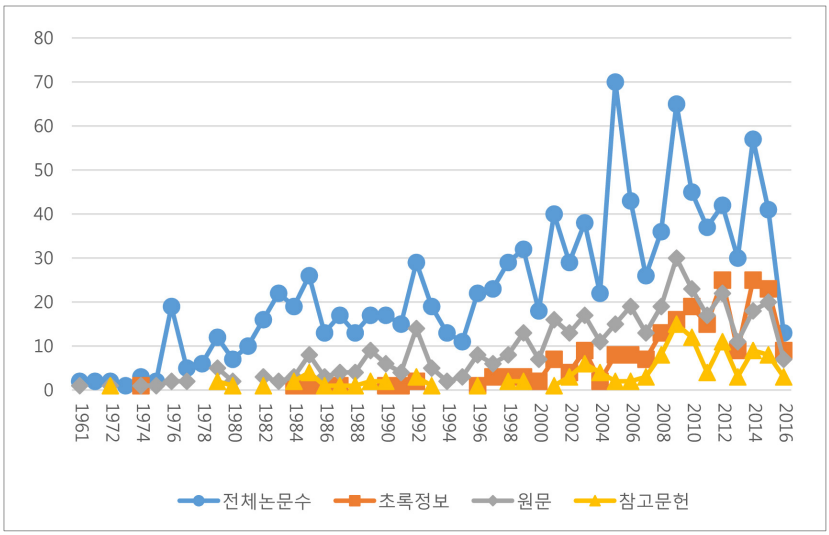
3.2 데이터 수집

분석대상으로 한 데이터는 윤동주에 관한 연구이며, '윤동주'를 키워드로 하여 국내 학술지, 학위논문 데이터베이스인 Dbpia와 RISS에서 검색하였다. 그 결과 Dbpia에서는 432건, RISS에서는 1,062건이 검색되었다. 영문 논문도 찾기 위해 Web of Science에서 Yun dong-ju, dongju, dong-ju라는 키워드로 검색하여 87건이 검색되었지만 이는 모두 한글 논문들의 영문 버전이었고, RISS에서 검색된 94건의 논문도 대부분 한글 논문들의 영문 버전이었던 관계로 데이터 중복을 방지하기 위해 영어 논문은 연구대상에서 제외하였다. 결과적으로 수집된 총 논문 1,581건 중에서 영어 논문 181건과 저자정보가 없는 논문을 제외한 1,076건의 데이터를 분석대상으로 설정하였다. 〈표 1〉에서 보듯이 전체 1,076건의 데이터에서 초록을 포함하고 있는 논문은 220건이었다. 또한 원문을 다운로드 받을 수 있었던 논문 398건 중 인용정보를 추출할 수 있었던 논문은 121건이었으며 이를 대상으로 2,569건의 인용정보를 추출할 수 있었다.

수집된 논문은 1961년부터 2016년까지 국내 학술지 및 학위논문이며, 〈그림 3〉에서 보는 것과 같이 연도별로 편차는 존재했으나 연구가 양적으로 꾸준히 증가하고 있음을 볼 수 있다. 또한 초록 분석, 인용 정보 분석에 사용된 문헌이 특정 연도에 치우치지 않고 연도별 연구의 양에 따라 비슷한 비율로 사용되었음을 볼 수 있다.

〈표 1〉 분석 방법별 데이터 건수

분석 기법	분석 대상	수집 건수	정보원
토픽모델링을 통한 연구 주제분석, 개체명 추출을 통한 관계 분석	초록 정보	1,076건 중 220건	Dbpia, RISS
저자동시인용 분석	참고문헌(인용정보)	121건의 논문에서 2,569건 추출	Dbpia



〈그림 3〉 분석에 사용된 연도별 논문 현황

4. 데이터 분석방법 및 분석결과

4.1 토픽모델링을 통한 연구 주제 분석

각각의 다른 윤동주 연구들 속에서 개체들이 갖는 의미를 추론하기 위해 윤동주를 검색어로 하여 추출된 초록 220건을 대상으로 LDA 기반 토픽모델링을 수행하였다. 한글 형태소 분석기 Komoran(Shineware 2014)을 이용하여 형태소 분석 및 불용어를 제거하고 명사만을 추출 하였다. 또한 '윤동주'와 같이 출현빈도가 과도하게 높아 성능을 저해할 수 있는 단어들은 제거하였다. 토픽모델링은 Mallet Topic Modeling Package를 사용하였으며, 5개, 10개, 20개, 30개의 토픽으로 모델링 후 도출된 토픽을 살펴보았다. 데이터 사이즈와 토픽이 표현하는 주제를 종합하여 5개의 토픽으로 최종 결정하였다. 토픽 수는 LDA 기반 토픽 모델링의 제한점 중 하나이다. 토픽 수는 연구자가 사전에 정의하게 되는데 개수가 적으면 토픽이 일반적이게 되며 다른 토픽과 잘 구별되는 장점이 있다. 또한 토픽의 개수가 많으면 다른 토픽과 미묘한 차이가 나거나 의미적으로 겹칠 확률이 높아진다. 따라서 이런 문제를 해결하기 위해 다양한 개수를 지정해서 모델링 한 후 연구자가 최적의 개수를 판단하는 방법을 사용할 수 있다(Asuncion 2010).

〈표 2〉는 토픽모델링 결과로 나타난 5개의 연구 영역과 주요 단어를 보여주고 있으며 크게 두 가지 영역으로 구분 지을 수 있다. 윤동주의 삶과 시대적 상황을 중심으로 연구한 토픽 1, 2와 주요 작품인 『간』, 『자화상』 및 다른 작품을 대상으로 한 시에 대한 연구이다. 첫 번째 토픽으로 나타난 식민지, 문학사, 강점기, 디아스포라 등은 윤동주의 출생과 살아온 시대를 대변하는 단어들이다. 특히 디아스포라는 흩어진 사람들이라는 뜻으로 북간도 또는 만주국, 즉 윤동주 출생에 대한 다문화주의 연구가 많이 나

〈표 2〉 윤동주 연구 토픽모델링

윤동주 생애와 관련된 시대적 배경	기독교 관점 윤동주 연구	『간』, 시적 세계, 조선족 문학	『자화상』을 중심으로 한 연구	교과서 수록 시 연구
topic 1	topic 2	topic 3	topic 4	topic 5
식민지	기독교	프로메테우스	교과서	텍스트
문학사	정지용	조선족	자화상	이미지
그리움	상상력	고등학생	서정주	학습자
아이러니	세계관	문제점	참회록	상상력
강점기	문학사	감상문	기념비	김소월
디아스포라	식민지	이야기	문학관	교과서
서정시	박두진	낭만주의	김수영	스스로
마지막	부끄러움	한국어	기념관	정지용
어머니	이미지	문익환	자의식	목소리
부끄러움	가운데	우크라이나	학습자	마지막

타나고 있다. 주요 연구로는 오문석(2012)의 윤동주와 다문화적 주체성의 문학이 있으며, 주로 윤동주가 태어나고 살았던 환경이 윤동주 시에 어떤 영향을 주었는지에 대한 주제를 다루고 있다.

두 번째 토픽으로 나타난 단어는 기독교, 세계관, 정지용으로 기독교 신자 윤동주의 삶과 기독교적 세계관에 대한 연구와 관련되어 있다. 주요 연구로는 이승하(1999)의 『일제하 기독교 시인의 죽음의식 - 정지용·윤동주』, 류양선(2011)의 『윤동주의 시에 나타난 기독교 신앙 - 「十字架」를 중심으로』 등이 있다. 또한 정지용과 박두진은 시와 기독교의 맥락에서 비교 연구가 이루어지고 있음을 알 수 있다.

세 번째는 윤동주의 시적 세계에 대한 연구, 특히 『간』에 대한 연구와 관련된 토픽이다. 윤동주의 『간』은 한국의 귀토설화와 그리스 신화를 모티브로 쓴 작품으로 혁명적 낭만주의 시인인 셸리의 영향을 받은 것으로 분석되고 있다(박호영 2012). 또한 문익환 목사에 끼친 영향에 대한 연구도 일부 나타나고 있다.

네 번째는 『자화상』을 중심으로 한 윤동주 작품을 연구한 토픽이다. 또한 자화상이 실린 교과서에 대한 연구도 포함하고 있다. 윤동주의 『자화상』은 윤동주의 대표작 중 하나로 작품 자체에 연구뿐만 아니라 다른 시인의 작품과도 많이 비교연구 되고 있다. 윤동주와 서정주는 『자화상』이란 작품으로 많이 연구되고 있다. 제목이 같고 작가가 다른 두 작품에서 자기인식, 자기표현 등에 대한 연구가 이루어지고 있다. 또한 노천명의 『자화상』과도 비교연구 되고 있다.

다섯 번째는 교과서에 수록된 시 연구이다. 윤동주 시는 김영랑, 김소월, 박두진, 김춘수 등의 다른 작가의 시와 함께 교과서에 실려 있다. 교과서에 실린 시의 연구에서 다른 시인과 비교 또는 윤동주 시 단독으로 연구되고 있다. 대표적으로 정은아(2016)의 『윤동주 시 교육 방법론 연구: 〈별 헤는 밤〉의 '읽기 전 활동'을 중심으로』와 같은 윤동주 시에 대한 교육 방법론적인 연구가 있다.

이러한 분석결과는 앞서 소개한 선행연구와

도 맥락을 같이 하고 있음을 알 수 있었다. 김형태(2015)의 연구에서는 윤동주에 관한 연구를 시기별로 특성을 나누었는데 70년대에 저항성의 문제를 중심으로 윤동주의 문학사적 위치에 대한 논의가 이루어 졌다고 분석한 것은 토픽 1과 맥락을 같이하고 있으며, 80년대에는 윤동주의 전기적 사실에 중점을 두었던 이전 연구들에 비해 개별 작품 해석에 중점을 두며 기독교 의식과 자의식 양상에 관한 연구가 진행되었다는 것은 토픽 2와 맥락을 같이한 것으로 볼 수 있다. 90년대는 기존의 연구성과를 바탕으로 비교연구가 활발하게 진행되었다고 분석하였는데 이는 토픽 5번과 같은 내용임을 볼 수 있었다.

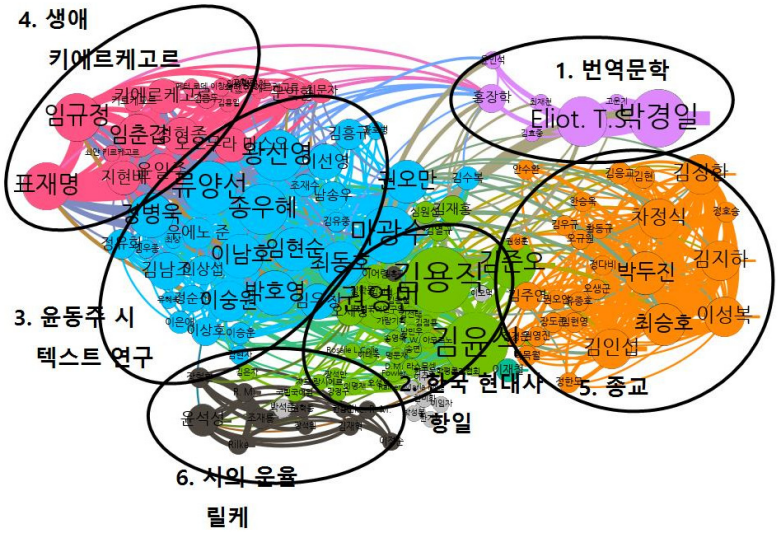
4.2 저자동시인용 분석을 통한 연구 주제 분석

저자동시인용은 어떤 연구에 있어 다른 논문 2건 이상을 동시에 인용하는 것을 나타낸다. 동시인용 빈도가 높다는 것은 두 저자의 연구가 인용하는 연구에 영향을 준다는 의미이므로, 동시인용 빈도를 이용하여 연구의 주요 개념이나 방법론을 유추해 낼 수 있다. 저자동시인용 분석을 위하여 윤동주 관련된 연구 원문 121건을 수집하였으며, 2,569건의 인용정보를 추출하였다. 이를 활용하여 45,961회의 동시인용정보를 추출하였으며, 5회 이상 동시인용된 저자를 대상으로 261개 노드와 1,075개의 엣지로 구성된 네트워크를 작성하였다. Average Degree는 8.679로 나타났으며, Density는 0.033으로 나타났다. 모듈리티는 0.542로 나타났으며 모듈리티를 활용하여 커뮤니티를 분석한 결과 12개로 나타났다. 그 중 유의한 6개를 대상으로 각 커뮤니티별 연결 중심성 상위에 있는 저자를 통하여 커뮤니티의 특성을 추출하고 이름을 부여하였으며, 〈그림 4〉처럼 커뮤니티 네트워크를 구성하였다.

각 커뮤니티의 라벨링을 위해 참고문헌 데이터를 바탕으로 각 커뮤니티별 연결 중심성 상위 저자의 논문 제목을 수작업으로 분석하였다. 첫 번째 커뮤니티에 속한 박경일, Eliot. T.S, 윤인석, 김효중, McCann. David. R, 리쾨르 폴

〈표 3〉 저자동시인용 커뮤니티 분석

번역문학	한국 현대사, 항일	윤동주 시 텍스트 연구	생애 키에르케고르	종교	시의 운율, 릴케
박경일	김용직	마광수	임규정	박두진	윤석성
Eliot. T.S.	김윤식	류양선	표재명	최승호	조재룡
홍장학	권영민	왕신영	임춘갑	김인섭	장철환
이재철	김준오	송우혜	윤일주	이성복	Rilke
윤인석	오세영	최동호	정현종	김지하	김재혁
김효중	김재홍	박호영	키에르케고르	김정환	R. M.
고운기	심원섭	정병욱	오오무라마스오	차정식	Rilke. R. M.
최재천	김열규	이남호	지현배	김주연	이정순
McCann. David. R.	김학동	임현순	문익환	김응교	권혁웅
리쾨르 폴	이어령	권오만	최문자	김현	장석원



〈그림 4〉 저자동시인용 네트워크

등은 인용된 참고문헌의 제목을 살펴본 결과 외국문학을 다루거나 번역을 주제로 논문을 작성했다는 공통점을 확인할 수 있었다. 두 번째 커뮤니티에 속한 김용직, 김윤식, 권영민 등 저자의 문헌들은 한국 현대사와 관련된 윤동주 연구 또는 일제 시대와 같은 시대상을 반영한 연구를 수행한 저자들이 모여 있었다. 세 번째 커뮤니티에 속한 마광수, 송우혜, 최동호, 박호영의 문헌은 윤동주 시의 상징적 표현에 관한 연구 및 문체론, 실증적 접근 등 윤동주 시 자체적인 텍스트를 분석한 논문들이었다. 네 번째 커뮤니티에 속한 임규정, 표재명, 임춘갑은 키에르케고르를 연구하였으며, 정현종, 오오무라마스오, 지현배, 문익환은 윤동주의 생애와 관련된 논문의 저자라는 공통점이 있었다. 다섯 번째는 박두진, 최승호, 김인섭, 김지하, 김정환, 차정식 등 기독교와 신학, 예수, 종교 등의 주제를 다룬 저자들인 것으로 확인할 수 있었으며, 여섯 번째 커뮤니티에 속한 윤석성, 조재룡, 장철환, 권혁웅, 장석웅은 시의 운율과 리듬에 대해서 공통적으로 논했으며, 김재혁, 이정순 릴케와 윤동주를 연구한 저자라는 공통점을 발견할 수 있었다.

이러한 분석결과 또한 선행연구(김형태 2015)에서 분석한 윤동주 연구의 동향과도 비슷한 맥락을 보여주고 있다. 윤동주의 연구들은 크게 주체성, 저항성, 종교성으로 나누어 볼 수 있는 정신사적 측면에서의 연구, 전기적 사실에 대한 연구는 두 번째와 다섯 번째 커뮤니티에서 찾아볼 수 있으며, 개별작품 해석에 중점을 둔 문학사적 위치에 관한 연구는 세 번째 커뮤니티에서 찾아볼 수 있다. 또한 비교문학적 연구는 첫 번째 커뮤니티와 네 번째, 여섯 번째 커뮤니티에서 찾아볼 수 있었다.

4.3 개체명 추출(인명, 작품명)을 통한 개체간의 관계 분석

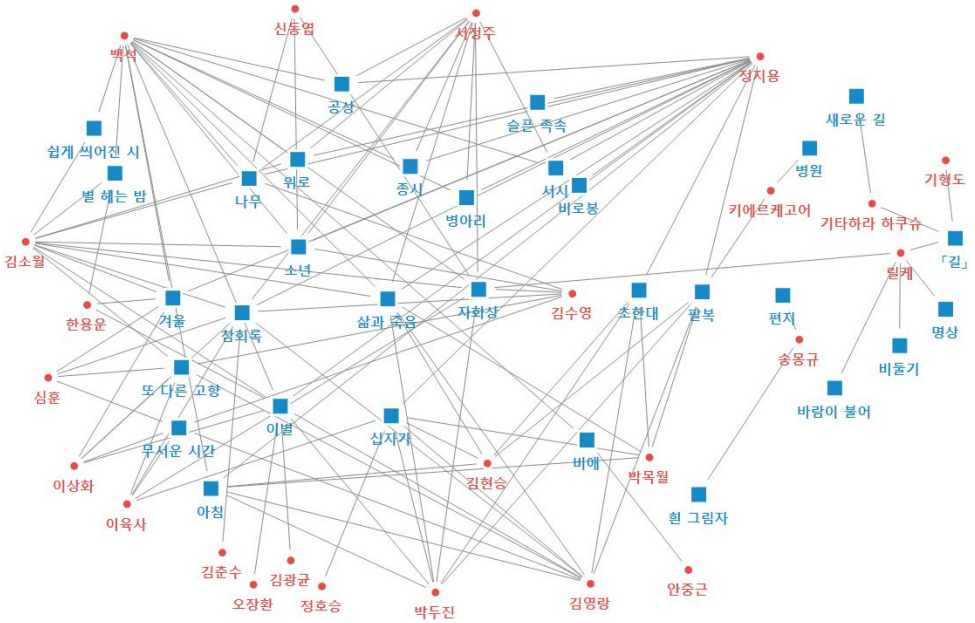
토픽모델링에서는 초록에 포함된 주요 단어

들을 대상으로 확률적 분석을 통해 토픽을 발견하는 분석방법이며, 저자동시인용 분석은 연구자를 중심으로 학문분야의 지적구조를 파악할 수 있는 분석방법이다. 본 논문에서는 이러한 분석방법과 더불어 문학작품 연구에 있어서 다루어지고 있는 인물 및 작품의 관계를 추론해 내기 위해 개체계량학적 분석방법을 시도하였다. 인물 및 작품 연구, 비교 연구가 주를 이루는 윤동주 연구의 초록에서 나타난 주요 개체는 인명, 작품명이다. 윤동주와 다른 시인 또는 작품을 비교연구 할 때 주로 윤동주의 작품에 나타난 내용과 다른 작가의 작품에 나타난 내용을 비교하는 방식으로 이루어지고 있다. 따라서 윤동주의 작품과 다른 작가의 인명이 초록에서 동시 출현하는 빈도를 분석하게 되면 윤동주의 작품 중 어떤 작품이 다른 작가와 많이 비교되는지를 추론할 수 있다.

작품명은 개체명 중에서도 독특한 형식을 가지고 있다. 박용민과 이재성(2014)은 한국어 제목 개체명 인식 및 사전 구축에서 제목 개체명의 특징을 다음과 같이 정의하고 있다. 첫째, 단어부터 문장까지 다양한 형태를 가지고 있으며, 제목이 다른 개체명이 되는 경우도 있다. 예를 들어 베를린은 지역명이면서도 제목 개체명이 될 수 있다. 또한 제목 개체명은 개체명 인식에 사용할 수 있는 특별한 자질을 가지고 있지 않다. 예를 들어 광역시, 특별시가 접미사로 붙으면 지역 개체명으로 인식할 수 있지만 제목은 이러한 특징을 가지고 있지 않다. 따라서 제목과 같이 형태의 다양성으로 인하여 개체명 인식이 불분명할 경우에는 사전을 구축하여 사전과 비교하여 해결하는 것이 효과적이라고 보았다. 따라서 본 연구에서는 윤동주 연구에서 나타나는 작품명 개체를 분석하기 위하여 사전 방식을 사용하였다. 사전 구축은 윤동주 연구의 초록에서 나타나는 개체를 대상으로 하였으며 사전 구축을 위한 기초데이터를 수집하기 위해 초록에서 [ ], “ ”, 「 」, 『 』, ‘ ’ 등과 같이 특수문자로 묶여 있는 단어들을 추출하였다. 작품명 사전은 1차로 특수문자를 근거로 추출된 목록에 윤동주 작품 목록을 추가하여 구축하였다. 인명 사전은 특수문자로 묶이지 않은 경우도 다수 발견되어 초록을 검토 후 일부 인명을 추가 하였다. 각각 구축된 건수는 인명사전은 52건, 작품명 사전은 109건이다. 이렇게 구축된 사전을 두 가지 사전을 기반으로 초록에서 일치하는 인명 개체명과 작품명 개체명의 동시 출현 빈도를 추출 하였다. 인명 개체는 추출 대상 초록 220건에서 656개의 개체를 추출하였으며, 33개의 고유한 개체가 추출되었다. 작품명은 총 374회 추출되었으며, 67개의 고유한 개체가 추출되었다. 하지만 ‘이상’이라는 작가는 ‘형이상학’, ‘이상적인’ 등의 단어 속에도 포함되어 있었기 때문에 윤동주와 비교 연구가 많음에도 불구하고 결과의 오류값이 커져 제외하였다. 마찬가지로 ‘바다’라는 작품도 ‘바다와 나비’, ‘바다는 누가 울은 눈물인가’ 등과 같이 ‘바다’라는 작품 하나를 대표할 수 없었기 때문에 삭제하였다.

〈그림 5〉는 추출된 개체명 동시출현을 에지로 하여 이진 네트워크를 구축한 결과이다. 네트워크는 총 93개 노드와 304개 엣지를 가지고 있으며, 윤동주를 중심으로 문학 작품과 연결되어 있으며, 또한 연결된 문학 작품을 매개로 다른 작가와 연결된 모습을 볼 수 있다.

윤동주 연구의 많은 부분은 다른 시인과 작



〈그림 5〉 작품명과 인명 네트워크

〈표 4〉 동시출현빈도 상위 20개 작가와 작품명

작가	작품명	출현빈도	작가	작품명	출현빈도
서정주	자화상	7	백석	아침	2
백석	위로	4	백석	병아리	2
백석	나무	4	백석	삶과 죽음	2
한용운	이별	4	백석	겨울	2
김광균	이별	2	백석	소년	2
백석	공상	2	백석	서시	2
백석	쉽게 씌어진 시	2	백석	종시	2
백석	이별	2	백석	자화상	2
백석	별 헤는 밤	2	서정주	소년	2
백석	비애	2	서정주	나무	2

품들의 내용 비교이다. 이 네트워크를 통해 다른 작가의 작품과 윤동주의 작품이 같이 연구될 때 작가마다 다른 작품을 중심으로 비교되고 있음을 보여 준다. 〈표 4〉는 다른 작가와 윤동주 작품의 동시출현 빈도를 나타낸 것이다. 이 표에서 다른 작가와 윤동주 작품의 동시출현 빈도는 서정주, 백석, 정지용을 중심으로 나타났다. 서정주의 경우에는 '자화상'이란 동일한 제목의 작품을 통해서 자연스럽게 비교연구가 많이 수행되었지만 백석, 정지용에 비해 연결되어 있는 작품이 다소 적다는 것을 그림에서 볼 수 있다. 이는 서정주라는 시인은 일부

작품에 한정하여 많은 분석이 이루어지고 있음을 나타내는 것이다. 이와는 대조적으로 백석의 경우 단일작품과 연결된 빈도는 낮지만 매우 다양한 작품과 매개되어 있는 것을 볼 수 있었다. 또한 윤동주의 작품을 매개로 한 다른 작가의 연결을 구체적으로 살펴본 결과 '백석'과 '한용운'은 '서시'와 '별 헤는 밤'이라는 작품과 매개되어 있는 모습을 볼 수 있는데 분석한 논문 중에서는 이들 간의 관계가 한번에 들어가 있는 논문은 없었다. '한국 근대시 연구(김병호 2002)' 논문에서 백석과 한용운의 이름만 함께 언급되었고, '한국 근대시의 만주체험 - 시적 형상화와 그 의미(윤여탁 2015)'라는 논문에서 '백석'과 '별 헤는 밤'이라는 작품의 연결을 볼 수 있었으며, '소리-뜻을 중심으로 구성되는 현대시의 리듬(권혁웅 2013)' 논문에서 '한용운'과 '별 헤는 밤'의 연결을 볼 수 있는 것으로 보아 시대적 상황과 시의 리듬이라는 토픽 안에서 윤동주의 별 헤는 밤 - 백석 - 한용운의 관계가 맺어지고 있다는 사실을 발견할 수 있었다. 즉 개체계량학적 분석을 통해 토픽모델링이나 저자동시인용 분석에서는 확인하기 어려웠던 작품과 작가의 관계를 추론할 수 있었으며, 이러한 방법론의 적용은 윤동주의 작품 중 다른 작가와 많이 비교되는 작품을 볼 수 있는 한편 다른 작가와 어떤 작품을 매개로 비교 연구 또는 공동 연구 되는지를 살펴볼 수 있을 것이다.

5. 결 론

본 논문에서는 텍스트마이닝을 이용한 개체계량학적 방법론을 윤동주 관련 연구에 한정한 인문학 분야에 적용해 봄으로써 윤동주라는 인물에 관해 데이터 기반으로 보다 객관적으로 분석해 볼 수 있었다. 분석방법은 크게 세가지였으며, 서지계량학적 연구에서 주로 사용되고 있는 토픽모델링분석, 저자동시인용 분석을 수행하는 한편 새로운 방법론으로 부각되고 있는 개체계량학적 분석을 적용해 봄으로써 전체적인 연구의 동향과 지적구조 분석은 물론 윤동주의 작품을 매개로 한 다른 작가와 작품들의 관계까지 확인해 볼 수 있었다.

이러한 연구결과는 선행연구에서 제시했던 기존 연구자의 질적 연구 결과와도 맥락을 같이했기 때문에 개체계량학적 방법론이 인문학 분야 연구자들의 연구 시작단계에 있어서 매우 효율적이고 정확한 결과를 제시해줌으로써 연구 방향을 잡는데 큰 기여를 할 것으로 판단된다. 하지만 개체계량학적 방법을 적용할 때에는 연구대상으로 하는 데이터 양이 풍부해야 하며, 텍스트마이닝 기술을 이용해 분석할 수 있는 데이터 형식을 갖추고 있어야 한다는 한계점이 있다. 특히 국내 인문학 분야 연구는 한자와 한글을 혼용 사용하는 경향이 있고, 오래되어 전자적 형태로 분석할 수 없는 논문들이 상당 수 있었으며, 참고문헌 정리 방법과 양식이 다양하기 때문에 정확한 분석을 위한 데이터 전처리에 소요되는 시간이 많았으며 실제로 이번 연구에서 분석에 활용된 데이터는 매우 제한적이었다.

본 연구를 시발점으로 인문학 분야 연구 논문에 대한 텍스트마이닝 기법을 활용한 다양한 개체계량학적 분석이 활성화되기를 기대한다.

참 고 문 헌

권혁웅. 2013. 소리-뜻을 중심으로 구성되는 현대시의 리듬: 「님의 침묵」, 「별헤는 밤」을 중심으로. 『한국문학이론과 비평』, 59: 27-48.

김병호. 2002. 『한국 근대시 연구』. 박사학위논문. 중앙대학교 대학원, 문예창작전공.

김학용. 2012. 대하소설 토지 등장인물 네트워크의 동적 변화 분석. 『한국콘텐츠학회논문지』, 12(11): 519-526.

김형태. 2015. 『윤동주 시의 실존의식 연구』. 박사학위논문. 한국교원대학교 대학원, 국어교육전공.

김희전, 조현양. 2010. 저자동시인용분석과 저자서지결합분석에 의한 지적 구조 분석. 『정보관리학회지』, 27(3): 283-306.

류양선. 2011. 윤동주의 시에 나타난 기독교 신앙. 『한국시학연구』, 31: 141-168.

박용민, 이재성. 2014. 한국어 제목 개체명 인식 및 사전 구축: 도서, 영화, 음악, TV 프로그램. 『정보처리학회논문지/소프트웨어 및 데이터 공학』, 3(7): 285-292.

박자현, 송민. 2013. 토픽모델링을 활용한 국내 문헌정보학 연구동향 분석. 『정보관리학회지』, 30(1): 7-32.

박호영. 2012. 일제강점기 바이런과 셸리의 수용과 의의. 『어문연구』, 40(4): 277-295.

오문석. 2012. 윤동주와 다문화적 주체성의 문학. 『한국근대문학연구』, 25: 149-176.

윤순근. 1992. 「황무지」연구의 계량서지학적 고찰. 『서지학연구』, 8: 135-206.

윤여탁. 2015. 한국 근대시의 만주 체험 - 시적 형상화와 그 의미 -. 『한중인문학연구』, 46: 121-140.

이승하. 1999. 일제하 기독교 시인의 죽음의식 - 정지용・윤동주의 경우. 『어문논집』, 27: 133-161.

이재윤. 2005. 문헌동시인용 분석을 통한 한국 문헌정보학의 연구 전선 파악. 『정보관리학회지』, 32(4): 77-106.

이재윤. 2008. 서지적 저자결합분석. 『정보관리학회지』, 25(1): 173-190.

정은아. 2016. 윤동주 시 교육 방법론 연구. 『우리문학연구』, 49: 375-402.

함정은, 송민. 2015. 인용정보를 고려한 미발견 공공 지식 추출 - Swanson의 ABC 모델 재현 및 확장. 『정보관리학회지』, 32(2): 87-103.

Asuncion, H. U., A. U. Asuncion, and R. N. Taylor. 2010. "Software traceability with topic modeling." *Proceedings of the 32nd ACM/IEEE International Conference on Software Engineering-Volume 1*, 95-104

Blei, D. M. 2012. "Probabilistic topic models." *Communications of the ACM*, 55(4): 77-84.

Blei, D. M., A. Y. Ng, and M. I. Jordan. 2003. "Latent dirichlet allocation." *Journal of Machine Learning Research*, 3(Jan): 993-1022.

Ding, Y., M. Song, J. Han, Q. Yu, E. Yan, L. Lin, and T. Chambers. 2013. "Entitymetrics: Measuring the impact of entities." *PloS one*, 8(8): e71416.

Griffiths, T. L. and M. Steyvers. 2004. "Finding scientific topics." *Proceedings of the National Academy of Sciences*, 101(suppl 1): 5228-5235.

Hammarfelt, B. 2011. "Citation analysis on the micro level: The example of Walter Benjamin's Illuminations." *Journal of the American Society for Information Science and Technology*, 62(5): 819-830.

McCain, K. W. 1983. "The author cocitation structure of macroeconomics." *Scientometrics*, 5(5): 277-289.

Shineware. 2014. 한글형태소분석기 Komoran 2.0.

Small, H. 1973. "Co-citation in the scientific literature: A new measure of the relationship between two documents." *Journal of the American Society for Information Science*, 24(4): 265-269.

Song, M. and S. Y. Kim. 2013. "Detecting the knowledge structure of bioinformatics by mining full-text collections." *Scientometrics*, 96(1): 183-201.

Tonussi, P. 2013. "Branwell Brontë and TS Eliot, April Rain and Aching Memories: History of a Reading?" *Brontë Studies*, 38(2): 139-144.

Wang, X. and A. McCallum. 2006. "Topics over time: a non-Markov continuous-time model of topical trends." *The 12th ACM SIGKDD international conference on Knowledge discovery and data mining*.

White, H. D. 1990. "Author co-citation analysis: Overview and defense." *Scholarly Communication and Bibliometrics*, 84: 106.

White, H. D. and B. C. Griffith. 1981. "Author cocitation: A literature measure of intellectual structure." *Journal of the American Society for Information Science*, 32(3): 163-171.

• 국문 참고자료의 영어 표기
(English translation / romanization of references originally written in Korean)

Chung, Eun-Ah. 2016. "A Study on the Methodology of Yun Dong Ju's Poetry Education - Focusing on a Pre-Reading Activity About the Poem: 〈Night of Counting Star〉." *The Studies of Korean Literature*, 49: 375-402.

Ham, Jung Eun and Song Min. 2015. "Detection of Hidden Knowledge Using a Citation-Based Approach Based on Swanson's ABC Model." *Journal of the Korean Society for Information*

Management, 32(2), 87-103.

Kim, Byung-Ho. 2002. *A Study on Modern Korean Poetry*. PhD diss. Chung-Ang University.

Kim, Hak-Yong. 2012. "Analysis of Network Dynamics from the Roman-Fleuve, Togi." *Journal of the Korea Contents Association*, 12(11): 519-526.

Kim, Hee-Jeon and Hyun-Yang Cho. 2010. "A Study on Intellectual Structure Using Author Co-Citation Analysis and Author Bibliographic Coupling Analysis in the Field of Social Welfare Science." *Journal of the Korean Society for Information Management*, 27(3): 283-306.

Kim, Hyung-Tae. 2015. *Study on the existential consciousness in the poetry of Yun Dong-Ju*. PhD diss. Korea National University of Education.

Kwon, Hyuk-woong. 2013. "The Study on the Rhythm which Consists of a Sound-Meaning (Prosodie) in Korean Modern Poetry." *Korean literary theory and criticism*, 59: 27-48.

Lee, Jae-Yun. 2005. "Identifying the Research Fronts in Korean Library and Information Science by Document Co-citation Analysis." *Journal of the Korean Society for Information Management*, 32(4): 77-106.

Lee, Jae-Yun. 2008. "Bibliographic author coupling analysis: a new methodological approach for identifying research trends." *Journal of the Korean Society for Information Management*, 25(1): 173-190.

Lee, Seung-ha. 1999. "The death ceremony of a Christian poet under Japanese rule: In the case of Jung, Ji-yong and Yun, dong-ju." *The Journal of Language and Literature*, 27: 133-161.

Oh, Moon-seok. 2012. "Multicultural Subjectivity in Yun Dong-ju's Literature." *Journal of Modern Korean Literature*, 25: 149-176.

Park, Ho-young. 2012. "Reception and Significance of Byron and Shelley during the Japanese Ruling Era of Korea." *The Society for Korean Language and Literary Research*, 40(4): 277-295.

Park, Ja-Hyun and Min Song. 2013. "A study on the research trends in library & information science in Korea Using topic modeling." *Journal of the Korean Society for Information Management*, 30(1): 7-32.

Park, Yong-min and Jae-Sung Lee. 2014. "Named Entity Recognition and Dictionary Construction for Korean Title: Books, Movies, Music and TV Programs." *Korea Information Processing Society*, 3(7): 285-292.

Ryu, Yang-seon. 2011. "The Christian faith in Yun, Dong-ju's poetry." *The Korean Poetics*

Studies, 31: 141-168.

Yoon, Yeo-Tak. 2015. "A study on Manchuria Erlebnis of Korean Modern Poetry." *Studies of Korean & Chinese Humanities*, 46: 121-140.

Yun, Soon-Keun. 1992. "A Study of the bibliometrics on 「the Waste Land」 by T.S. Eliot." *The Institute of Bibliography*, 8: 135-206.