인문논총 제78권 제1호 (2020.2.28), pp. 221~262
[DOI] https://doi.org/10.17326/jhsnu.78.1.202102.221

『개벽』 논조의 사회주의화에 관한 새로운 접근*

— 토픽 연결망 분석을 중심으로

허 수**

[초 록]

본 논문에서는 『개벽』 후기 논조의 사회주의화가 개벽 주도층에게도 비슷한 영향을 끼쳤는지 여부를 규명하였다. 분석 방법으로는 토픽 모델링과 연결망 분석 및 통계적 검정 방법을 사용하였다. 분석 절차

* 이 연구는 2019년 서울대학교 미래기초학문분야 기반조성사업으로 지원되는 연구비에 의하여 수행되었음.
이 논문은 2018년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF2018S1A6A3A01022568).
이 논문은 2020년 10월 10일 성균관대학교 동아시아학술원 대동문화연구원・한림대학교 한림과학원 및 아시아문화연구소가 공동으로 개최한 '개벽의 백년 백년의 개벽 — <개벽>으로 다시 여는 매체인문학 연구의 새 지평'(화상 학술대회) 때 발표한 '텍스트 마이닝을 활용한 『개벽』의 논조 분석 — 거시적 구조의 변화를 중심으로 —'를 대폭 수정한 것이다.
** 서울대학교 인문대학 국사학과 부교수

주제어: 개벽, 논조, 사회주의, 개벽 주도층, 토픽 모델링, 연결망 분석, 통계적 검정, 주요 논설
Gaebyeok, Tone, Socialism, *Gaebyeok* Leadership, Topic Modeling, Network Analysis, Statistical Test, Main Editorial

는 크게 세 단계를 밟았다. 첫째, 사회주의화를 판별할 지표를 정하였다. 『개벽』 주요 논설 334개의 전산 자료를 전처리한 뒤, 토픽 모델링을 사용하여 7개의 주제와 104개의 토픽 단어를 추출하였다. 그 중에서 토픽1에 속하는 20개의 단어를 사회주의 판별의 가장 중요한 지표로 간주하였다. 둘째, 논조의 사회주의화 양상을 살펴보았다. 26,060개의 '문서 — 토픽 단어 — TFIDF값'을 입력하여 7개 토픽 간의 관계를 그린 토픽 연결망 지도를 산출하였다. 그 결과 『개벽』 전기에는 연결망의 중심이 개벽의 개조론에 있었으나, 후기에는 그 중심이 사회주의로 이동하였음을 확인할 수 있었다. 셋째, 사회주의가 개벽 주도층에 미친 영향을 살펴보았다. 먼저 후기의 기명(記名) 논설 121개를 '개벽 주도층'과 '일반 필자층'으로 양분하였다. 다음으로 이 자료를 사용하여 두 집단 사이에서 토픽1과 토픽8의 비중에 각각 유의미한 차이가 있는지 여부를 'T검정'으로 살펴보았다. 그 결과 개벽 주도층에 대한 사회주의의 영향은 여타 필자들에 비하여 통계적으로 유의미하게 낮았음이 드러났다.

1. 머리말

『개벽』(1920~1926)에서 1923~1924년 무렵부터 사회주의적 논설이 급증하는 사실은 대부분의 학자들이 인정하고 있다.[1] 또한 이러한 증가 현상이 당시 한국 사회에서 사회주의가 유행하던 사실을 반영한 것이라는 데에도 이견은 없다. 그런데 이런 '사회주의화'가 『개벽』의 편집과 운영을 주도한 인물들의[2] 사상적 변화까지 동반한 것이었나

1) 최수일(2008), 『개벽 연구』, 서울: 소명출판사, p. 484 ; 김정인(2007), 「'개벽'을 낳은 현실, '개벽'에 담긴 희망」, 임경석 · 차혜영 외, 『『개벽』에 비친 식민지 조선의 얼굴』, 서울: 모시는 사람들, p. 244 참조.

2) 본 논문에서는 『개벽』을 이끌어 간 인물들로 이돈화, 김기전, 박달성, 차상찬의 네 명에 주목한다. 이하에서는 이들을 편의상 '개벽 주도층'이라 부른다.

에 관해서는 견해 차이가 있다.

허수(2011)는 이 문제를 다룬 선행연구들의 주장을 양분하여, 각각의 입장을 '외부의 상황 변화에 대한 수동적 동조'와, '능동적 대응을 통한 사회주의의 자기화'로 요약하였다.[3] 나아가 두 입장을 절충하는 견지에서 제3의 입장을 내놓았다. '개벽 주도층의 종교적 이상주의가 사회주의와 친연성을 가졌지만,『개벽』 후기에 와서도 주도층의 사상이 사회주의로 기울지는 않았다'는 것이 핵심 요지이다.[4] 그런데 허수(2011)에서는 표지와 목차를 소재로 한 형태적 분석을 중심으로 하였으므로, 내용 분석을 통해 그 주장을 검증하고 보완할 필요가 있다.

본 논문은 이러한 제3의 입장을 보완・발전시키는 취지에서『개벽』 논조의 사회주의화 양상을 좀 더 분석적으로 다루고자 한다. 지금까지『개벽』 논조의 사회주의화에 주목한 연구들은 대부분 이 주제를 서술적(descriptive)으로 다루는 데 그쳤다. 선행 연구들은『개벽』에서 사회주의 필진의 참여가 많아지고, 논설에서 '사회주의'와 '혁명', '계급', '무산'(無産) 등의 단어가 증가하는 양상 등을 중요한 근거로 삼았다. 이런 근거가 사회주의적 성향을 판별하는 데 중요한 것임은 분명하다. 그렇지만 필자가 보기에 선행연구들은 다음과 같은 핵심적 사항을 소홀하게 다루었다. 첫째, 사회주의적 성향을 판단할 수 있는 '객관적' 지표는 무엇인가,[5] 둘째,『개벽』의 '논조'는 무엇을 통해 살필 수 있으며, 그것은 사회주의 논설의 증가와 어떤 관계에 있는가, 셋째, 개벽

3) 허수(2011),『식민지 조선, 오래된 미래』, 서울: 푸른역사, 2011, pp. 80-82 및 pp. 114-116.

4) 허수(2011), pp. 80-82 및 pp. 114-116.

5) 본 논문에서 필자는 '객관적'이라는 단어를, '불편부당한'이나 '옳은'이라는 가치평가의 맥락보다는 '근거를 구체적으로 제시한'이라는 기능적 지표 판별의 맥락에서 사용한다. 이에 따라 필자가 '주관적'이라는 말을 사용할 경우에도, 그 단어를 '옳지 않은'이나 '치우친'이라는 의미보다는, '근거를 구체적으로 제시하지 않은'이라는 맥락에서 사용할 것이다.

주도층이 사회주의로 기울었는지 여부를 어떻게 판단할 수 있을까 등이다. 『개벽』의 사회주의화 문제를 학문적으로 다루기 위해서는, '사회주의', '논조 변화', '영향력' 등을 객관적 지표를 통해 판단하는 일이 중요하다.

이에 본 논문에서는 『개벽』 논조의 사회주의화 문제에 다음과 같이 접근하고자 한다. 첫째, 필자는 '사회주의'를 판단하는 근거를 사회주의 주제를 나타내는 단어들의 출현 빈도에서 구한다.[6] 단 그러한 사회주의적 단어의 추출은 '토픽 모델링'을 통해 『개벽』의 논설에서 내재적으로 획득할 것이다. 이는 곧, 그 근거를 사회주의 사상 사전이나 기타 '권위' 있는 외부 텍스트에 의존하지 않는다는 의미이다.[7] 둘째, '논조'를 '주제들의 합성(合成, vector)'으로 규정한다. 이에 따라 '논조의 사회주의화'는 여러 주제 중 사회주의가 중심 주제로 진입하는 것으로 간주한다. 또한 이러한 중심화 양상은 주제 간의 유사도를 시각화한 '토픽 연결망'을 통해 판별한다.[8] 셋째, 개벽 주도층의 사회주의

6) 본 논문에서는 '주제'라는 단어를, 영어의 'subject'나 'theme'보다는 더 좁고 구체적인 성격을 가진 'topic'에 가까운 의미로 사용하였다. 따라서 이 '주제'는 '변별적 의미를 이루는 단어 집합'으로 규정할 수 있다. 이 점에서 '주제'는 '토픽 모델링'의 '토픽'과 거의 동일하다. 토픽 모델링에 관하여는 2장에서 상술한다.

7) 필자는 기존 연구에서 사회주의 성향 판별이나 사회주의를 집약하는 단어 선정에 오류가 있었다고 생각하진 않는다. 그렇지만 다음 두 가지 면에서 기존 연구의 성과나 관행적 연구방법으로는 불충분하다고 보았다. 첫째, 사회주의자라고 해서 자신의 논설에서 항상 사회주의적 성향을 노출하거나 그러한 소재의 글만 쓴다고 볼 순 없다. 또한 사회주의적 입장을 개진하는 논설에서도 사회주의적 성향이 모든 문장이나 단락에 균질하게 분포하지도 않는다. 둘째, 연구자가 사회주의를 대표하는 단어를 현재나 당시의 권위 있는 자료에서 뽑아내어 지표로 사용한다고 해도 그러한 시도가 성공적일지에 대해서는 회의적이다. 왜냐하면 그렇게 선정한 단어들이, 실제 『개벽』에 나타난 사회주의적 논의를 잘 반영한다고 확신할 수 없기 때문이다.

8) '토픽 연결망'이라는 용어는 필자가 논지 전개의 편의상 만든 것이다. 이것은, 토픽 모델링으로 도출한 토픽들의 상호간 유사도를 산출한 뒤, 이 값을 기초로 연결망(Network)을 그린 것이다.

화 문제는 ‘T검정’과 같은 통계적 방법을 활용한다. 개벽 주도층의 논설에서 사회주의적 단어들이 여타 집단의 논설에 비해 현저하게 낮게 나오는지 여부가 핵심이 될 것이다.

『개벽』 분석에 토픽 모델링을 활용한 선구적 연구로는 이재연(2016)이 있다. 이재연은『개벽』에 관한 선행연구가 ‘신경향파 비평을『개벽』 전체의 주제적 맥락으로 오해’했다고 비판하였다. 그리고 토픽 모델링을 도입하여 ‘『개벽』의 전체 주제 구조를 파악’하였다.[9] 이런 관점과 방법은 크게 보아 본 논문의 문제의식과 접근법을 선취한 것이라 할 수 있다. 그렇지만 간과할 수 없는 차이점도 있다. 이재연(2016)에서 토픽 모델링의 활용은『개벽』의 전체 논지를 공시적으로 분석하는 데 주안점을 두었다. 통시적 분석은 ‘생명’과 ‘생활’이라는 두 단어에 관한 정성적 접근에 그쳤다.

이와 달리 본 연구에서는 토픽 모델링 방법을『개벽』의 논조 변화라는 포괄적이고 통시적인 문제를 해명하는 데 사용할 것이다. 2장에서는 사회주의를 판별할 지표를 정한다. 이를 위하여 토픽 모델링으로 논설의 전산 자료에서 주제군을 뽑아낸다. 특히 그 중에서도 사회주의 주제에 해당하는 단어에 주목하고 이를 사회주의를 판별할 핵심 지표로 간주할 것이다. 3장에서는 논조의 사회주의화 양상을 살펴본다. 우선, 토픽 단어들의 빈도 가중치를 근거로『개벽』의 논조 변화 마디를 2개 시기로 나눈다. 다음으로 각 시기별 토픽 연결망을 산출하고 이를 분석하여『개벽』 논조의 전・후기 양상을 비교한다. 4장에서는 사회주의가 개벽 주도층에 미친 영향을 살펴본다. 먼저 그 영향력을 관찰하는 데 필요한 지표를 정한 뒤, 후기의 논설을 개벽 주도층과 그 밖의 집단으로 구분한다. 그 다음에, 주요 지표가 되는 토픽들의 논설 내 비중 면에서 두 집단 사이에 유의미한 차이가 있는지 여부를 통계적 검

9) 이재연(2016),「토픽 모델링으로 본 <개벽>의 주제 지도 분석」,『상허학보』 46, 상허학회, pp. 296-297.

정 방법으로 살펴볼 것이다.

2. 『개벽』의 논조를 구성하는 주제들

2.1. 코퍼스 형성과 토픽 추출

본 논문에서는 『개벽』의 논조를 살펴볼 표본을 334개의 논설 기사로 한정하였다.[10] 이는 선행연구에서 '주요논설'이라 부른 것과 일치한다.[11] 주요논설에 해당하는 전산자료는 국사편찬위원회에서 제공하는

10) 『개벽』의 기사 수 및 논설의 규모는 연구자별로 집계에 다소 차이가 있다. 판본별 차이나 삭제 기사 포함 여부, 연구자들의 분류 기준 차이 등이 작용하였기 때문이다. 이에 관한 엄밀한 비교・검토는 본 논문의 범위를 넘어선다. 다만 전체적인 이해를 돕기 위하여 국사편찬위원회 제공 한국사데이터베이스의 관련 정보를 다음과 같이 제시해 둔다. 총 기사 수는 2,257개이며, 기사 분류 항목은 총 17개이다. 100개 이상의 기사를 가진 분류 항목 8개를 기사 수와 함께 내림차순으로 제시하면, 논설(570), 소식(358), 시(252), 사고・편집후기(249), 문예기타(185), 잡저(148), 소설(147), 문예평론(104) 순이다. 한편, 이 분류에서는 권두언이나 사설을 논설에 포함하기도 하고, 잡저나 사고・편집후기 등으로 분류하여 일관성이 떨어진다. 그렇지만 권두언과 사설은 『개벽』의 논조를 잘 반영한다고 간주하여, 본 논문에서는 검토 대상인 '주요 논설'의 범위에 포함시켰다. 334개 논설에서 권두언과 사설을 제외한 순수 '논설'은 303개로 이 값은 전체 570개 논설의 53.2%에 해당하는 규모이다.

11) 허수는 '주요논설'이 개벽 주도층의 입장과 메시지를 가장 잘 담고 있으며, 주로 목차나 실제 잡지 공간에서 앞부분에 위치한다고 했다. 또한 이런 이유로 주요논설은 다른 기사보다 더 중요하다고 강조하였다[허수(2011), p. 89]. 본 논문에서도 주요논설이라는 용어를 이런 입장에서 사용한다. 한편, 주요 논설을 선택하는 실제 작업에서는 각 호별 기사배치 상황의 차이를 고려하였다. 통상적으로는 1호부터 72호까지의 목차에 일단 의존하여, 맨 앞에서 적게는 2개, 많게는 5~6개의 논설을 골랐다. 특집으로 논설이 평소보다 많을 경우에는 평균보다 조금 더 많은 논설을 선별하기도 했다. 50~60호에서는 간혹 기사의 배열방식이 변하여, 본래 주요 논설이 있던 자리를 시사 기사가 차지하는 대신, 논설은 뒷부분에 위치하는 경우도 있었다. 이때에는 잡지의 앞부분이라는 위치 기준보다는,

한국사데이터베이스에서 구하였다.[12] 이 자료를 다음 순서대로 가공하여 '개벽 주요논설 코퍼스'(이하에서는 '코퍼스'로 줄임)를 만들었다.

전산자료[13] → 문장 길이 균질화[14] → 표기 통일[15] → 형태소 분석[16] → 한글 동형이의어(同形異義語) 구분[17] → 불용어 제거[18]

논설이라는 분류를 더 중요시하여 뒷부분에 있더라도 '논설'을 선택하였다.

12) 다음 장소에서 웹스크래핑을 통해 수집하였다.「한국근현대잡지자료: 개벽」, 국사편찬위원회 한국사데이터베이스, 2020.05.23. http://db.history.go.kr/item/level.do?sort=levelId&dir=ASC&start=1&limit=20&page=1&pre_page=1&setId=-1&totalCount=0&prevPage=0&prevLimit=&itemId=ma&types=&synonym=off&chinessChar=on&brokerPagingInfo=&levelId=ma_013&position=-1.

13) 334개 논설, 37,189개 문장이다.

14) 문장의 길이에서 너무 큰 편차가 나지 않도록 조정하였다. 문맥을 크게 해치지 않는 범위에서, 너무 길다고 판단한 문장은 180자 기준(공백 제외)으로 나누었다. 그 결과 37,189개 문장은 38,812개로 늘어났다.

15) 번역어, 국명, 순서 전도된 단어 등에 대하여 시행하였다. 특히 순서 전도된 단어의 경우, 현대어와 같은 뜻을 가지면서 어순만 역전된 2음절 단어는 현대어 용법의 표기로 통일하였다. 예컨대 '호상'(互相), '상호'(相互)는 '상호'로, '치사'(侈奢)와 '사치'(奢侈)는 '사치'로 통일하였다.

16) 형태소 분석은 파이썬에서 한국어형태소 분석기 '코모란 3.0'을 사용하여 진행하였다. 품사는 명사와 고유명사에 한정하였다. '○○하다'의 표현에서 '○○'에 한자어가 있으면 그 부분을 포함하였다.

17) 예를 들어 '유지'(有志, 維持, 遺志 등)처럼 동일한 표기가 둘 이상의 변별적 의미를 가지는 한글 단어의 경우, 실제 용례에 맞는 한자로 표기하여 단어만 봐도 서로 구분이 되도록 하였다.

18) 불용어는 연구자가 분석에서 불필요하다고 생각한 단어에 해당한다. 본 논문에서는 시간・수량 표시 단어의 경우 그 자체로는 의미 파악이 불분명한 경우가 많아서 불용어에 포함시켰다. '現在', '現下', '古今', '古來' 등이 여기에 해당한다. 동일한 이유로 一大, 一, 一個 등도 포함하였다. 단 '현대', '근대', '고대', '중세' 등 역사적 시기구분 용어는 불용어에서 제외하였다. 한편, 上・的・下・中・內・外 등도 불용어로 처리하였다. 다만 '전적'(全的)처럼 的 앞에 1음절 글자가 오면 이때의 的은 불용어에서 제외하였다. 1음절로 된 한글 및 한자도 동일한 취지에서 불용어로 간주하였다. 품사를 제한하고 불용어를 제거한 결과, 유효 문장 수는 38,812개에서 34,030개로 줄었으며 단어는 5,782종 271,532개를

→ 복합어 분리[19] → ‘문맥’ 범위 설정[20] → 코퍼스 산출

이상의 과정을 거쳐 3,426개 문서와 그 속에 있는 6,707종 158,394개 단어를 코퍼스로 확정하였다.[21] 머리말에서 언급하였듯이 본 논문의

확보하였다.

19) 복합어를 하위 성분으로 구분하는 것은 국립국어원의 표준국어대사전 표제어 형식에 따랐다. 하이픈(‘-’)로 구분된 단위가 그것이다. 예컨대 ‘기독교회’는 ‘기독/교/회’처럼 세 마디의 단위로 분해하고, 단위를 각각 별개의 독립적 단어로 간주하여 빈도를 계산하였다. 이와 별도로, 바이그램(2-gram), 트라이 그램(3-gram)을 적용하여 기독, 기독교, 기독교회도 각각 검색 대상에 포함시켰다. 이와 비슷한 경우를 더 제시하면 다음과 같다. ‘경제학자’ → 경제/학/자(경제, 경제학, 경제학자), ‘경제지식’ → 경제/지식(경제, 지식, 경제지식). 단 복합어를 하위 단위로 분해한 뒤 그 단위를 표기할 때 두음법칙 적용 여부가 애매할 경우에는, 해당 단위를 앞 단어 쪽에 붙여서 표기하는 것을 원칙으로 하였다. 즉 ‘변증론법’의 경우에는 ‘변증/론/법’으로 구분할 수 있으며, 복합어를 표기할 땐 ‘변증’, ‘변증론’, ‘변증론법’으로 하며, 맨 마지막 단어의 경우 ‘변증논법’으로 하지 않았다. 이러한 n-gram 사용으로 단어의 종류와 저빈도 단어 수도 증가하였다. 이에 빈도가 10회 미만인 단어는 일괄 제외하였다. 그 결과 단어 종류는 크게 증가한 대신 단어 개수는 오히려 이전보다 약간 감소하였다. 34,030개 문장에 6,707종 255,178개 단어가 되었다.

20) ‘문맥’은 단어 간의 공기(共起) 여부를 살필 때 기준으로 삼는 텍스트 공간이라는 점에서 중요하다. 문맥은 문장이 될 수도 있고 문단이나 논설 기사 전체가 될 수도 있다. 그런데 문맥을 논설 기사 전체로 삼을 경우, 그 범위가 너무 넓어서 공기어가 너무 많아진다. 반대로 문장으로 할 경우, 토픽 모델링으로 토픽을 추출하기에는 공기 범위가 너무 짧은 감이 있다. 따라서 본 논문에서는 10개 문장을 하나로 묶어서 이를 ‘문서’라 부르고, 이 문서를 ‘문맥’의 범위로 설정하였다. 문맥을 문서로 할 경우 논설 전체를 문맥으로 삼을 때보다 텍스트 길이가 상대적으로 균질해지는 이점도 있다. 문장을 통합하여 문서를 만들 때, 34,030개 문장을 앞에서부터 10개씩 끊어서 만들었다. 다만 두 논설 기사의 경계가 하나의 문서 속에 공존하는 일은 없도록 하였다. 의미 맥락상의 단절이 클 수 있기 때문이다. 이를 위하여 논설 기사의 범위 내에서 문서를 생성하다가, 자투리 문장이 5개 이하면 앞 ‘문서’에 통합하였다. 6개 이상이면 별도의 ‘문서’를 만들었다.

21) 단어 수가 이전의 255,178개에서 158,394개로 감소한 것은, 단어 집계 범위의 변경에 따른 결과이다. 이전까지는 잠정적으로 동일 단어의 중복 불허 범위를 문장으로 상정하였으나, 전술하였듯이 현 단계에서 문맥을 10개 문장을 합한

목적은『개벽』논조의 사회주의화가 개벽 주도층에 미친 영향을 검토하는 것이다. 그 출발점은 '사회주의화'를 판별할 지표를 정하는 것이다. 방금 획득한 코퍼스도 이 작업을 위한 재료이다. 토픽 모델링을 활용하면 코퍼스로부터 상호 응집력이 높은 n개의 단어 집합, 즉 토픽들을 얻을 수 있다. 선행연구의 성과나『개벽』논설의 내용상, 그 속에는 사회주의 주제도 들어 있을 가능성이 크다. 그렇다면 우리는 그 사회주의 주제에 속한 단어들을, 사회주의를 판별하는 내재적 지표로 삼을 수 있을 것이다. 게다가 사회주의를 더 넓은 범위, 즉『개벽』의 전체 주제 속에서 살펴볼 수 있는 단서를 얻게 될 것이다. 토픽 추출 과정은 다음 단계를 밟았다.

코퍼스 단어 수치화[22] → 토픽 모델링의 모델 선택[23] →

문서로 확정하면서 중복 불허 범위가 10배로 늘어난 셈이 되었다. 이에 따라 전체 텍스트 규모에서 본다면, 단어 종류는 그대로이나 단어 개수는 크게 줄어드는 결과가 되었다.

22) 문서의 특징을 추출해서 컴퓨터로 처리하려면, 먼저 그 문서를 컴퓨터가 식별할 수 있는 모습으로 바꿔야 한다. 가장 일반적인 방법은 문서를 수치화하는 것이다. 가장 간단한 수치화는 문서에 출현한 단어의 빈도를 활용하는 것이다. 이를 '백오브워즈'(bag of words) 방식이라고 한다. 그런데 이 방식을 사용한 문서 특징 추출은, 개별 문서에 집중적으로 출현하여 그 문서의 특징을 잘 나타내는 단어와, 문서들 전체에 공통적으로 많이 출현하는 단어를 잘 구분해 내지 못하는 한계가 있다. 이런 한계를 극복하기 위한 대안으로 나온 것이 '단어빈도-역문서빈도'(TF-IDF, Term Frequency-Inverse Document Frequency) 방법이다. 이 방법도 동일한 빈도 기반 수치화이지만 문서의 특징을 나타내는 데 더 효과적이기 때문에 본 논문에서는 이 방법을 채택하였다. 이 방법은 우선, 문서에서 많이 등장한 단어(TF)를 중요하게 생각한다. 이 점에서는 백오브워즈와 같다. 그러나 가장 큰 차이점은 그 단어가 여러 문서에 걸쳐서 등장할수록 해당 단어의 빈도를 일정하게(IDF) 낮추는 데 있다. 본 논문에서는 토픽 모델링 진행 과정에서 '사회', '조선' 같은 고빈도 단어가 다수의 토픽에 공동으로 등장하는 일이 잦아서, 이를 방지하기 위하여 TF-IDF 방식의 수치화 방식을 채택하였다. TF-IDF에 관한 사항은 다음을 참조하였다. 이기창(2020),『한국어 임베딩 — 자연어 처리 모델의 성능을 높이는 핵심 비결, Word2Vec에서 ELMo, BERT까지』, 서울:

세부 설정값 조정[24] → 토픽 추출

필자는 추출 과정에서 세부 설정값을 다양한 조합으로 반복 입력하였다. 그리고 그 중에서 『개벽』의 전체 내용을 잘 분류하고 본 논문의 연구 목적에 부합할 것 같은 결과를 선택하였다. 그렇다고 하여 이 작업이 연구자가 기계학습이나 통계라는 이름 아래 실제로는 주관적 의도를 관철한다고 볼 순 없다. 또한 이와 반대로 이 과정이 절대적 타당성을 가졌다거나, 알고리즘 때문에 연구자의 선택 여지가 거의 없다고 속단할 수도 없다. 오히려 그것은 『개벽』에 관한 필자의 사전 지식 및 연구 목적과, 통계적 추론 알고리즘에 따른 기계학습이 상호 협업한 결과에 더 가깝다.[25]

에이콘, pp. 60-64.

23) 토픽 모델링은 "문서 모음에서 토픽을 추출하기 위한 비지도 방식 머신러닝 기술"이다[벤자민 벵포트・레베카 빌브로・토니 오제다(2019), 박진수 역, 『파이썬으로 배우는 응용 텍스트 분석』, 파주: 제이펍, p. 122]. 그 대표적 모델인 '잠재 디리클레 할당(LDA, Latent Dirichlet Allocation)'을 사용하면 토픽을 "주어진 용어 집합이 발생할 확률로 표현"할 수 있고, 이에 따라 문서를 "이러한 토픽들의 혼합체(mixture)로 표현"할 수 있다[벤자민 벵포트・레베카 빌브로・토니 오제다(2019), p. 122].

24) 토픽 종류는 8개, 토픽 1개당 단어 수는 20개로 지정하였다. 연산 반복 회수는 1,000회로 하였다. 그 외의 값은 다음과 같이 주었다. doc_topic_prior = 0.1, topic_word_prior = 0.0005, random_state = 59. 맨 처음 것은 통상 알파(α)값이라 부르는데, 문서들의 토픽 분포에 영향을 미친다. 두 번째는 베타(β)값이라 하는데, 각 토픽에서 단어가 분포하는 양상에 영향을 미친다. 알파값과 베타값이 클수록 토픽 간의 분포는 균등해지며, 작을수록 특정 토픽이 크게 나타난다. 맨 마지막 값은 토픽 모델링 결과에 영향을 주는 변수는 아니다. 다만 연구자가 특정 조합의 입력값을 최종적으로 선택해서 그 결과를 연구에 활용하였을 경우, 다른 연구자가 동일한 자료로 동일한 토픽 결과를 산출하려할 때 필요한 값이다.

25) 토픽 모델링이 결과의 가변성을 가지지만 과학적 탐구방법의 필수 요소인 반복성과 예측가능성도 확보한다는 점에 대해서는 다음 기사를 참조할 필요가 있다. 권오성(2019),「변덕꾸러기 토픽 모델링 어떻게 다뤄야 하나」, 한겨레 웹페이지, 한겨레신문, 2020.09.30. http://www.hani.co.kr/arti/science/technology/907134.html#

이러한 경로로 토픽을 추출하여 우선, 8개의 토픽과 토픽별 20개씩 총 160개의 토픽 단어를 얻게 되었다.[26] 그 면면은 <표 1> 내용 중 음영 및 대각선 처리 이전 상태에 해당한다.

〈표 1〉 LDA로 추출한 8개의 토픽과 토픽 보정 결과

토픽 순위	토픽1 (20)	토픽2 (2)	토픽3 (11)	토픽4 (19)	토픽5 (6)	토픽6 (8)	토픽7 (20)	토픽8 (20)
1	자본	個性	당쟁	소작	동학	朱子	일본	사람
2	계급	규범	도착	지주	전봉준	천지	朝鮮	사회
3	노동	김정필	이조	농촌	대회	태극	경제	생활
4	생산	효도	태조	농업	인터내셔널	만물	미국	朝鮮
5	무산	旣成	정권	토지	정치	우주	영국	主義
6	운동	부모	高麗	농민	갑오	인류_구제	전쟁	사상
7	主義	생활_표준	한양	소작_人	관군	道心	프랑스	민족
8	부르주아	로맹롤랑	도착_點	朝鮮	임시	天理	독일	自己
9	자본_主義	生道	정국	노동	정치_문화	한울	평화	인류
10	경제	詩村	宣祖	지방	사회_실상	佛陀	유럽	운동
11	마르크스	부르주아_문화	출발	도회	朝鮮_운동	하나_님	세계	정신
12	사회	죽음	사화	경작	식민_정책	獨夫	朝鮮_人	자유
13	무산_계급	쾌락	남북	서당	물질_조건	일월	국제	민중
14	제도	문화_발전	즉위	교육	公州	知者	인구	문화

csidx45abb92d7e27dd4b65c5203ceef0fce.

26) 본 논문에서는 파이썬 머신러닝 라이브러리 중에서도 사이킷런(Scikit-learn)을 사용하였다.

토픽 / 순위	토픽1 (20)	토픽2 (2)	토픽3 (11)	토픽4 (19)	토픽5 (6)	토픽6 (8)	토픽7 (20)	토픽8 (20)
15	자본_家	생활_양식	임진	개량	농민_朝鮮	불교	會議	개인
16	착취	작자	당파	경성	몽룡	가절	외국	세계
17	노동_者	민주	세력	학교	근세	음양	대전	자연
18	공업	독서	형세	商工業	강령	인내천	중국	인간
19	혁명	新_문학	불평	금융	격언	神人	帝國	정치
20	노동_계급	측정	이성계	收入	식민	한자	국가	의미

비고: 1. 밑줄 표시('_')가 들어 있는 단어는 n-gram으로 포함시킨 복합어이다.
2. 표의 전체 단어는 토픽별로 20개씩 추출한 것이다. 이 중 음영에 대각선으로 표시한 곳은 토픽 보정으로 제외한 경우이다.
3. 토픽번호 아래의 괄호 속 숫자는 토픽 보정을 거쳐 최종적으로 남은 토픽단어 수이다. 토픽 2번은 2개 단어만 남아서 토픽 자체를 제외하였다.
4. 왼쪽 '순위'의 숫자는, 각 단어가 토픽에서 차지하는 비중이 가장 높은 것을 1번으로 하여 내림차순으로 정렬한 것이다.

2.2. 토픽 보정과 명명(命名)

그런데 토픽 모델링에 대해서는 논자들이 다음과 같은 사용상의 주의점을 지적한 바 있다. 원래 하나로 묶을 수 있는 토픽이 여러 개의 토픽에 걸쳐 있는 '토픽 중복' 현상이나, 이와 반대로 하나로 추출된 토픽 안에 여러 개의 상이한 토픽이 병존하는 '토픽 혼재' 현상이 종종 발생한다는 것이다.[27] 이런 한계를 극복하기 위하여 토픽 모델링으로 추출한 토픽을 수정・보완하는 '토픽 보정'을 진행하였다.[28] 토픽 보정 작업은 토픽 분리와 토픽 병합의 두 단계로 구분할 수 있다. 하나의 토픽을 단위로 하여 진행 과정을 요약하면 다음과 같다.

27) 김동욱・이수원(2017),「단어 유사도를 이용한 뉴스토픽 추출」,『정보과학회 논문지』44(11), p. 1139.

28) 토픽 보정의 진행은 다음 연구에 크게 의존하였다. 김동욱・이수원(2017), pp. 1141-1144.

① 토픽 분리[29]: 토픽 단어를 순위별로 정렬 → 1순위 단어(편의상 '중심 단어')는 고정시켜 놓고 이것을 나머지 단어(편의상 '대응 단어')와 일대일로 대응시켜 PMI를 산출[30] → PMI가 0과 음수일 경우, 해당 대응 단어는 제거 → 생존 단어 중 그 다음 순위의 단어를 중심 단어로 설정 → 이 단어를 그 이하 순위의 단어와 일대일로 대응시켜 PMI를 산출한 뒤, 그 값이 0과 음수일 경우 해당 대응 단어는 제거 → 이런 식으로 마지막 단어까지 진행한 뒤 생존 단어들로 해당 토픽의 '1순위 TC'를 생성 → 동일한 과정을 2순위 단어에게도 적용하고, 이런 식으로 '2순위 TC'부터 '20순위 TC'까지 생성 → 생성된 20개의 TC 중에서 단어 집합이 동일한 것(순서는 고려하지 않음)은 제거

② 토픽 병합[31]: 동일 토픽 내 TC를 두 개씩 조합하여 새 TC들을 생성 → 새 TC별로 '토픽 분리'와 비슷한 방식으로 단어 간 PMI를 산출 → PMI가 0과 음수에 해당하는 경우가 전체 PMI 수의 0.3 이하인지 여부를 확인[32] → 0.3 이하이면 해당 새 TC는 토픽 병합 후보로 남기고, 0.3 이상이면 제외 → 최종적으로 남은 토픽 병합 후보 중에서 소속 토픽 단어들의 PMI 값 평균을 산출 → 평균이 가장 높은 TC를 최종 토픽으로 확정

---

29) '토픽 분리'란 각 토픽을 단위로 하여, 응집력이 높은 단어들의 결합인 TC (Topic Clique)를 생성하는 과정이다. 연결망 분석에서 clique를 '파당'(派黨)으로 부르기도 한다. 그런데 TC에 대응하는 번역어는 찾지 못했다. TC는 토픽별로 한 개만 만드는 것이 아니라, 토픽 단어 각각을 중심으로 하는 것이므로 본 논문에서는 토픽별로 20개씩 만들었다.

30) PMI는 '점별 상호 정보량'(PMI, Pointwise Mutual Information)이다. PMI는 '두 단어가 하나의 문서에서 동시에 출현할 확률'을, '두 단어가 각각 출현할 확률을 서로 곱한 것'으로 나눈 값이다. 두 단어의 연관성이 높을수록 PMI 값이 높다.

31) '토픽 병합'은 각 토픽별로 토픽 분리로 생성된 TC 중에 유사성이 높은 것을 통합하는 작업이다.

32) 이 0.3은 임계값(Threshold)이다. 다음 연구에서 그것이 토픽 보정의 임계값임을 증명하였다. 본 논문의 작업도 기본적으로 동일한 경우로 판단했으므로 이 임계값을 그대로 사용하였다. 김동욱・이수원(2017), p. 1146.

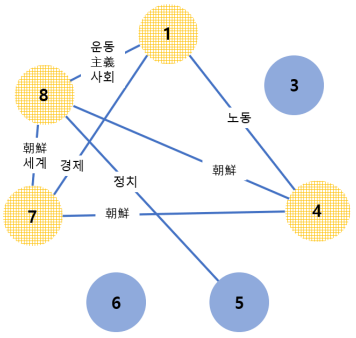
토픽 보정으로 제거한 단어와 주제는 <표 1>에서 대각선 및 음영으로 표시하였다. 그리하여 최종적으로 7개 토픽에 95종 104개 토픽 단어를 획득하였다. 토픽1, 토픽4, 토픽7, 토픽8은 토픽 단어가 19~20개인 대규모 토픽이다. 토픽3, 토픽5, 토픽6은 토픽 단어가 각각 11개, 6개, 8개로 중소규모 토픽이다.[33)]

한편 토픽 모델링은 당초에 연구자가 지정한 주제 수 및 토픽 단어 개수만큼만 상호 응집도 높은 단어를 산출할 뿐이며 그 토픽이 어떤 성격을 가졌는가는 표시하지 않는다. 따라서 필자가 그 결과물을 보고 적절한 이름을 붙였다. 그 결과는 [그림 1]의 좌측과 같다.

그런데 95종 104개 토픽 단어 중 일부는 적어도 두 군데 이상의 토픽에 속해 있다. 그 용례의 밀도나 의미 폭이, 다른 토픽 단어보다 높고 넓기 때문일 것이다. 편의상 이를 '복속형'(複屬形)으로 분류한다. 이와 달리 소속이 하나 뿐인 토픽 단어는 '단속형'(單屬形)이라 부른다. 단속형 토픽 단어는 87종 87개이다. 복속형 토픽 단어는 8종 17개이다. [그림 1]의 우측은 복속형 토픽 단어와 토픽 간의 관계를 표시한 것이다.

33) 독자들 중에는 '개벽 주도층은 천도교청년층인데 천도교 관련 토픽이 없는 것이 이상하다'라는 의문을 가질 수 있다. 이재연도 '생명', '개벽사' 등 천도교와 관계가 있는 토픽을 제시하였다[이재연(2016), p. 298]. 필자가 보기에 가장 직접적인 이유는 이재연이 『개벽』 기사 전체를 대상으로 토픽을 50개 추출한 데 비해, 본 논문에서는 검토 대상을 『개벽』의 주요 논설에 한정하였고, 토픽 수도 7~8개로 설정했기 때문이다. 더 근본적으로는, 『개벽』의 편집과 운영을 천도교청년층이 주도하였지만 『개벽』의 핵심 논조에는 종교성이 크게 드러나지 않았기 때문이다. 이에 관하여는 다음 연구들을 참조할 필요가 있다[허수(2015), 「『개벽』의 종교적 사회운동론과 일본의 '종교철학'」, 『인문논총』 72 (1), 서울대학교 인문학연구원, pp. 339-341; 허수(2011), p. 103; 최수일(2008), pp. 371-376]. 이런 점을 고려하면 <표 1>의 토픽6에서 당초 '인내천'과 '神人' 등 천도교의 종교사상에 관한 용어가 들어 있었으나 토픽보정 과정에서 제거된 이유도 유추할 수 있다. 토픽6 내에서도 '朱子' 등에 비해 주요 논설에서 천도교 관련 단어의 영향력이 상대적으로 약했기 때문이다. 한편 본 논문에서는 분석대상을 『개벽』으로 설정했으므로, 잡지 '개벽'과 출판사 '개벽사' 등은 그다지 의미 있는 용례가 아니라고 보아 제외하였다.

토픽id / 토픽명 순위	tp1 사회주의	tp3 조선정치	tp4 경제,교육	tp5 동학농민 전쟁	tp6 성리학	tp7 국제관계	tp8 개조론
1	자본	당쟁	소작	동학	朱子	일본	사람
2	계급	이조	지주	전봉준	전지	朝鮮	사회
3	노동	태조	농촌	정치	태극	경제	생활
4	생산	高麗	농업	갑오	만물	미국	朝鮮
5	무산	한양	토지	관군	우주	영국	主義
6	운동	宣祖	농민	公州	道心	전쟁	사상
7	主義	사화	소작_人		天理	프랑스	민족
8	부르주아	즉위	朝鮮		음양	독일	自己
9	자본_主義	임진	노동			평화	인류
10	경제	세력	지방			유럽	운동
11	마르크스	형세	도회			세계	정신
12	사회		서당			朝鮮_人	자유
13	무산_계급		교육			국제	민중
14	제도		개량			인구	문화
15	자본_家		경성			會議	개인
16	착취		학교			외국	세계
17	노동_者		商工業			대전	자연
18	공업		금융			중국	인간
19	혁명		收入			帝國	정치
20	노동_계급					국가	의미



비고: 1. tp1은 '토픽1'을 뜻한다.
2. 좌측 그림의 음영 부분은 '복속형' 토픽 단어이며, 이 단어와 토픽 간의 관계는 우측 그림과 같다.
3. 우측 그림에서 옅은(노랑) 원은 대규모 토픽을, 진한(파랑) 원은 중소규모 토픽을 나타낸다.

[그림 1] 토픽의 명명 및 복속형 토픽단어 현황

지금까지 살펴본 토픽 모델링 결과에 따르면 문서를 구성하는 단어들은 크게 세 부류로 나눌 수 있다. '단속형 토픽 단어', '복속형 토픽 단어', '일반 단어'가 그것이다.[34] 이 가운데 앞의 두 부류, 즉 토픽 단어는 해당 문서의 의미나 성격을 파악할 수 있는 유용한 특징이 될 수 있다. 문서 내 토픽 단어 중에서 토픽1에 속한 단어가 많다면, 그 문서는 사회주의적 성향을 띤다고 말할 수 있는 것이다.

3. 개조론 중심에서 사회주의 중심으로

2장에서는 문서를 단위로 하여, 사회주의 성향을 판별할 토픽 단어를 토픽 모델링으로 도출하였다. 머리말의 문제의식을 끌어와서 말하

34) 본 논문에서는 토픽 단어 이외의 단어를 편의상 '일반 단어'라 부른다.

자면 사회주의의 영향을 판단할 수 있는 지표를 확보한 것이다. 이 장에서는 그것을 밑천으로『개벽』논조의 사회주의화 문제를 살펴보고자 한다. 머리말에서는『개벽』의 논조를 주제들의 합성, 즉 어우러짐의 효과라 가정하였다. 그렇다면 논조의 구조를 파악하는 일은 7개 토픽 간의 관계를 형상화하는 데 있을 것이다. 이는 얼핏 생각해도 토픽 7개를 점(node)으로 하는 토픽 연결망 지도를 그려보면 될 것이다. 그런데 이 지도 작성은 점과 함께 선(link), 즉 토픽 간의 관계를 숫자로 표현할 수 있을 때 가능하다. 토픽 간 관계를 무엇으로 포착하며 어떻게 수치화할 수 있을까?

이에 관한 본 논문의 아이디어는 토픽 단어를 토픽 간 유사도를 파악하는 매개물로 사용하는 것이다. 이 생각을 풀어 쓰면 다음과 같다. 첫째,『개벽』의 논조는 334개의 주요 논설로 접근할 수 있다. 또한 토픽 모델링의 결과 주요 논설의 골간은 7개의 토픽으로 표현할 수 있게 되었다. 그리고 '문서'의 특징은 토픽 단어들로 표현할 수 있다. 둘째, 문서를 구성하는 토픽 단어들은 서로, 그 문서에서 공기(共起) 즉 동시 출현하는 관계로 볼 수 있다. 따라서 '문서 — 토픽 단어 — TFIDF 값'을 이용하면 '토픽 단어 — 토픽 단어' 간의 연결망 지도를 그릴 수 있다.[35] 셋째, 토픽 단어 연결망 지도에서 토픽 단어를 소속 토픽으로 치환하면, 앞서 목표로 삼았던 토픽 연결망 지도를 도출할 수 있다.[36]

35) 본 논문에서는 TF-IDF를 '값'과 붙여서 사용할 경우 편의상 하이픈('-')을 생략하고 'TFIDF값'으로 표기한다.

36) 어떤 독자는 이런 과정이 불필요하거나 너무 번거롭다고 생각할 수도 있다. 3장 2절에서 상술하겠지만, 본 논문의 방식은 1) '문서1 — 단어1 — TFIDF값', '문서1 — 단어2 — TFIDF값' 등을 입력하고, 2) 이로부터 '단어1 — 단어2'의 공기어 연결망을 도출하며, 3) 이런 공기어 연결망을 이루는 각 단어들을, 각각 그 단어가 소속된 토픽 번호로 치환하여 토픽 연결망을 도출한다. 그런데 이처럼 에둘러 가는듯한 방법 대신, 1) '문서1 — 토픽1 — 토픽1의 문서 내 비중', '문서1 — 토픽2 — 토픽2의 문서 내 비중' 등을 입력하고, 2) 이로부터 '토픽1 — 토픽2'의 토픽 연결망을 곧바로 산출하는 것이 더 간단명료하지 않은가라는 의구심을

3.1. 토픽 단어의 소속 판별과 시기 구분

전술하였듯이 95종 104개 토픽 단어 중 8종 17개는 두 개 이상의 토픽에 공통으로 속한 '복속형'이다. 따라서 각 문서에 출현한 복속형 토픽 단어의 경우 일률적으로 그 소속을 정할 수 없다. 이 유형의 단어는 8종에 불과하다. 그러나 그 소속을 마저 판별하지 않으면 문서의 특징 파악을 완료할 수 없다. 또한 8종에 불과해도 '조선'(朝鮮), '사회' 등 대부분 고빈도 단어라『개벽』에서 그 영향력이 크다.[37] 그럼 어떻게 해야 합리적으로 그 소속을 판별할 수 있을까.

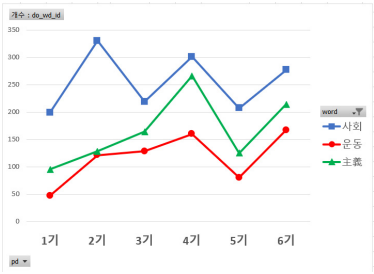
본 논문에서는 복속형 주변에 있는 단속형들의 소속을 단서로 삼았다. 문서에서 '조선'은 토픽4와 토픽7, 토픽8의 세 가지 가능성을 가진다. 1번 문서에서 '조선'이 주로 토픽4에 속하는 단어들과 함께 나타났다고 가정하자. 그렇다면 이 문서에서 '조선'의 의미는 토픽4에 해당

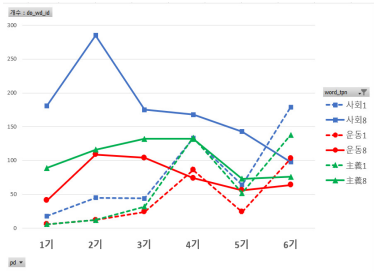
가질 수 있다. 그러나 후자의 방법대로 하면 제대로 된 토픽 연결망을 얻을 수 없다. 왜냐하면 이럴 경우 텍스트 전체에서 각 문서의 특징을 나타내는 인자가 최대 7개(= 토픽 수)로 지나치게 단순화되어 문서의 변별성이 극도로 줄어들기 때문이다. 그 결과 이 문서를 매개로 도출한 토픽과 토픽 간의 1항 관계도 7개 토픽 대부분이 서로 연결된 형태가 되어 버린다. 이렇게 하는 대신 토픽 '단어'의 형태를 유지하면 텍스트 전체에서 각 문서의 특징을 나타내는 인자는 최대 104개(= 토픽단어 수)가 된다. 그러므로 문서의 변별성을 유지하려면 본 논문에서 실행한 것처럼, '단어' 형태를 고수하되, 마지막 단계에 와서야 그 단어를 소속 토픽으로 치환하는 것이 필요하다.

37) 토픽 판별의 필요성은 다음 사례에서도 확인할 수 있다. '사회', '운동', '주의'(主義)는 토픽1과 토픽8 간의 복속형 토픽단어이다. 이 단어들의 단순 빈도 변화를 토픽 구별 없이 살펴보면 아래의 좌측 그래프와 같다. 1~6기 구분은 각각『개벽』12개 호(1년)를 단위로 하였다. 이 그래프만으로는 시기별 변화에서 특별한 모습을 관찰하기 어렵다. 그런데 세 토픽 단어의 소속을 반영한 것이 우측 그래프이다. 토픽1은 점선으로 표시하였다. 여기서는 토픽8 소속 단어는 3기부터 감소한다. 반면 토픽1 소속 단어는 3기 이후 급변하고 전체적으로 증가 추세를 보인다. 이런 차이는『개벽』의 논지 파악에서 토픽 단어의 소속 판별이 중요한 또 하나의 이유라 할 수 있다.

할 개연성이 높다는 발상이다. 다만 이 발상을 구체화할 때에는 한 가지를 더 고려하였다.[38] 선택 가능한 토픽에서 해당 단어가 차지하는 중요도를 반영하는 것이다. 실제 전개과정은 다음의 순서를 밟았다.[39]

> 토픽 단어를 소속별로 집계[40] → 복속형 토픽 단어의 토픽별 중요도 계산[41] → 소속 토픽 판정[42]





38) 토픽 단어의 소속을 판별하는 방법은 토픽 모델링에서 토픽을 산출하는 원리와 동일하다. 그러나 이런 발상은 곽기영의 다음 자료에서 영향을 받아서 응용한 결과이다. 곽기영(2020),「데이터 애널리틱스. 텍스트마이닝7: 토픽 모델링」, 유튜브, 2020.09.15. https://www.youtube.com/watch?v=CuW7-QkNMNE. 이 중에서 필자가 참고한 곳은 앞부분의 설명(처음~22:10)이다.

39) 본 논문에서는 토픽 단어 이외의 단어, 즉 일반 단어는 문서의 특징을 나타내는 유의미한 요소로 상정하지 않았다. 이에 토픽 단어의 소속 판별에서 일반 단어의 개수는 무시하였다.

40) 24번 문서에 다음과 같이 8개의 토픽 단어가 있다.
농민(4), 농촌(4), 개량(4), 생활(8), 농업(4), 자본(1), ***朝鮮(4/7/8)***, 자본_家(1)
괄호 속 숫자는 소속 토픽 번호이다. 토픽4, 토픽7, 토픽8에 소속된 단어를 찾는다. 토픽4에 소속된 단어는 '농민', '농촌', '개량', '농업'이므로 모두 4개이다. 토픽7 소속 단어는 없으므로 판별 후보에서 제외한다. 토픽8 소속은 '생활' 1개이다.

41) 토픽4와 토픽8에서 '조선'이 차지하는 중요도를 계산한다. 중요도는 코퍼스 전체에서 '조선'이 출현한 빈도를, 동일 범위에서 '조선'이 속한 토픽의 모든 단어 빈도를 합산한 값으로 나누어 구할 수 있다. 24번 문서에서는 토픽4와 토픽8에서 '조선'이 가진 중요도를 계산할 필요가 있다. 토픽4에서 이 단어의 중요도는 0.528, 토픽8에서는 0.147로 나왔다.

42) 토픽별로 소속 단어 개수와 중요도를 서로 곱해서 높은 값이 나온 쪽으로 '조선'

토픽 단어의 소속을 모두 확정하였으므로 이제 토픽 단어로 문서의 특징을 온전히 나타낼 수 있게 되었다. 본 논문의 논지 전개에서 토픽 단어의 사용이 가장 시급한 부분은 시기 구분이다. 지금까지『개벽』의 논조가 전・후기에 차이가 있다고 한 것은 선행 연구의 성과를 일단 따른 것이다. 그러나 이 구분은 아직 잠정적인 것이다. 변화를 관찰하려면 몇 개의 시기로 구분하면 좋을지, 어디가 적절할지 등을 검토할 필요가 있다.

시기 구분은 논조 변화를 살펴볼 시기적 단위를 정하는 일이라 매우 중요하다. 본 논문에서는 두 가지 대상에 주목하였다. 문서의 토픽 구성과 문서 유사도이다.

먼저, 문서의 토픽 구성이 크게 변하는 지점을 살폈다. '문서의 토픽 구성'은 토픽의 상대적 비중을 나타낸 것이다. 문서의 토픽 단어에서 표기 형태는 무시하고 그 소속만 토픽별로 집계하였다. 각 문서에서 토픽의 비중을 모두 합하면 1이 된다.[43] [그림 2]의 위쪽 그래프는 토픽1과 토픽8만 표시했는데, 토픽1의 비중이 후반부에 높아지는 것을 알 수 있다. 비슷한 양상을 아래쪽 그래프에서도 관찰할 수 있다. 이것은 7개 토픽 전체 구성을 백분율로 표시한 것이다. 다른 토픽에 비해 토픽1의 증가세가 두드러진다. 토픽8은 대체로 이와 반대되는 양상을 보인다. 토픽1의 경우 특히 1~12분기와 13~24분기의 차이가 육안으로도 두드러진다.

의 소속 토픽을 정한다. 토픽4의 경우 4 × 0.528 = 2.112이고, 토픽8의 경우 1 × 0.147 = 0.147이므로 이 문서에서 '조선'의 소속은 토픽4로 판별한다. 말하자면 이 문서에서 '조선'은 토픽4, 즉 '경제・교육'의 맥락에 있다고 볼 수 있는 것이다. 만약 24번 문서에서 판별의 근거를 찾을 수 없다면, 적용 범위를 23번 및 25번 문서로 확장해서 해결한다.

43) 다시 24번 문서로 예를 들자면, 이 문서를 구성하는 8개의 토픽 단어는 토픽1 소속이 2개, 토픽4 소속이 5개, 토픽8 소속이 1개이다. 그러므로 토픽 구성은 각각 토픽1이 0.25(2/8), 토픽4가 0.625(5/8), 토픽8이 0.125(1/8)이다.

비고: 1. 위쪽 산점도(Scatter Plot)의 가로축은 문서 번호로 시간 순서대로 되어 있다. 세로축은 문서의 토픽 구성이다. 연한 색(하늘색) 점은 토픽8의 분포를, 진한색(빨강색) 점은 토픽1의 분포를 나타낸다.

2. 아래쪽 '100% 기준 누적 세로 막대형 그래프'의 가로축은 3개월을 분기로 하여 구분한 24개 분기이다. 세로축은 각 문서의 토픽 구성을 분기별로 합산한 뒤 그것을 백분율로 표시하였다. 범례의 색깔은 맨 아래부터 위쪽으로 토픽1에서 토픽8까지의 7개 토픽을 배열하였다.

[그림 2] 토픽 구성의 시기별 변동

다음으로 문서와 토픽 단어의 관계를 이용하여 분기 간의 유사도를 살펴보았다. 본 논문에서는 이를 위해 첫째, '문서 — 토픽 단어 — TFIDF값'을 24개 분기 단위로 재구성하였다. 그 결과가 '분기 — 토픽 단어 — TFIDF값'이다.[44] 둘째, 이 값을 '넷마이너'에 입력하여 '문서

44) 문서에 출현한 토픽 단어의 빈도를 단순히 등장 회수만큼 집계하는 대신, '단어 빈도-역문서빈도'(TF-IDF)라는 상대 빈도로 나타내는 취지는 본 논문 2장 1절의

— 문서'로 된 네트워크를 작성하였다.[45] 셋째, 이 소프트웨어의 '커뮤니티 분석' 기능을 사용하여 문서들을 2개로 군집화(Clustering)하였다.[46] 이 커뮤니티 분석은 '계층적 군집화' 방식에 해당하며, 군집 개수

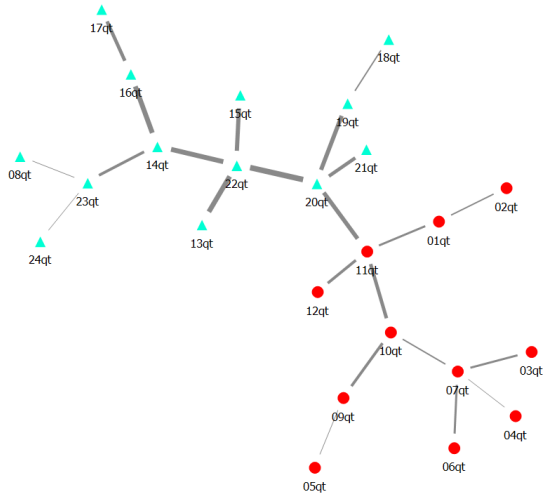
각주 22번에서 밝힌 내용과 일맥상통한다. 본 논문에서는 분석의 필요상 텍스트 혹은 문맥의 단위를 '문서'에서 '분기'나 '논설' 등으로 확장하더라도 토픽 단어를 TF-IDF라는 상대 빈도로 나타내는 원칙을 고수하였다. 한 가지 특기해 둘 사항은 당초 '문서' 기준으로 도출한 토픽 단어의 TF-IDF 값을 '분기'나 '논설' 규모에 맞게 단순하게 통합하는 계산 방식은 부적절하다는 점이다. 문맥의 범위가 달라짐에 따라 TF-IDF 계산에서 고려해야 할 단어 수와 빈도가 상이해지기 때문이다. 따라서 비교 단위가 24개 분기로 변했을 경우 기존의 상대 빈도는 무시하고 다시 코퍼스 단어로 돌아가서, 사이킷런의 라이브러리를 활용하여 분기를 단위로 하는 (토픽) 단어의 상대 빈도를 재산출하였다.

45) 사용한 소프트웨어의 정보는 다음과 같다. Cyram (2021), NetMiner v4.3.2.d, Seoul: Cyram Inc. 신서인은 네트워크 연구에서 연결망 분석을 주로 절대 빈도에 의존해 온 관행을 비판하였다. 본 논문에서 TF-IDF를 사용한 것도 이런 비판을 수용한 측면이 크다. 신서인(2017), 「네트워크 분석을 이용한 복지 담화 연구」, 『개념과 소통』 20, 한림대학교 한림과학원, pp. 220-221.

46) 연결망 지도는 '분기' 간 코사인 유사도에 기반을 두었고, 연결망 산출은 패스파인더(PFnet, PassFinder Net) 알고리즘을 사용하였다. 패스파인더 방식은 '전체적인 지적 구조를 잘 표현하면서도 개별 개체에 관한 세부 구조도 비교적 잘 나타내는 강점'이 있어서, 계량 서지적 분석에서 많이 이용해 오고 있다[이재윤(2006), 「지적 구조의 규명을 위한 네트워크 형성 방식에 관한 연구」, 『한국문헌정보학회지』 40(2), 한국문헌정보학회, pp. 351-352 참조]. 이재윤에 따르면 연결망 형성 방식은, '생성 혹은 제거할 링크를 선택하는 기준에 따라' 절대적 가중치 기준과 상대적 가중치 기준의 두 방법이 있다. 전자에 속하는 '기준값 절단'은 가장 간단한 방법으로, 특정 가중치를 기준으로 삼아 그 이상의 값을 가진 링크만 남긴다. 이 방식은 "기준값을 낮출수록 링크의 수가 많아져서 그래프가 복잡해진다." 또한 "기준값을 높이면 그래프가 분할되어 서브그래프나 고립 노드"가 나타나, "전체적인 연결 흐름이 나타나지 않는 경우가 있다."[이재윤(2006), p. 338]. 이재윤은 이런 문제점에 대한 대안으로 패스파인더 알고리즘을 제안한다. 패스파인더 방식은 '상대적 가중치 기준'에 속한다. 이 기준은 '절대적 가중치 기준'과 달리 "가중치가 더 높으면서도 상대적인 중요도에 따라서 생략되는 링크가 있을 수 있다."[이재윤(2006), p. 337]. 필자는 이재윤의 이러한 지적에 깊이 공감하였다. 본 논문의 목적에 비추어 필자는 일단 104개의 토픽 단어들이 연결망 지도에 누락 없이 남으면 좋겠다고 생각하였다. 또한 그것들이 전체적으로

2개는 이 알고리즘이 제안하는 최적의 개수를 따랐다. 2개 군집은 [그림 3]의 연결망에 색깔과 형태로 구분하였다.

[그림 3]의 결과도 [그림 2]의 분석 결과와 대동소이하다. 24개 분기를 12분기와 13분기 사이를 경계선으로 양분할 수 있기 때문이다. 연결망 지도의 좌측 상단에 8분기가 후반부 분기들과 붙어 있긴 하다. 그러나 대세에 지장을 줄 정도는 아니다.



[그림 3] 24개 분기를 2개로 군집화한 결과

모두 연결해 있으면서도, 전체 구조를 파악하기 어려울 정도로 복잡하지는 않기를 희망하였다. 따라서 패스파인더 방식이 적절하다고 판단하였다. 실제 사용은 넷마이너 메뉴의 'Analyze\Connection\PFnet'를 활용하였다. 먼저 '분기 — 토픽 단어 — TFIDF값'을 입력하여 '분기 — 토픽 단어' 간 2항(2-mode) 연결망을 만들었다. 유사도를 비교할 단위로 설정한 '분기'는 24개에 불과하므로 이 개수를 줄이진 않는다. 유사도는 코사인 유사도를 선택하였다. 다음은 2항 연결망을 24개 분기를 중심으로 하는 1항 연결망으로 전환하였다. 그 다음의 '선'(link) 절삭 단계에서 패스파인더 방식을 사용하였다. 이 방식을 사용하면 연구자가 특별히 절삭 기준값을 제시할 필요는 없다.

지금까지 문서의 토픽 구성을 기초로 시기별 변동과 분기 간 유사도를 살펴보았다.『개벽』36호와 37호 사이가 가장 뚜렷한 구분선이 된다는 점을 알 수 있었다. 이는『개벽』발행 기간의 한가운데에 해당하는 지점이기도 하다. 따라서『개벽』의 논조변화는 전기(1~36호)와 후기(37~72호)로 구분해서 고찰하는 것이 합리적이다.[47)]

3.2. 전・후기 연결망의 구조 변동

이 절에서는 토픽 연결망 지도를 그려서 전기와 후기의 논조를 살펴볼 것이다. 나아가 전・후기 사이의 구조 변동을 지속과 변화에 함께 유의하며 검토하려 한다. 3장 첫머리에서 언급하였듯이 먼저 '문서 — 토픽 단어 — TFIDF값'을 이용하여 '토픽 단어 — 토픽 단어' 간의 연결망 지도를 그릴 것이다. 다음으로 토픽 단어 연결망 지도에서 토픽

47) 익명의 심사위원 한 분이 '『개벽』이 36호를 기준으로 전기와 후기로 나뉘며, 후기의 경우 논조의 사회주의화가 진행되었다는 점은『개벽』을 일별하는 것만으로도 알 수 있는 너무 분명한 결론이다'라는 요지의 심사평을 해 주었다. 필자도 머리말 첫 단락에서 '1923~1924년 무렵부터 사회주의적 논설이 급증'하는 사실은 대부분의 연구자들이 동의하는 바라고 언급하였다. 그러나 필자가 '1923~1924년 무렵'이라고 표현한 것은, 선행연구들이 사회주의 논설의 급증 시점을 조금씩 달리보고 있는 점을 감안한 서술이다. 따라서 필자는 전・후기 시기구분에 관한 본 논문의 결론이, 그렇게 자명하거나 '일별'만으로도 알 수 있는 '너무 분명한 결론'이라 생각하지 않는다. 예컨대 김정인은 '좌경적 성향의 논설과 기사'가 1923년부터 비중 있게 나타난다고 하면서, 구체적으로는 1923년 2월의『개벽』32호 기사를 인용하였다[김정인(2007), p. 259]. 허수는『개벽』41호를 계급 담론 대두 시점으로 보면서 이때부터 사회주의 소개 기사가 본격적으로 등장한다고 보았다[허수(2011), p. 97, p. 109]. 최수일은 논조의 큰 변화가 감지되는 시점을 37호 전후한 시기로 보았다[최수일(2008), p. 462]. 적어도 본 논문의 방법에 따라 선행연구를 살펴보면 이상의 견해 중에서는 최수일(2008)의 주장과 본 논문의 분석 결과가 일치한다. 필자는 이처럼 상이한 견해가 병존하는 상황에서, 나름의 근거와 방법론에 기반을 두면서『개벽』의 시기구분을 전기와 후기의 2개로 시기구분하고 그 시점을 분명하게 제시하였다. 또한 당연하게도 이런 작업이 별로 새롭지 않거나 지나치게 세밀한 일로 생각하지는 않는다.

단어를 소속 토픽으로 치환하여 토픽 연결망 지도를 도출할 것이다.

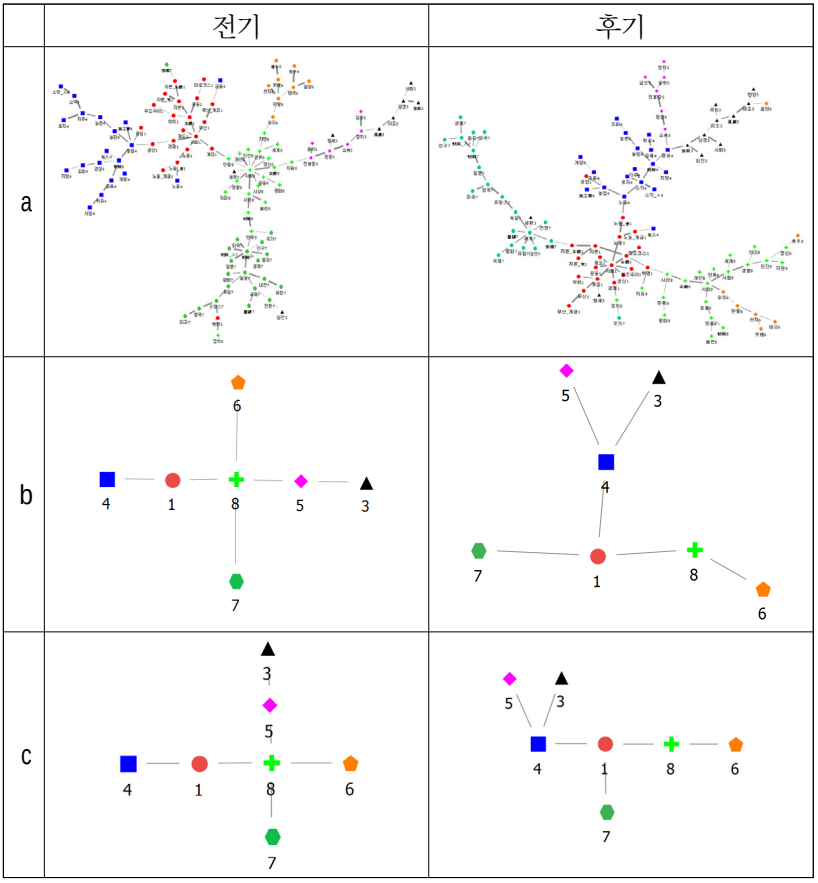
우선, 지금까지 사용해 오던 '문서 — 토픽 단어' 정보를 앞 절에서 확정한 전기와 후기로 나누었다. 전기와 후기의 텍스트 분량은 <표 2>와 같다.

〈표 2〉 텍스트 단위별 전·후기 분포 상황 단위: 개

시기 / 텍스트 단위	전기	후기	전체
논설	177	157	334
문서	1,802	1,624	3,426
토픽단어	12,696	13,364	26,060
전체단어	83,565	74,829	158,394

'문서 — 토픽 단어 — TFIDF값'을 전·후기로 나누어 각각 넷마이너에 입력하였다.[48] 전술한대로 이 단어들은 나중에 토픽으로 치환할 것이며 단어 개수도 아주 많지는 않다. 따라서 상대적으로 빈도가 낮은 토픽 단어들도 절삭하지 않고 연결망에 모두 남도록 하였다. 그리고 '문서 — 토픽 단어'로 된 2항 연결망을 코사인 유사도에 기반을 둔 '토픽 단어 — 토픽 단어' 의 1항 연결망 형태로 바꾸었다. 토픽 단어를 잇는 연결망은, 지나치게 복잡하거나 반대로 고립·단절 없이 전체 구조를 잘 드러내는 패스파인더 방식으로 산출하였다. 이렇게 산출한 결과가 [그림 4]의 'a'이다. 일부 예외는 있지만 대부분 토픽 단어들이 자신과 같은 소속의 단어들과 밀집해 있음을 확인할 수 있다. 이는 토픽 모델링의 토픽 산출 과정이 그러한 단어 간 응집도에 따른 것이므로 예상 가능한 결과이다. 하지만 동일한 토픽에 속한 단어 간의 응집성을 연결망 지도에서 눈으로 확인하는 의의가 있다.

48) 토픽 단어는 95종 104개인데, 이 중 전기에는 101개가, 후기에는 102개가 출현하였다.



비고: 6개 연결망 모두에서 각각 7개 점(노드)의 색깔과 형태는 7개 토픽에 일대일로 대응한다.

[그림 4] 『개벽』 전·후기의 토픽 연결망 비교

다음 단계로 이러한 토픽 단어 연결망 지도를 단순화한 토픽 연결망 지도를 그렸다. 이 작업은 수작업으로 진행하였다. 토픽 단어들이 모여 있는 곳을 해당 토픽을 나타내는 한 개의 '점'으로 치환하였다. 다만 이렇게 치환한 점들 간의 연결 관계도 잘 드러나도록 유의하였

다. 그 결과가 [그림 4]의 ‘b’이다. ‘b’는 ‘a’와 비교하기 쉽도록 토픽의 위치를 ‘a’의 토픽 단어 밀집 위치에 대응시켜 그린 것이다. 전체 연결망 속 ‘점’의 위치나 ‘선’의 개수는 연결망의 중심 여부를 판단하는 데 핵심 기준이 된다. 이 연결망을 보면 전기에는 토픽8이, 후기에는 토픽1이 각각 연결망의 중심에 위치하고 있음을 알 수 있다.

이상에서『개벽』논조의 중심은 전기에는 토픽8, 즉 개조론이었다가, 후기에는 토픽1, 즉 사회주의로 옮겨갔다고 말할 수 있다. 이러한 사실은 선행 연구의 성과를 재확인하는 의의가 있다. 그렇다면 이 토픽 연결망 지도에서 기존 연구에서 드러나지 않은 정보는 없을까.

전기와 후기의 논조를 좀 더 비교하기 쉽게 조정한 것이 [그림 4]의 ‘c’이다. 점과 선의 구조는 손대지 않고 그 위치 표시만 조정하였다. 이렇게 하면 연결망 구조의 변경과 지속을 한눈에 파악할 수 있다. 이것으로 두 시기를 비교하면, ‘4 — 1 — 8 — 6’을 잇는 횡축에는 변동이 없다. 변동은 종축에서 일어났다. 전기에는 토픽8이 각각 토픽7, 토픽5와 접속하였고, 토픽5는 다시 토픽3과 접속해 있었다. 그 결과 토픽8이 중심점이 되었다. 후기가 되면 토픽7은 이동하여 토픽1과 접속하였다. 또한 토픽5와 토픽3은 각각 토픽4와 접속하였다. 그 결과 연결망의 중심, 즉 논조의 중심은 토픽1로 변했다. 이런 사실은 후기 논조의 사회주의화를 더 넓은 맥락에서 보여주는 이점이 있다. 논조 변화는『개벽』을 구성하는 여러 주제들 간의 접속과 단절을 동반하면서 일어난 것이다.[49]

49) 익명의 심사위원 한 분이 이에 관하여 다음과 같은 요지의 질문을 해 주었다. ‘논조의 옮겨감은 분절인가 수평이동인가, 아니면 어떤 다른 의미화의 형태인가’, ‘토픽연결망의 구조적 형태변화를 논조라는 중심주제의 이동과 더불어 어떻게 다르게 이해할 수 있을까’라고 하였다. 이 지적의 의미를 정확하게 이해하진 못했지만, 논조의 이동, 그리고 이와 연결망 구조 변화의 관계에 대하여 보충설명이 필요할 듯하여 다음을 부연한다. 본 논문에서는 일단 토픽 연결망의 점(= 노드)에 해당하는 토픽들은 전기와 후기가 동일하다고 상정하였다. 물론 각

4. 사회주의 영향의 제한성

3장에서는 토픽 연결망 지도를 통하여『개벽』후기 논조의 사회주의화를 확인하였다. 이 장에서는 그러한 사회주의화가 개벽 주도층에게 중요한 영향을 미쳤는지 여부를 살펴볼 것이다. 그런데 '영향력의 정도'를 판단하는 것은 까다로운 일이다. 그 기준이 모호하기 때문이다. 머리말에서는 그것을 '사상적 변화 여부'로 표현했지만 그렇다고 모호함이 줄어들지는 않는다. 본 논문에서는 '영향력'의 척도를『개벽』의 다른 논설과 비교하는 데서 구하려 한다. 전술하였듯이 당시 한국사회의 식자층에게서 사회주의는 '대세'였다.『개벽』의 후기 논조도 사회주의가 우세를 점했다. 그렇다면 핵심 관찰 지점은 개벽 주도층의 논설에서 사회주의에 해당하는 토픽1의 단어가 현저히 적게 등장하는가 여부이다. 물론 이 때에도 '현저히'를 어떻게 판별하는가가 관건이다. 이 지점에서 필자가 선택한 방법은 통계적 검정이다. 우선『개벽』의 후기 논설을 개벽 주도층의 것과 여타 사람들의 것으로 양분한다. 그리고 각각에서 토픽1의 단어가 비슷한 정도로 나타나는지 여부를 살펴본다. 만약 개벽 주도층의 논설에서 토픽1의 빈도가 특별히 낮다는 사실이 통계적으로 드러난다면, 우리는 개벽 주도층에 대한 사회주의의 영향력이 제한적이었다고 말할 수 있을 것이다.

토픽별로 전기와 후기에 토픽 단어의 빈도나 구성이 변할 수도 있으나 그 차이는 무시하였다. 그리고 논조는 토픽 연결망 그 자체이다. 조금 더 구체적으로 표현하자면 논조는 토픽 연결망의 배치 효과라고 할 수 있을 듯하다. 그 연장선상에서 토픽 연결망의 중심이 곧 논조의 중심이라 할 수 있다. 한편, 전・후기 논조변화를 발생시킨 가장 핵심 요인은 토픽 간의 연결 관계 변동이다. 그리고 이 변동은 각 토픽 간의 힘 관계 변화를 총체적으로 반영한다. 예컨대 전기에서 후기로 논조 중심이 바뀐 것은, 전기의 토픽8이 토픽7 및 토픽5와 각각 맺고 있던 관계가 약화된 측면도 있지만, 이와 함께 토픽4와 토픽5 및 토픽3, 그리고 토픽1과 토픽7이 가까워진 측면도 동시에 작용했다고 할 수 있다.

4.1. 관찰 단위와 관찰 지표의 확장

지금까지는『개벽』의 전체시기를 대상으로 '문서'라는 텍스트 단위에 집중하였다. 여기서부터는 관찰의 시공간적 단위를 '후기'의 '논설'로 조정할 것이다. 개벽 주도층에 대한 사회주의의 영향력 문제를 살펴보기 위함이다. 논설이라는 텍스트 단위는 지금까지 진행해 온 문서 기반의 토픽에 관한 정보를 포함하고 있다. 이와 더불어 그것은 필자라는 상위 범주와도 밀착해 있다. 이에 문서 단위의 정보를 논설 단위로 재조정하였다. 가장 중요한 작업은 '문서 — 토픽 단어 — TFIDF 값'이라는 핵심 정보를 '논설 — 토픽 단어 — TFIDF값'으로 재구성하는 일이다.[50]

이와 관련하여『개벽』의 후기 논설을 선별하고 양분하였다. 먼저『개벽』후기의 '기명(記名) 논설'을 주요 검토 대상으로 선별하였다. <표 3>을 3장 2절의 <표 2>와 함께 살펴보면,『개벽』후기 논설 중 필자 미상이 36편이나 되지만, 개벽 주도층의 글인지 여부를 가려낼 수 없어서 제외하였다.[51] 후기의 논설 157편에서 36편을 제외한 121편의 논설을, 다시 개벽 주도층의 29편과 '일반 필자층'의 92편으로 나누었다.[52]

다음에는 관찰 지표로 토픽1 이외에 토픽8을 추가하였다. 이 역시 지금까지는 실용적으로 토픽1, 즉 사회주의 주제만 거론하였다. 토픽1

50) 문서와 달리 논설 간에는 텍스트의 규모에 편차가 크기 마련이다. 그러나 사이킷런으로 TF-IDF를 산출할 때 각 논설의 길이를 고려한 정규화 작업도 포함하므로 큰 문제는 없으리라 보았다.

51) 최근에 기계학습을 통한 문체 판별로『개벽』논설의 저자를 판별하는 성과가 나왔다. 본 논문에서는 그 성과를 반영하지 못했으나 주목할 만한 내용이라 할 수 있다. 최지명(2018),「기계학습을 이용한 역사 텍스트의 저자판별: 1920년대 개벽 잡지의 논설 텍스트」,『언어와 정보』22(1), 한국언어정보학회.

52) 본 논문에서 사용하는 '일반 필자층'은 '개벽 주도층'의 비교군을 가리키려고 개벽 주도층 이외의 필자들을 통칭한 것이다. 따라서 '일반 필자층'이 반드시 단일한 정체성을 지닌 사람들이라고 할 수는 없다.

〈표 3〉 『개벽』 주요 논설의 필자군과 필자별 논설 수의 분포 양상 단위: 편[명]

시기 / 필자군	전기	후기	합계
미상	49[1]	36[1]	85[1]
	미상(49)	***미상***(36)	
개벽 주도층	62[3]	29[4]	91[4]
	이돈화(29), ***김기전***(24), ***박달성***(9)	***이돈화***(14), ***김기전***(6), 차상찬(5), ***박달성***(4)	
일반 필자층	66[32]	92[52]	158[79]
	선우전(10), ***이광수***(9), 강인택(7), ***이동곡***(5), 윤익선(3), 동양실주인(2), 신식(2), ***이성환***(2), 이종린(2), 황의돈(2), 권동진(1), 김규호(1), 김병준(1), 김영갑(1), 김윤식(1), 김찬영(1), 박사직(1), ***배성룡***(1), 석진형(1), 실명생(1), 안석응(1), 윤필균(1), 이상재(1), 이인영(1), 임규(1), 임주(1), 장덕수(1), 정규선(1), 최승만(1), 최중갑(1), 현상윤(1), 황석우(1)	이성태(6), 정지현(5), 김기진(4), 박형병(4), ***배성룡***(4), 김경재(3), 김명식(3), 박진순(3), 양명(3), 이순탁(3), 주종건(3), 최화운(3), 견지동인(2), 김양수(2), 신일용(2), 옥천생(2), ***이동곡***(2), 이민창(2), ***이성환***(2), 청진학인(2), BSL생(1), O民(1), PSL생(1), XY生(1), 권덕규(1), 김성(1), 김자립(1), 김정설(1), 김철산(1), 덕월산인(1), 박영희(1), 박은식(1), 박현영(1), 반구실주인(1), 부지암(1), ***선우전***(1), 안재홍(1), 양건식(1), 원종린(1), 월평인(1), 윤주(1), ***이광수***(1), 이창림(1), 이철(1), 이청우(1), 인사학인(1), 임장화(1), 전영택(1), 첩기생(1), 최남선(1), 허죽재(1), 황영환(1)	
합계	177[36]	157[57]	334[84]

비교: 1. 필자 미상은 편의상 1인으로 간주하였다.
2. 인명 뒤 소괄호(‘( )’) 속 숫자는 그 필자의 논설 수이며 단위는 ‘편’이다.
3. 대괄호(‘[ ]’) 속 숫자는 필자 수이며 단위는 ‘명’이다.
4. 전・후기에 공통적으로 등장한 필자는 글자 굵기와 밑줄, 이탤릭체로 강조하였다.

이 가장 중요하긴 하지만, 혹시 다른 토픽도 함께 살펴볼 순 없을까 라는 의문이 들었다. 『개벽』의 논설을 나타내는 특징으로, 이론적으로는

토픽1~8의 7개 토픽 소속 단어들이 모두 가능하다. 그렇다면 이 논설 중 사회주의 영향력을 나타내는 지표는 무엇일까. 토픽1일 뿐일까? 표로 정형화한 데이터에서 특정 요인과 관계가 있는 다른 요인을 판별하는 지표가 있다. 상관 계수가 그것이다. 가장 대표적으로는 '피어슨 상관계수'(Pearson Correlation Coefficient, PCC)를 많이 사용한다. 본 논문에서는 이를 이용하여 토픽1과 상관성이 높은 토픽을 찾아보았다. 그 결과는 <표 4>와 같다.

〈표 4〉 '사회주의' 토픽(토픽1)과의 상관성 비교

시기 / 순위	전체	전기	후기
1	-0.220 [8]	0.148 [4]	-0.243 [8]
2	-0.118 [6]	-0.117 [8]	-0.121 [6]
3	0.097 [4]	-0.105 [6]	-0.114 [3]
4	0.047 [7]	0.090 [7]	0.104 [4]
5	-0.037 [3]	0.017 [3]	-0.063 [5]
6	-0.031 [5]	0.006 [5]	0.009 [7]

비고: 1. '[ ]' 속 숫자는 토픽 번호이다.
2. 순위는 절대값이 큰 것을 1위로 하여 내림차순으로 정렬한 것이다.
3. 전체적으로 토픽1과 가장 가까운 토픽8에는 음영을 표시하였다.

<표 4>에서 알 수 있듯이 전체적으로 토픽1과 가장 상관성이 높은 토픽은 토픽8, 즉 개벽의 개조론이다. 후기에 오면 그 상관성은 더욱 커진다. 양자의 관계는 음의 상관관계에 있다. 즉 토픽8은 사회주의의 영향력이 클수록 낮아지고, 그 반대일수록 높아지는 경향이 있다는 것이다. 물론 이 계수는 통상 절대값이 0.3 이상일 때 뚜렷한 상관관계가 있다고 말할 수 있다고 한다. 후기의 토픽8은 그 값이 0.243으로 0.3에 미달한다. 그러나 다른 토픽에 비하여 그 값이 크므로 일단 사회주의 영향력을 판단하는 지표에 포함시켰다.

4.2. 통계적 검정

앞 절에서는 개벽 주도층의 논설과 일반 필자층의 논설을 양분하여 사회주의의 영향력을 서로 비교할 집단으로 삼았다. 이 절에서는 두 집단에서 사회주의 관련 지표들(토픽1・토픽8 단어)의 평균값에 유의미한 차이가 있는지 여부를 살펴보고자 한다.

첫 순서로, 어떤 값을 가지고 비교할 것인가를 정해야 한다. 본 논문에서는 논설 내 토픽의 상대 빈도 비중이 적절하다고 보았다. 상대 빈도란 TFIDF값을 가리킨다. 이를 위해서는 우선, 특정 논설에 있는 토픽 단어의 상대 빈도를 모두 합산하였다. 다음으로 토픽1로 두 집단을 비교하기 위해 토픽1 소속 단어들의 상대 빈도를 합했다. 전자를 분모에, 후자를 분자에 할당해서 나온 결과가 이 절에서 사용할 값이다. 이렇게 해서 토픽1에 관하여는 개벽 주도층의 경우 29개의 값을, 일반 필자층의 경우 92개의 값을 얻었다. 토픽8에 대해서도 동일하게 진행하였다.

둘째, 두 집단 간 평균 비교를 'T검정'으로, 특히 그 중에서도 '독립표본 T검정'으로 수행하였다. T검정은 두 집단 간의 평균이 통계적으로 유의미한 차이를 보이는지 여부를 따질 때 널리 사용하는 분석 방법이다.[53] 이 비교를 '독립표본 T검정'으로 진행한 이유는, 개벽 주도층과 일반 필자층이 사회주의 영향의 측면에서 서로 간섭하는 관계가 아니라고 보았기 때문이다. 또한 이 121개 논설이 비록 『개벽』 후기의 기명 논설을 망라한 것이지만, 당초에 선별한 334개의 주요 논설 그

53) '검정'(檢正, testing)은 "(가설이) 옳은지 검사한다"는 뜻이다. 비슷한 용어로 '검증'(檢證, validation)이 있다. 이 용어는 '증거를 조사한다'라는 뜻을 가진다. 두 단어의 의미가 어느 정도 상통하지만 "통계적 가설의 시비를 논할 때는 '검정'이 더 명확한 표현"이다. 이에 관하여는 다음 저서를 참고하였다. 조엘 그루스(2016), 박은정・김한결・하성주 역, 『밑바닥부터 시작하는 데이터 과학 — 데이터 분석을 위한 파이썬 프로그래밍과 수학・통계 기초』, 서울: 인사이트, p. 83.

자체가『개벽』의 전체 기사 중 일부이므로 '표본'이라 할 수 있다.

실제 검정 과정은 통계 전문 소프트웨어인 SPSS로 진행하였다.[54] 과정은 그리 복잡하지 않지만 논리적으로는 다음 세 단계의 과정을 거쳤다. 단계별로 간단하게 설명하고자 한다. 또한 설명은 토픽1을 기준으로 하며, 토픽8은 토픽1과 거의 동일한 과정이므로 결과와 해석 중심으로 간략하게 부연할 것이다. 첫 단계는 두 집단의 입력값에 대한 정규성 검정을 하였다. 정규성 검정이란 해당 값이 정규분포를 띠는지 여부를 검정하는 일이다. 왜냐하면 T검정을 적용한 결과값이 의미를 가지려면, 대상이 되는 입력값이 정규 분포를 가져야하기 때문이다. 통상 이러한 검정은 일반적 가설을 전제하면서 출발하되, 그 가설을 부정할 만한 드문 결과가 나올 때 비로소 그 가설과 상반된 결론을 승인하는 방식이다. 사회과학 연구에서는 통상 그 드문 경우를 5% 이하, 즉 '유의확률' 0.05 이하로 설정한다. 필자는 이에 따라 입력값을 넣어서 <표 5>와 같은 정규성 검정 결과를 얻었다.[55]

<표 5>의 '콜모고로프-스미르노프'(Kolmogorov-Smirnov, KS)와 '샤피로-윌크'(Shapiro-Wilk, SW)는 정규성 검정에서 많이 사용하는 검정

〈표 5〉 토픽1 소속 단어의 토픽 비중값에 대한 정규성 검정 결과

	Class	Kolmogorov-Smirnova			Shapiro-Wilk		
		통계	자유도	CTT 유의확률	통계	자유도	CTT 유의확률
tp1	개벽주도층	.091	29	.200*	.959	29	.317
	일반필자층	.051	92	.200*	.977	92	.105

비고: 이하의 설명은 SPSS의 결과값을 그대로 옮겼다.
*. 이것은 참 유의성의 하한입니다.
a. Lilliefors 유의확률 수정

54) 'IBM SPSS Statistics'를 이용하였다.

55) SPSS의 메인메뉴에서 '분석\기술통계량\데이터 탐색'으로 들어가 진행하였다. 이하에서 제시하는 결과값 표 및 그래프 등은 SPSS의 데이터 탐색 및 T검증의 결과에서 인용하였다.

방법이다. 데이터 규모에 따라 양자 중 적당한 것을 선택하기도 하지만, 통상 둘 중 하나만 통과하면 된다고 한다. '통과'란 '유의확률'값이 0.05보다 크게 나와 해당 입력값이 정규분포를 가지게 되는 경우를 말한다. <표 5>를 보면 두 집단 모두 0.05를 넘어 양 검정을 통과하였다. 따라서 정규분포를 가지는 사실을 확인하였으므로 T검정을 진행할 수 있게 되었다. 토픽8의 경우 동일 과정을 진행하였을 때 처음에는 한쪽 검정만 통과하였으나, 이상치(Outliar) 몇 개를 제거하고 다시 실시하니 모두 통과하였다.[56]

둘째 단계는 등분산 여부를 검정하였다. 이 검정은 비교하는 두 집단의 분산에 차이가 있는지 여부를 판별하는 것이다.[57] SPSS에는 이 검정과 T검정 결과가 함께 출력된다. <표 6>를 보면 가장 일반적 방법의 하나인 레빈(Levene) 등분산 검정 결과가 나와 있다. 이 검정도 마찬가지로 '두 집단이 등분산이다.'를 가정한다. 실제로 나온 '유의확

〈표 6〉 토픽1에 관한 두 집단의 등분산 검정 및 T검정 결과

		Levene의 등분산 검정		평균의 동일성에 대한 T 검정						
		F	유의확률	t	자유도	유의확률(양측)	평균차이	표준오차차이	차이의 95% 신뢰구간	
									하한	상한
tp1	등분산을 가정함	4.528	.035	-3.133	119	.002	-.109787	.035047	-.179183	-.040390
	등분산을 가정하지 않음			-3.718	65.543	.000	-.109787	.029525	-.168744	-.050830

56) 개벽 주도층의 유의확률은 각각 0.200 (KS), 0.360 (SW)로 나왔다. 일반 필자층의 경우 각각 0.099 (KS), 0.067 (SW)가 나왔다.

57) 등분산 검정은 앞의 정규성 검정과는 달리 검정을 통과하지 않더라도 '이분산'으로 된 T검정 값을 이용하면 되므로 큰 문제는 없다. 다만 양자를 구별하는 것은 등분산 여부에 따라 T검정 과정과 그 값이 상이하기 때문이다.

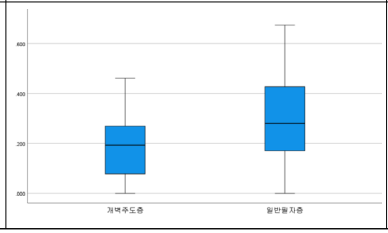
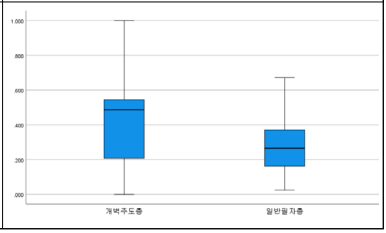
률’은 0.035로 0.05보다 작아서 이 가정은 더 이상 유효하지 않다. 따라서 두 집단은 ‘등분산을 가정하지 않음’에 해당한다. 토픽8의 ‘유의확률’은 토픽1보다 더 작은 0.01 이하가 되었으므로, 토픽8 역시 이분산, 즉 ‘등분산을 가정하지 않음’에 해당하였다.

셋째 단계는 T검정 결과를 도출하는 것이다. 이 역시 <표 6>에 나와 있다. ‘등분산을 가정하지 않음’에 해당하는 T검정의 ‘유의확률’을 보면 0.000이다. T검정도 ‘두 집단의 평균은 동일하다.’라는 가정에서 출발한다. 그래서 ‘유의확률’이 0.05보다 작으면 그 가정은 유효하지 않게 되어 두 집단 간 평균은 동일하지 않다고 볼 수 있다. 실제로 이번 결과는 0.05보다 작은 0.000이 나와서 동일성 가정은 더 이상 유효하지 않다. 따라서 개벽 주도층의 논설과 일반 필자층의 논설은 토픽1의 평균에서 유의미하게 차이가 난다고 할 수 있다. 토픽8도 동일한 과정으로 진행하여 0.005의 유의확률을 얻었다. 이 역시 0.05보다 작아서 이 경우에도 두 집단 간의 차이가 유의미하게 난다고 말할 수 있다.

지금까지『개벽』후기 기명 논설에서 토픽1과 토픽8의 상대 빈도 평균을 각각 비교한 결과, 두 경우 모두 ‘유의 수준 0.05에서 개벽 주도층의 논설과 일반 필자층의 논설에 차이가 있다’는 점을 확인할 수 있었다. 그렇다면 마지막으로 이 검정 결과에서 나온 ‘차이’가 본 논문이 목표로 한 사회주의의 영향력 문제에 어떤 시사점을 주는지 개괄하겠다.

[그림 5]는 개벽 주도층과 일반 필자층의 상대빈도 평균을 토픽1과 토픽8로 구분해서 나타낸 것이다. ‘평균’은 두 집단의 평균값을 부등호로 나타내었다. ‘상자그림’은 최대값, 최소값, 중간값 등 기술통계량을 나타낸 것이다.[58] 토픽1에서는 일반 필자층이 더 높고 토픽8에서

58) 이 상자그림은 각 집단의 기본적인 통계값을 집약해서 나타내는 흔한 방식이다. 상자 속 가로줄이 중간값이며, 상자와 연결된 위쪽 선의 끝이 최대값, 그 반대쪽이 최소값이다.

	토픽1			토픽8		
평균	개벽주도층 0.18672	<	일반필자층 0.29651	개벽주도층 0.41586	>	일반필자층 0.27498
상자그림						

[그림 5] T검정으로 구한 두 집단의 평균 비교

는 개벽 주도층이 더 높다. 이는 앞 절에서 확인한 바와 같이 토픽1과 토픽8이 음의 상관관계를 가진다는 사실과 부합한다. 다만 여기서는 더 적극적인 의미로 해석할 수 있다. 『개벽』 후기 두 집단의 논설 사이에 통계적으로 유의미한 차이가 있음이 드러났다. 이와 더불어 개벽 주도층에게 토픽1의 평균은 더 낮게, 토픽8의 평균은 더 높게 나왔다. 양자를 종합하면, 결국 『개벽』 후기에 사회주의가 논조의 중심이 될 만큼 유행했으나, 적어도 개벽 주도층에 대한 영향력은 상대적으로 낮았다고 말할 수 있다.[59]

59) 심사위원들이 '결론을 정해 놓고 통계적 검증을 하는 것에 대한 인문학적 문제의식과 내용을 보강'(a)할 것을 주문하였다. 또한 '사회주의 영향이 통계적으로 낮았다는 것이 무엇을 의미하는 것인지 분명하지 않다'(b)고 하면서, '이를 비교하기 위해서는 개벽주도층의 전기와 후기 논설을 서로 비교할 필요가 있다'(c)라고 제안하였다. 독자들도 이런 생각을 할 수 있을 듯하여 이에 관한 의견을 덧붙인다. 'a'의 경우 통계적 검정이라는 방법 자체가 상식적인 '가설'을 전제로 해서 가설을 무너뜨릴 만한 이유를 확인할 때 그 가설과 상반된 새로운 주장을 하는 논증 절차이다. 따라서 '결론'을 정해놓기 보다는 '가설'을 제시한 것으로 봐 주었으면 한다. 물론 이런 방법적 의미는 본문에서 이미 밝혀 놓았다. 또한 이런 통계적 검정을 하는 '인문학적' 문제의식은 머리말에서 어느 정도 밝혔다고 생각한다. 즉 현재 학계에서는 '『개벽』 논조의 사회주의화가 개벽주도층의 사상적 변화까지 동반하였나'라는 문제를 둘러싸고 견해가 상충한다. 이런 상

5. 맺음말

『개벽』 후기 논조가 사회주의로 변한만큼 개벽 주도층의 입장도 사회주의로 기울었을까? 이 질문이 본 논문의 출발점이었다. 개벽 주도층은 이돈화, 김기전, 박달성, 차상찬의 네 명으로 한정하였다. 그 과제를 풀기 위하여 질문을 다시 세 개의 하위 질문으로 나누었다. 첫째는 사회주의화를 무엇으로 판단할 수 있을까, 둘째는『개벽』의 후기 논조는 과연 사회주의 중심으로 전환하였나, 셋째는『개벽』 후기에 주도층의 성향은 사회주의로 크게 기울었나 라는 질문이 그것이다. 본문의 3개 장에서 이 세 질문을 하나씩 다루었다. 이하에서는 본문 내용을 요약하는 것으로 맺음말을 대신하고자 한다.

2장에서는 사회주의를 판별할 지표를 확정하였다.『개벽』의 주요

황에서, 본 논문은『개벽』 후기에 논조의 사회주의화가 뚜렷하게 나타났지만, 개벽주도층의 논설에서는 그 영향이 다른 집단에 비해 '유의미하게' 낮았다는 것을 증명한 것이다. 또한 이런 결론은 전술한 견해들 중에서 '제3의 입장', 즉『개벽』 후기에 가서도 개벽주도층은 사회주의로 기울지 않았다는 주장과 가장 밀접하다고 말할 수 있다. 이 설명으로 'b'에 대한 답변도 되었으리라 생각한다. 한편 'c'의 제안에 대한 필자의 답변은 먼저 3장 1절의 [그림 2]를 염두에 두면서 시작하고 싶다.『개벽』 후기에 사회주의적 경향이 급증한 것은 주지의 사실이며 따라서 개벽주도층의 논설에도 토픽1의 단어들이 전기에 비해 증가하였다. 필자도 이 점을 부인하지 않는다. 그러나 이런 유행담론을 개벽주도층이 자신의 논설에 반영하는 것과, 사회주의를 적극적으로 수용하는 것은 다르다는 것이 필자의 문제의식이다. 여기에 관한 정성적인 접근은 필자의 선행연구에서 전개한 바 있고, 이 점은 머리말에서 논거와 함께 제시하였다. 본 논문에서는 정량적 접근으로 이 문제를 다루고자 하였다. 그럴 경우 개벽주도층의 전기와 후기 논설을 비교해서는 유의미한 결론을 내리기 어렵다. 당연히『개벽』의 모든 논설에서 전기에 비해 후기에 토픽1 관련 단어들이 증가할 것이기 때문이다. 이전보다 사회주의적 단어가 늘어났으므로 사회주의의 영향이 증가하였다고 평가하는 것은, 틀린 진술은 아닐지 모르지만 이 문제에 관한 현재의 연구 지형상 별로 의미 있는 해석은 아니다. 따라서 사회주의 영향의 상대적 차이를 살펴보기 위하여, 시점을『개벽』 후기에 고정하되 개벽주도층과 여타 필자층 간의 비교에 역점을 두었다.

논설 334개를 선별하고 전처리한 뒤 토픽 모델링으로 7개의 주제를 추출하였다. 단어 빈도로는 상대 빈도, 즉 TF-IDF값을 이용하여 문서의 특징이 잘 드러나도록 하였다. 이렇게 하여 총 6,707종 158,394개 단어 중 95종 104개의 단어를 토픽 단어로 획득하였다. 이를 다시 '토픽 보정'이라는 다소 엄격한 절차를 거쳐 정리 정돈하였다. 그 중 토픽1에는 '자본', '계급', '노동'부터 '공업', '혁명', '노동_계급'에 이르는 20개의 단어가 들어 있었다. 이를 다른 토픽과 비교하여 '사회주의' 주제로 평가하였다. 그리고 이 토픽 단어를 사회주의 판별의 가장 중요한 지표로 간주하였다.

3장에서는『개벽』후기 논조의 사회주의화 양상을 살펴보았다. 이를 위하여 우선, 토픽 소속이 미확정인 '복속형' 토픽 단어의 소속을 확정하였다. 다음으로 문서의 토픽 구성을 산출하고 이를 바탕으로 시기구분을 하였다. 여기에는 토픽 구성의 분기별 변화 관찰과, 토픽 단어의 상대빈도를 활용한 분기 연결망 산출 등이 필요하였다. 그 결과『개벽』의 논조 변화를 당초와 같이 전기와 후기로 나누어 살펴보는 것이 적절함을 확인하였다. 마지막으로 전기와 후기의 토픽 연결망 지도를 그려 논조의 구조와 변동 양상을 살펴보았다. 26,060개의 '문서 — 토픽 단어 — TFIDF값'을 입력하여 7개 토픽 간의 관계를 시각화한 토픽 연결망 지도를 산출하였다. 그 결과『개벽』전기에는 개벽의 개조론이 연결망의 중심에 있었으나, 후기에는 그 중심이 사회주의로 이동하였음을 확인하였다.

4장에서는 사회주의가 개벽 주도층에 미친 영향을 살펴보았다. 우선『개벽』의 후기 논설 중 필자가 드러난 논설 121개를 개벽 주도층의 29개와 '일반 필자층'의 92개로 양분하였다. 사회주의의 영향력을 살펴볼 지표로 토픽1과 토픽8을 선택하였다. 비교 단위를 논설과 확대하였으므로, 토픽 단어의 상대 빈도도 그에 맞게 재산출하였다. 이를 바탕으로 토픽1과 토픽8이 각 논설의 토픽 단어 전체에서 차지하는 비중

을 각각 계산하였다. 다음으로 이 값을 사용하여 두 집단에서 토픽1과 토픽8의 비중에 각각 유의미한 차이가 있는지를 T검정이라는 통계적 방법으로 분석하였다. 그 결과 개벽 주도층에 대한 사회주의의 영향은 여타 필자들에 비하여 통계적으로 유의미하게 낮았음이 드러났다.

참고문헌

【자 료】

「한국근현대잡지자료: 개벽」, 국사편찬위원회 한국사데이터베이스, 2020.05.23. http://db.history.go.kr/item/level.do?sort=levelId&dir=ASC&start=1&limit=20&page=1&pre_page=1&setId=-1&totalCount=0&prevPage=0&prevLimit=&itemId=ma&types=&synonym=off&chinessChar=on&brokerPagingInfo=&levelId=ma_013&position=-1.

【논 저】

곽기영(2020), 「데이터 애널리틱스. 텍스트마이닝7: 토픽 모델링」, 유튜브, 2020.09.15. https://www.youtube.com/watch?v=CuW7-QkNMNE.

권오성(2019), 「변덕꾸러기 토픽 모델링 어떻게 다뤄야 하나」, 한겨레 웹페이지, 한겨레신문, 2020.09.30. http://www.hani.co.kr/arti/science/technology/907134.html#csidx45abb92d7e27dd4b65c5203ceef0fce.

김동욱・이수원(2017), 「단어 유사도를 이용한 뉴스토픽 추출」, 『정보과학회논문지』 44(11).

김정인(2007), 「‘개벽’을 낳은 현실, ‘개벽’에 담긴 희망」, 임경석・차혜영 외, 『『개벽』에 비친 식민지 조선의 얼굴』, 서울: 모시는 사람들.

벤자민 벵포트・레베카 빌브로・토니 오제다(2019), 박진수 역, 『파이썬으로 배우는 응용 텍스트 분석』, 파주: 제이펍.

신서인(2017), 「네트워크 분석을 이용한 복지 담화 연구」, 『개념과 소통』 20, 한림대학교 한림과학원.

이기창(2020), 『한국어 임베딩 — 자연어 처리 모델의 성능을 높이는 핵심 비결, Word2Vec에서 ELMo, BERT까지』, 서울: 에이콘.

이재연(2016), 「토픽 모델링으로 본 <개벽>의 주제 지도 분석」, 『상허학보』 46, 상허학회.

이재윤(2006), 「지적 구조의 규명을 위한 네트워크 형성 방식에 관한 연구」, 『한국문헌정보학회지』 40(2), 한국문헌정보학회.

조엘 그루스(2016), 박은정・김한결・하성주 역, 『밑바닥부터 시작하는 데이

터 과학 — 데이터 분석을 위한 파이썬 프로그래밍과 수학・통계 기초』, 서울: 인사이트.
최수일(2008),『개벽 연구』, 서울: 소명출판사.
최지명(2018),「기계학습을 이용한 역사 텍스트의 저자판별: 1920년대 개벽 잡지의 논설 텍스트」,『언어와 정보』22(1), 한국언어정보학회.
허수(2011),『식민지 조선, 오래된 미래』, 서울: 푸른역사.
____(2015),「『개벽』의 종교적 사회운동론과 일본의 '종교철학'」,『인문논총』72-1, 서울대학교 인문학연구원.

원고 접수일: 2021년 2월 9일
심사 완료일: 2021년 2월 20일
게재 확정일: 2021년 2월 20일

ABSTRACT

A New Approach to Socialization in *Gaebyeok*'s Tone:

Focusing on Topic Network Analysis

Hur, Soo*

This paper aims to investigate whether the phenomenon in which *Gaebyeok* (開闢) acquired a socialistic in tone in its late years impacted on the leadership of *Gaebyeok*. The analytical approach used to address this question includes three steps: topic modeling, network analysis, and statistical testing. The indicators used to define socialization are first identified and determined. 7 topics and 104 topic words are selected at the stage of topic modeling after the data pre-processing of 334 main editorials of *Gaebyeok*. Twenty of these topic words that are classified into topic 1 are the most important indicators for defining socialism. This paper then examines how the *Gaebyeok* became socialistic in tone. By inputting 26,060 'document-topic words-tfidf value', a topic network map was computed, which reflects the relationships among the seven topics. The result reveals that the network centrality moved from Kaejoron (改造論) in the early *Gaebyeok* years to socialism in the late *Gaebyeok* periods. Lastly, this paper evaluates the influence of socialism on the leadership of

* Associate Professor, Department of Korean History, Seoul National University

Gaebyeok. One hundred twenty-one identified editorials published in the late *Gaebyeok* periods could be divided into two types: one written by the '*Gaebyeok* leaders' and one written by the 'ordinary authors'. Using these sources, this study examines whether there is a significant difference between two groups in terms of the proportions of topic 1 and topic 8 via the 'T-test', respectively. The result shows that socialism's influence on the leaders of *Gaebyeok* was significantly weaker than that of other authors.