빅데이터 분석 기반 문화콘텐츠 학위논문 연구 동향 (2004년~2020년)

오정심

[국문초록]

이 논문은 빅데이터 분석 방법을 통해 2004년부터 2020년까지 문화콘텐츠학 전공자들의 석·박사 학위논문 1,812편의 국문초록과 서지정보를 분석하여 문화콘텐츠 학위논문의 주요 연구 영역, 연구 주제, 연구 흐름을 연구한 결과물이다.

본 연구를 통해 얻은 결과를 요약하면 첫째, 문화콘텐츠 학위논문의 연구 영역은 크게 '사회와 문화', '문화콘텐츠 활용', '문화콘텐츠 장르', '문화콘텐츠 산업', '문화콘텐츠 제작 기술' 5개로 구분된다. 이중에서 활용은 인문학 등 다른 학문과 구별되는 독특한 영역으로 내세울 수 있다.

둘째, 5개 영역별로 논문 비율을 검토한 결과, 문화콘텐츠학 전공자들은 사회와 문화와 같은 인문·사회과학 기반의 연구(9.1%)보다 자원 활용, 콘텐츠 제작과 같이 실용성을 목적으로 하는 연구(90.1%)를 압도적으로 많이 했던 것으로 나타났다.

셋째, 문화콘텐츠 학위논문의 연구 영역은 박사학위논문이 본격적으로 발표되기 시작한 2008년부터 2012년까지 사이에 인문·사회과학 기반의 '문화' 영역과 실용성 목적의 '콘텐츠' 영역을 중심으로 형성됐다. 2013년부터 2016년까지 사이에 '활용' 영역이 추가됐고 2016년부터 2020년까지 사이에 '사회', '문화', '활용', '콘텐츠' 4개 연구 영역으로 확대 발전되어 오늘날의 연구 지형도가 완성됐다.

본 논문의 가치는 지금까지 쌓여 있기만 했던 문화콘텐츠 학위논문의 데이터를 분석하여 연구 영역, 연구 주제, 연구 흐름을 도출했다는 것과 이를 통해 학문으로서 문화콘텐츠의 특징을 밝히는 일에 기초를 제공했다는 것이다.

[주제어] 문화콘텐츠, 연구 동향, 빅데이터 분석, 시맨틱 네트워크 분석, 토픽 모델링

투고일: 2021. 10. 5. 심사일: 2021. 10. 27. 게재 확정일: 2021. 11. 15.
https://doi.org/10.16937/jcp.2021.35.3.185

오정심_문화콘텐츠학 박사/주저자(ruaths0802@naver.com)

Ⅰ. 서론

1. 연구 목적 및 필요성

이 논문의 목적은 2000년부터 2020년까지 발표된 문화콘텐츠 학위논문의 서지정보와 초록을 수집하여, 빅데이터 분석방법을 통해 문화콘텐츠 학위논문 분야의 연구 주제, 연구 영역, 연구 흐름을 연구하는 것이다. 이러한 일을 통해 학문으로서 문화콘텐츠의 특징을 밝히는 일과 위상을 세우는 일에 이바지하는 것이다.

국내에서 문화콘텐츠학과는 2002년에 처음으로 생겼다. 2002년 3월에 한국외국어대학교 대학원에서 문화콘텐츠학과를 만들고, 신입생 11명을 뽑았다(임영상, 2017). 뒤이어 문화콘텐츠 산업이 크게 발달하자 국내 대학들은 문화콘텐츠학과를 앞다퉈 만들었다. 국내 고등교육기관에 개설된 문화콘텐츠학과의 수는 2004년에 1,124개, 2017년에 1,483개로 조사됐다.(KOCCA, 2017). 문화콘텐츠학과가 2002년에 처음으로 생겼던 점을 고려하면 그 수가 2004년까지 가파르게 늘어난 것이다.

그런데 문화콘텐츠가 무엇인지, 학문 체계를 어떻게 마련할지와 같은 고민을 먼저 하지 않고 학과부터 만들다 보니 적잖은 문제가 나타났다. 학교에서는 제대로 된 교육 자료가 부족해 인문학 등 기존 학과의 것을 가져와 썼고, 강의계획서는 강사 역량으로 마련됐다(정창권, 2007). 그리고 문화콘텐츠 용어의 개념과 뜻을 정리하지 않은 채 여기저기서 쓰다 보니 혼란이 생겼다. 같은 단어를 쓰고도 여러 가지 뜻으로 해석하는 일이 벌어진 것이다(박상천, 2007). 이러한 상황이 이어지다 보니 급기야 문화콘텐츠가 아닌 것이 없다는 말까지 나오게 됐다.

일부 학자들은 문화콘텐츠가 다른 학문과 구별되는 연구 대상과 이론들이 없기 때문에 학문으로 볼 수 없다고 말했다. 다른 쪽에서 그 주장을 반대하며 문화콘텐츠는 기존 학문 체계와 다르며 여러 분야를 아우르는 융합 학문이라고 말했다. 그러자 또 다른 쪽에서 여러 분야를 아우르는 것은 백과사전 지식이지 새로운 지식을 만드는 학문은 아니라고 말하면서 논란이 일었다(박치완・유제상, 2015). 전국에 1,000개가 넘는 문화콘텐츠학과가 이미 개설됐는데, 학문으로 볼 수 없다는 주장이 제기된 것이다.

이러한 논란 가운데 문화콘텐츠학과를 졸업한 신진 연구자들은 학문으로서 문화콘텐츠의 정체성을 밝히기 위해 노력을 기울였다. 관련 논문으로 태지호(2005) 등 10편이 있다. 하지만 대부분의 연구가 구체성이 부족하고 당위성을 강조하는 수준에서 내용이

그치고 말았다. 요즘에 와서 계량 방법을 통해 문화콘텐츠의 특징을 구체적으로 연구하는 논문들이 발표되고 있다(오정심, 2020b).

한편 신광철(2014)은 학문으로서 문화콘텐츠의 특징을 밝하기 위해서 이제까지 문화콘텐츠 연구 분야에 발표된 모든 논문들을 수집해 메타 분석해야 한다고 말했다. 여기서 메타 분석이란 문헌을 검토하여 얻은 정보를 기호로 바꾸고, 기술통계방법으로 분석하는 방법을 말한다. 그리고 일의 범위가 크기 때문에 혼자 할 수 없으며, 프로젝트로 진행해야 한다고 말했다.

그러나 아무리 많은 사람이 투입된다고 하더라도 메타 분석 방법만 가지고 원하는 결과를 얻을 수 없다. 사람이 수많은 내용을 빠짐없이 읽고 분석 항목에 따라 분류할 수 없기 때문이다. 이러한 한계 때문에 오늘날 연구자들 사이에서 내용 모두를 읽지 않고도 파악할 수 있는 '빅데이터 분석 방법'이 선호되고 있다.

빅데이터 분석이란 대규모 데이터에서 새로운 정보나 인사이트 따위를 얻어 이를 바탕으로 문제를 해결하는 방법을 말한다. 빅데이터 분석 방법의 장점은 첫째, 데이터 수가 아무리 많아도 짧은 시간 안에 분석할 수 있다는 것이다. 실제로 미국 스탠퍼트 대학교에서 빅데이터 분석 방법을 활용해 천문학적 숫자의 별 이미지를 연구하고 있다[1]. 둘째, 데이터에서 표본을 추출하지 않고 전체를 분석하기 때문에 데이터 누락과 같은 오류가 적다. 셋째, 기존 방법으로 분석하기 어려웠던 자연어 텍스트, 이미지와 같은 비정형 데이터도 분석할 수 있다(조성준, 2019). 이러한 이유로 문화콘텐츠 학위논문 자료를 연구하는 데 빅데이터 분석 방법을 활용할 것이다. 분석 결과를 살펴보기 전에 빅데이터 분석 방법부터 알아보자.

2. 방법론과 선행 연구 검토

1) 빅데이터 분석 방법 이해

빅데이터 분석은 아주 큰 규모의 데이터에서 새로운 정보와 인사이트 따위를 얻어 문제를 해결하는 방법을 말한다. 빅데이터 분석은 시각화, 연관성, 클러스터링, 예측과 같은 작업으로 이루어지며, 이를 구현하는 방법론에 통계학, 기계학습 등이 있다(조성준, 2019). 본 논문에서 기계학습 알고리즘에 기반을 둔 소프트웨어를 써서 빅데이터 분석을 했다. 기계학습(Machine Learning)은 인공 지능의 한 분야이며, 컴퓨터에게

1) 원문 출처:https://www.symmetrymagazine.org/article/studying-the-stars-with-machine-learning

빅데이터를 반복적으로 학습시키고, 귀납적 추론을 통해 명제를 도출하게 하는 기술이다(두산백과).

빅데이터 분석 방법을 활용한 연구는 크게 '데이터 수집', '데이터 전처리', '데이터 분석', '분석 결과 해석', '인사이트 도출'과 같은 단계로 진행된다. 여기서 가장 중요한 부분은 데이터 전처리이다. 데이터 전처리 정도에 따라 분석 결과의 질이 달라지기 때문이다. 데이터 전처리는 형태소 분석기와 같은 프로그램을 통해 비정형 구조의 데이터를 정형화된 구조로 바꾸는 작업이다. 컴퓨터는 논문, 인터넷 텍스트처럼 사람이 쓴 불규칙한 형태의 텍스트를 해석할 수 없다. 그래서 컴퓨터가 이러한 데이터를 해석할 수 있게 정형화된 구조로 바꿔주는 작업이 필요하다.

데이터 전처리 작업에서 시소러스(thesaurus) 사전 기능을 활용하면 분석 결과의 질을 한 번 더 높일 수 있다. 시소러스란 컴퓨터에 기억된 용어 사전을 뜻하며, 컴퓨터에게 문장이나 단어의 경계를 알려주는 것이다.[2] 한국어는 조사와 어미가 발달한 교착어이기 때문에 시소러스 사전 기능을 적용하지 않으면 추출되는 단어의 수가 기하급수적으로 늘어나고 연산의 비효율이 발생한다(이기창, 2019; 오정심, 2020b). 그래서 데이터 전처리 작업에서 시소러스 사전 기능을 적용해 불필요한 단어들을 미리 걸러낸다.

빅데이터 분석에 활용되는 기법에 여러 가지가 있다. 본 논문에서 LDA 토픽 모델링과 시맨틱 네트워크 분석 기법을 써서 시각화, 연관성, 클러스터링과 같은 작업을 했다. 토픽 모델링(Topic Modeling)은 말뭉치에서 유사성을 가진 단어들을 토픽 묶음으로 분류해주는 기법이다. 토픽 모델링의 알고리즘 가운데 가장 많이 쓰이는 LDA(Latent Dirichlet Allocation)는 특정 토픽에 단어들이 포함될 확률을 계산하여, 비슷한 확률로 나타난 단어들을 하나의 묶음 단위로 분류해 준다(윤효준 외, 2019; 이기창, 2019).

시맨틱 네트워크 분석(Semantic Network Analysis)은 비정형 텍스트에서 형태소 단위로 단어를 추출하고, 단어의 연결 관계로 네트워크를 형성하여 텍스트의 다양한 특징을 분석하는 기법이다. 이 방법을 활용해 텍스트 구조, 단어 사용 패턴, 중심 주제 등을 파악할 수 있다. 그리고 분석 결과를 네트워크 맵 따위로 그릴 수 있기 때문에 분석 결과를 직관적으로 파악할 수 있다(사이람, 2019).

---

2) 시소러스 사전에 유의어(synonym), 지정어(directive), 제외어(exception)가 있다. 유의어는 뜻이 비슷한 말을 하나로 통일해 주는 기능이다. 지정어는 고유 명사나 개체 명사 등 형태소 분절 없이 명사를 그대로 추출해 주는 기능이다. 제외어는 관사, 전치사 등 의미가 없는 불용어와 불필요한 단어들을 제거해 주는 기능이다(오정심, 2020a).

시맨틱 네트워크 분석 결과는 주요 지표를 통해서 해석한다. 본 논문에서 주로 활용한 주요 지표는 '문 동시 출현 빈도수', '연결 중심성', '위세 중심성'이다. 〈표 1〉은 주요 지표에 대한 설명 내용을 요약한 것이다. 이 지표를 선정한 이유는 본 논문은 텍스트의 핵심어, 중심 주제들을 분석하는 것을 목적으로 하기 때문이다. 연결 중심성과 위세 중심성은 네트워크에서 중요한 역할을 하는 노드를 알고 싶을 때 활용한다. 반면에, 매개 중심성과 근접 중심성은 노드를 세트 단위로 정의하고, 세트 사이의 관계를 알고 싶을 때 활용한다.

〈표 1〉 시맨틱 네트워크 분석 주요 지표

주요 지표	설명	해석 방향
문서 동시 출현 빈도수 (Frequency)	단어들이 일정 범위에 얼마나 함께 자주 등장했는지 횟수 계산	-저자들이 텍스트를 작성하면서 공통적으로 자주 썼던 말
연결 중심성 (Degree centrality)	네트워크에서 서로 붙어있는 이웃 노드들 개수	-텍스트 구성에 중요한 역할을 하는 단어는? (텍스트에서 중심성이 높은 단어를 제거하면 텍스트 구성이 어렵게 됨)
위세 중심성 (Eigenvector centrality)	노드의 연결 개수와 함께 영향력까지 계산 (연결 중심성 결과 보완)	

자료: 사이람, 2019 재구성.

2) 선행 연구 검토

요사이 빅데이터 분석 방법은 시장 동향이나 여론 동향과 같은 연구에 자주 활용되고 있다. 조작이 꽤 간단한 소프트웨어가 나오면서 생기는 일이다. 이제 공학 전문가가 아니더라도 소프트웨어를 쓸 수 있는 사람이라면 빅데이터 분석을 할 수 있게 됐다(조성준, 2019). 빅데이터 분석 방법을 활용해 문화콘텐츠의 연구 동향을 연구한 논문으로 황동열・황고은(2016) 등 4편이 있다.

이중에서 오정심 연구(2020)는 필자의 선행 연구이다. 본 논문은 후속 연구의 결과물로서 이전에 한계로 지적된 점을 보완하고, 데이터 수집 범위를 학위논문 분야로 확대해 분석한 것이다. 주의 깊게 살펴볼 또 다른 논문으로 민요한・김지영・박옥남의 연구(2021)가 있다. 오정심 연구(2020)와 같이 문화콘텐츠로 검색되는 학술논문의 데이터를 수집하고, 넷마이너(NetMiner) 소프트웨어를 사용해 네트워크 분석과 토픽 모델링을 했다. 2개의 선행 연구를 비교해 시사점을 도출하면 다음과 같다.

첫째, 두 연구는 데이터 수집 방법에서 차이가 있다. 오정심 연구(2020)는 한국학술

〈표 2〉 관련 선행 연구 현황

발행 월	논문 제목 (저자명)	주요 내용
2016. 12.	빅데이터 기술을 활용한 인문콘텐츠 분야의 의미연결망 분석 (황동열 · 황고은)	-2003년~2015년《인문콘텐츠》 게재 논문 510편 국문초록 수집 -의미 연결망 분석, R프로그램 패키지 사용
2020. 03.	인문콘텐츠 분야 연구의 경향 분석: 토픽모델링과 의미 연결망 분석을 중심으로 (황서이 · 박정배 · 김문기)	-2003년~2018년《인문콘텐츠》 게재 논문 622편 국문초록 수집 -의미 연결망 분석, 토픽모델링, R프로그램 패키지 사용
2020. 08.	빅데이터 토픽모델링 및 네트워크 분석을 통한 문화콘텐츠학 지식구조 연구 (오정심)	-2000년~2020년 4월 KCI학술지 논문 중에 "문화콘텐츠"로 검색되는 논문 3,685편 국문초록 수집 -텍스트 네트워크 분석, 토픽 모델링, NetMiner프로그램 사용
2021. 04.	〈토픽모델링과 키워드 네트워크 분석을 활용한 '문화콘텐츠' 연구 경향 분석 (민요한 · 김지영 · 박옥남)	-2002년~2019년 11월 KCI학술지 논문 중에 문화콘텐츠 검색 논문 1,693편 국문초록 수집 -키워드 네트워크 분석, 토픽 모델링, NetMiner프로그램 사용

지인용색인(www.kci.go.kr)에서 2000년부터 2020년 4월까지 "문화콘텐츠"로 검색되는 논문 3,685편의 초록과 서지정보를 수집해 분석했다. 그러나 민요한 · 김지영 · 박옥남의 연구(2021)는 같은 사이트에서 2002년부터 2019년 11월까지 큰 따옴표("") 없이 문화콘텐츠로 검색되는 논문 1,693편의 데이터를 수집해 분석했다. 일반적으로 큰 따옴표("") 안에 키워드를 넣어 검색하면 논문 제목, 주제어, 초록 따위에 키워드가 반드시 포함된 자료만 나타난다.

둘째, 시기 설정 방법에서 차이가 있다. 오정심 연구(2020)는 대통령 임기에 따라 시기를 1시기(2000~2007), 2시기(2008~2012), 3시기(2013~2016), 4시기(2017~2020)로 구분했다. 이렇게 시기를 나눈 까닭을 문화콘텐츠가 정책 필요에 따라 만든 개념이고, 정부 주도 아래 성장했기 때문에 관련 이슈도 달라졌을 것이라고 설명했다. 그러나 민요한 · 김지영 · 박옥남 연구(2021)는 시기를 임의대로 5년씩 나누고, 데이터 수의 균형을 맞추기 위해 첫 번째 기간만 8년으로 늘렸다고 설명했다.

시기를 구분하는 방법에는 여러 가지가 있을 수 있다. 민요한 · 김지영 · 박옥남 연구(2021)처럼 임의로 5년 단위 또는 10년 단위로 나눌 수 있고, 오정심 연구(2020)처럼 가설을 세워 나눌 수도 있다. 본 논문에서 몇 가지 가설을 세우고, 가설에 따라 데이터를 분석한 후에 의미 있는 결과가 나온 방식을 최종 채택하기로 한다.

셋째, 데이터 정제 방법에서 차이가 있다. 오정심 연구(2020)는 검색어인 '문화콘텐츠'를 분석 대상에 함께 넣었으나, 민요한 · 김지영 · 박옥남 연구(2021)는 넣지 않았

다. 일반적으로 검색어는 TF-IDF 값이 적게 나타나기 때문에 분석 대상에 넣지 않는다. TF-IDF 값이 적게 나타나면 중요도가 낮다고 판단하기 때문이다[3]. 이에 대해 오정심 연구(2020)는 학술논문 저자들이 문화콘텐츠 용어의 개념과 뜻을 어떻게 쓰는지 알아볼 필요가 있어 분석 대상에 넣었다고 설명했다.

넷째, 두 연구는 분석 결과에서도 차이가 있다. 오정심 연구(2020)는 문화콘텐츠 학술논문의 연구 영역을 '한국 사회', '활용', '장르', '기술', '산업', '이론 체계' 6개로 나누고, 토픽 모델링으로 분류한 40개 주제를 영역별로 분류했다. 그러나 민요한·김지영·박옥남 연구(2021)는 연구 영역을 '스토리', '개발', '교육', '한국', '지역', '전통'과 같이 6개 항목으로 나눴으며, 하위 주제를 제시하지 않았다. 그런데 민요한·김지영·박옥남 연구(2021)에서 제시한 '스토리', '개발', '교육', '지역'은 오정심 연구(2020)에서 '활용' 영역의 하위 주제로 분류된다.

3. 연구 대상 및 방법

본 논문의 목적은 빅데이터 분석 방법을 활용해 문화콘텐츠 학위논문 분야의 연구 주제, 연구 영역, 연구 흐름 따위를 분석하고, 이를 바탕으로 다른 학문과 구별되는 특징이 무엇인지 살펴보는 것이다. 이러한 목적을 이루기 위해 2004년부터 2020년까지 발표된 문화콘텐츠 석·박사 학위논문의 국문초록과 서지정보 모두를 분석했다. 초록을 수집해 분석한 까닭은 문화콘텐츠 전공자가 문화콘텐츠 분야에서 중요하게 다루는 문제를 연구하고, 그 결과물을 요약해 쓴 글이기 때문이다. 수집한 데이터를 시맨틱 네트워크 분석과 토픽 모델링 기법을 통해 시각화, 연관성, 클러스터링과 같은 분석을 하여 연구 주제, 연구 영역, 연구 흐름들을 도출했다. 그리고 구체적인 연구 결과를 얻기 위해 5가지 문제를 가지고 분석을 진행했다.

첫째, 저자들이 문화콘텐츠 학위논문을 작성하면서 공통적으로 가장 자주 썼던 단어는 무엇인가?

둘째, 2004년부터 2020년까지 문화콘텐츠 학위논문의 주요 주제는 무엇인가?

3) Term Frequency-Inverse Network(TF-IDF, 단어의 문서 내 중요도)는 단어 빈도수와 문서 빈도수의 역수를 곱한 값이다. 중요하게 쓰는 표현들은 TF-IDF가 높게 나타나고 관용적으로 쓰는 표현들은 낮게 나타난다(사이람, 2019).

셋째, 문화콘텐츠 학위논문의 핵심 연구 영역은 무엇인가?
셋째, 주요 연구 영역과 주제는 시기별로 어떻게 변했나?
넷째, 위 결과를 단어 클라우드와 지식 지도 형태로 그리면 어떤 특징이 발견되는가?
다섯째, 위 결과를 바탕으로 '문화콘텐츠 학위논문의 주요 연구 영역과 주제'를 분류하면 어떤 특징이 발견되는가?

이러한 연구 목적과 문제를 가지고 연구를 진행했으며, 선행 연구에서 얻은 시사점을 참고해 연구 방법을 보완했다. 첫째, 데이터 수집은 학술연구 정보 서비스(www.riss.kr)에서 했다. 수집 기간을 2004년부터 2020년 12월까지로 했다. 시작 기간을 2004년으로 설정한 이유는 2004년에 문화콘텐츠학 석사학위 논문이 처음으로 발표됐기 때문이다. 그리고 문화콘텐츠학을 전공한 학생의 학위논문만 수집하기 위해서 상세검색 부분에서 학위논문 항목을 선택한 후에 '학과 정보'가 문화콘텐츠로 검색되는 자료만 수집했다. 이렇게 해서 조사된 자료들 중에서 초록이 없는 것, 중복된 것을 제외하고 총 1,812편 논문의 국문초록과 서지정보를 분석 대상으로 선정했다.

둘째, 시기를 나누기 전에 가설을 세웠다. 가설에 따라 분석을 하고 의미 있는 결과가 나온 방식을 최종 선택했다. 가설을 '정권의 정책이 학계 연구에 영향을 미치기까지 약 2년 정도 걸릴 것'과 '문화콘텐츠 연구는 정권의 정책과 사회 이슈의 변화에 민감하게 반응할 것'으로 세웠다.

셋째, 분석 대상에 검색어인 '문화콘텐츠'를 넣었다. 저자들이 문화콘텐츠 학위논문을 작성할 때 문화콘텐츠 용어의 개념과 뜻을 어떻게 쓰는지 살펴볼 필요가 있기 때문이다.

넷째, 데이터를 분석하는 것에서 그치지 않고 분석 결과를 활용해 '문화콘텐츠 학위논문의 주요 연구 영역과 주제 분류 방안'을 제시했다.

〈표 3〉과 〈표 4〉는 데이터 수집 방법과 분석 방법의 내용을 요약한 것이다. 연구는 크게 4단계 '데이터 수집', '데이터 전처리', '데이터 분석', '종합 및 해석' 으로 나눠 차례대로 진행했다. 데이터 수집과 분석에 넷마이너(NetMiner 4.3) 소프트웨어를 사용했다.

〈표 3〉 데이터 수집 방법

구분	내용
수집 대상	문화콘텐츠 석·박사 학위논문 국문초록과 서지정보 (문화콘텐츠학 전공 학생의 학위논문만 수집)
수집처	학술연구 정보 서비스(www.riss.kr)
검색어	문화콘텐츠 (상세 검색 학과정보에서 문화콘텐츠로 검색)
수집기간	2004년 ~ 2020년
1차 필터링	국문초록 없는 것, 중복된 것 제외
▼	
최종 연구 대상	문화콘텐츠 석·박사 학위논문 1,812편의 국문초록과 서지정보

〈표 4〉 데이터 분석 방법

구분	주요 내용		비고
데이터 전처리	-오탈자 검토, 영어와 한자 변환 -시소러스 사전 적용, 데이터 필터링 -넷마이너 한국어 형태소 분석기에 데이터 입력 (비정형 텍스트 해체→ 정형화 구조의 명사형 형태소 추출)		넷마이너 한국어 형태소 분석기
데이터 분석	시맨틱 네트워크 분석	-네트워크 형성 -문서 동시 등장 빈도수 분석 -연결중심성, 위세중심성 분석 -쿼리(Query)로 시기별 데이터 분류 -시기별 데이터 분석 결과 비교 -분석 결과 시각화 등	넷마이너 (NetMiner 4.3, Topic Modeling Plug-In)
	토픽 모델링	-토픽 수 결정 -LDA 토픽 모델링 -클러스터링 결과 확인 등	

Ⅱ. 본론

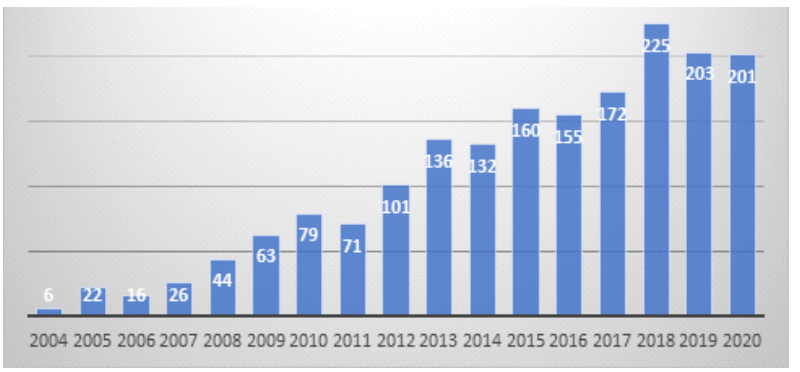
1. 기초 통계 결과

문화콘텐츠 학위논문 1,812편의 기초 통계 결과부터 살펴보자. [그림 1]과 〈표 5〉에서 보듯이 문화콘텐츠 학위논문은 2004년에 처음으로 6편이 발표됐다. 6편 모두는 석사학위논문이었다.[4] 박사학위논문은 2006년에 처음으로 1편 발표됐다. 2006년부터

4) 2004년에 처음으로 발표된 문화콘텐츠학 석사 학위논문의 제목은 다음과 같다.
-《문화원형의 매체를 통한 콘텐츠化 연구 : 브램 스토커의 드라큘라를 중심으로(조정연)》

2008년까지 문화콘텐츠 석·박사 학위논문은 50편 안팎으로 발표됐고 2009년부터 60편 이상씩 발표됐다. 2012년에 처음으로 100편 이상 발표됐는데, 2012년은 국내에서 문화콘텐츠학과가 처음으로 생긴지 10년째 되던 해였다. 그 뒤로 100편 이상씩 꾸준히 발표되다가 2018년에 처음으로 200편 이상 발표됐으며 현재까지 200편 이상씩 발표되고 있다.

[그림 1] 연도별 문화콘텐츠 학위논문 발표 현황(2004년~2020년)



〈표 5〉 연도별 문화콘텐츠 학위논문 발표 현황(2004년~2020년)

발표 연도	석사 학위논문 수(편)	박사 학위논문 수(편)	합계(편)
2004	6	0	6
2005	22	0	22
2006	15	1	16
2007	24	2	26
2008	42	2	44
2009	55	8	63
2010	73	6	79
2011	60	11	71
2012	84	17	101

〈표 계속〉

- 《무형문화재를 소재로 한 디지털 영상콘텐츠 개발에 관한 연구(오정심)》
- 《e-Learning을 이용한 외국어 교육 연구 : 태국어 교육 사례 비교와 분석(남경민)》
- 《한국문화콘텐츠 정체성 확립과 지역협력에 관한 연구(이승현)》
- 《인터넷 신조어를 통해서 본 문화적 트렌드 연구(표남숙)》
- 《Dynamic pricing에서 통제력 착각과 간접적 교환 관계가 가격 공정성 인식에 미치는 영향(이지혜)》

발표 연도	석사 학위논문 수(편)	박사 학위논문 수(편)	합계(편)
2013	119	17	136
2014	112	20	132
2015	134	26	160
2016	130	25	155
2017	138	34	172
2018	186	39	225
2019	155	48	203
2020	154	47	201
합계	1,509	303	1,812

학교별로 문화콘텐츠 학위논문의 발표 현황을 살펴보자. 문화콘텐츠 학위논문을 1편 이상이라도 발표한 학교는 51개로 조사됐다. 이중에서 10편 이상을 발표한 학교는 24개, 20편 이상을 발표한 학교는 19개로 나타났다. 〈표 6〉은 20편 이상을 발표한 학교의 목록이다. 〈표 6〉에서 보듯이 문화콘텐츠 학위논문을 가장 많이 발표한 학교는 건국대학교였다(329편). 이어서 한국외국어대학교(277편), 중앙대학교(197편), 한양대학교(163편), 동국대학교(123편) 순으로 문화콘텐츠 학위논문을 많이 발표했다.

〈표 6〉 학교별 문화콘텐츠 학위논문 발표 현황(2004년~2020년)

학교명	발표 논문 수(편)	학교명	발표 논문 수(편)
건국대학교	329	성균관대학교	44
한국외국어대학교	277	숙명여자대학교	36
중앙대학교	197	한국방송통신대학교	32
한양대학교	163	성신여자대학교	30
동국대학교	123	가톨릭대학교	28
고려대학교	96	청주대학교	28
연세대학교	58	경희대학교	25
동방문화대학원대학교	49	경기대학교	22
홍익대학교	47	경상대학교	20
안동대학교	46	-	-
합계		1,812	

2. 시맨틱 네트워크 분석 결과

문화콘텐츠 학위논문 1,812편의 국문초록을 넷마이너 형태소 분석기를 통해 전처리 작업하여 명사형 형태소 15,242개를 추출했다. 이 단어들을 네트워크로 연결하여 '문서 동시 출현 빈도수', '연결 중심성', '위세 중심성' 3개 지표로 분석했다. 지금부터 분석 결과와 그 의미를 살펴보자.

문서 동시 출현 빈도수(co-occurrence Frequency)는 단어들이 일정한 범위에 얼마나 함께 자주 등장했는지를 계산한 값이다. 빈도수가 크게 나온 단어들은 저자들이 관련 텍스트를 작성하면서 공통적으로 자주 썼던 말이라고 해석한다(오정심, 2020b).

분석 결과, 15,242개 단어들 중에서 문서 동시 출현 빈도수가 가장 크게 나온 단어는 '문화(918)'였다. 이어서 '활용(811회)', '콘텐츠(731회)', '한국(693회)', '사회(691회)', '가치(602회)', '방법(592회)', '사람(549회)', '산업(497회)' 순으로 값이 크게 나왔다. 문화콘텐츠학 전공자들은 학위논문을 작성하면서 문화, 활용, 콘텐츠, 한국, 사회, 가치, 방법, 산업과 같은 단어들을 가장 많이 썼던 것으로 나타났다. [그림 2]는 분석 결과를 직관적으로 알아볼 수 있게 단어 클라우드 맵으로 나타낸 것이다.

[그림 2] 단어 클라우드 (문서 동시 출현 빈도수 분석 결과(2004년~2020년))

연결 중심성(degree centrality)은 네트워크에서 서로 연결된 이웃 노드들의 개수가 얼마인지 계산한 값이다. 연결 중심성이 크게 나온 단어들은 다른 단어들과 많이 연결돼 있으며 텍스트에서 중요한 단어라고 해석한다. 만약 연결 중심성이 크게 나온 단어를 텍스트에서 없애면 텍스트 구성이 어렵게 된다(오정심, 2020b). 연결 중심성 분석 결과, '문화(0.362390)', '활용(0.337906)', '콘텐츠(0.325794)', '사회(0.316579)', '가치 (0.312778)', '한국(0.299538)' 순으로 값이 크게 나타났다.

위세 중심성(eigenvector centrality)은 노드들의 연결 개수와 함께 영향력까지 계

산한 값이다. 위세 중심성이 크게 나타난 단어들은 텍스트 구성에 중요한 역할을 하는 단어라고 해석한다(오정심, 2020b). 위세 중심성 분석 결과, '문화(0.164130)', '활용(0.152315)', '콘텐츠(0.146936)', '사회(0.14332)', '가치(0.14332)', '한국(0.136254)' 순으로 연결 중심성 결과와 비슷하게 나타났다.

앞에서 말했듯이 문서 동시 출현 빈도수가 크게 나온 단어는 해당 분야의 저자들이 텍스트를 작성하면서 공통적으로 자주 썼던 말로 해석한다. 그리고 연결 중심성과 위세 중심성이 크게 나온 단어는 텍스트 구성에 중요한 역할을 하는 단어로 해석한다. 이를 근거로 하여 3개 지표 값이 모두 크게 나온 노드는 해당 연구 분야에서 중요하게 다뤄진 대상(이하 중심 연구 대상)으로 이해할 수 있다.

문화콘텐츠 학위논문 분야의 중심 연구 대상을 알아보기 위해 3개 지표 값이 모두 크게 나온 노드를 조사했다. 먼저 데이터 분석 결과 값에서 노드별로 순위를 매기고 상위 70위 안에 있는 노드를 추려서 〈표 7〉과 같이 정리했다. 그리고 3개 지표 값이 모두 크게 나온 노드를 검토했다.

〈표 7〉 문서 동시 등장 빈도수·연결 중심성·위세 중심성 분석 결과(2004년~2020년)

문서 동시 등장 빈도수		연결 중심성		위세 중심성	
단어	측정 값	단어	측정 값	단어	측정 값
문화	918	**문화**	0.362390	**문화**	0.164130
활용	811	**활용**	0.337906	**활용**	0.152315
콘텐츠	731	**콘텐츠**	0.325794	**콘텐츠**	0.146936
한국	693	**사회**	0.316579	**사회**	0.143332
사회	691	**가치**	0.312778	**가치**	0.142035
가치	602	**한국**	0.299538	**한국**	0.136254
방법	592	**방법**	0.281499	**방법**	0.127497
사람	549	**사람**	0.279453	**사람**	0.126614
산업	497	**산업**	0.276924	**산업**	0.125520
시대	471	다양	0.272140	다양	0.123427
방식	464	시대	0.268125	시대	0.121987
인식	449	방식	0.267856	방식	0.121317
다양	447	개념	0.265942	개념	0.120661
제작	434	기술	0.260284	기술	0.117751
개념	433	인식	0.258898	인식	0.117347
지역	425	예술	0.254158	예술	0.115686
이론	424	제작	0.253536	역사	0.114827
기술	424	환경	0.251204	제작	0.114759

〈표 계속〉

문서 동시 등장 빈도수		연결 중심성		위세 중심성	
단어	측정 값	단어	측정 값	단어	측정 값
지속	422	역사	0.249717	환경	0.113773
예술	420	지속	0.24926	지속	0.113229
시장	414	공간	0.247569	지역	0.112475
환경	407	시장	0.247330	공간	0.112275
역사	405	세계	0.245436	세계	0.112004
전략	390	지역	0.245026	시장	0.111672
국가	375	이론	0.243477	문화콘텐츠	0.111581
세계	374	문화콘텐츠	0.242756	이론	0.110347
시작	371	전략	0.240946	국가	0.109914
상황	371	국가	0.238485	경제	0.109055
중국	369	형태	0.237905	전략	0.108862
형태	366	경제	0.237647	형태	0.108144
공간	362	시작	0.236329	시작	0.107769
형성	361	대중	0.236225	대중	0.107487
문화콘텐츠	361	형성	0.235769	형성	0.107208
자료	360	상황	0.234941	상황	0.106626
경험	358	전통	0.230782	전통	0.105578
정보	353	이미지	0.226738	현대	0.103879
프로그램	349	현대	0.226537	이미지	0.102685
전통	349	자료	0.223404	자료	0.101258
경제	349	경험	0.223275	기능	0.101160
법제도	347	기능	0.222752	경험	0.100778
영화	344	성장	0.221517	성장	0.100316
성장	329	정보	0.220631	성공	0.099796
대중	328	성공	0.218795	정보	0.099371
작품	326	매체	0.218642	매체	0.099084
이미지	326	법제도	0.216372	법제도	0.098986
성공	325	존재	0.216227	존재	0.098866
교육	325	장르	0.214745	장르	0.097562
현대	314	작품	0.214632	작품	0.097518
활성	312	활동	0.212638	활동	0.096721
활동	312	기획	0.210342	기획	0.096007

〈표 계속〉

문서 동시 등장 빈도수		연결 중심성		위세 중심성	
단어	측정 값	단어	측정 값	단어	측정 값
시간	309	활성	0.210166	활성	0.095989
존재	307	디지털	0.209008	중국	0.095964
참여	303	프로그램	0.208918	현상	0.095335
유형	303	교육	0.208601	교육	0.095289
기능	291	현상	0.208456	프로그램	0.095272
장르	289	중국	0.208287	디지털	0.095008
선정	284	영화	0.207239	참여	0.094279
등장	278	참여	0.207157	영화	0.094229
기획	275	등장	0.206451	등장	0.093402
매체	272	소비	0.203644	융복합	0.092345
미디어	255	시간	0.203294	시간	0.092233
현상	254	융복합	0.202736	소비	0.092210
생활	252	유형	0.202378	생산	0.092129
개인	251	미디어	0.202033	유형	0.091636
시각	248	생산	0.200897	미디어	0.091621
디지털	247	개인	0.199768	소통	0.090737
소재	241	선정	0.199430	선정	0.090689
인터넷	238	소통	0.198572	개인	0.090495
이야기	237	소재	0.196652	소재	0.089547
생산	235	상품	0.196119	상품	0.089464

※문서 동시 출현 빈도수, 연결 중심성, 위세 중심성이 모두 크게 나타난 단어들을 밑줄로 표시하였다.

조사 결과, 3개 지표 값이 모두 크게 나온 노드는 '문화', '활용', '콘텐츠', '사회', '가치', '한국', '방법', '사람', '산업'이었다. 이것은 문화콘텐츠 학위논문 분야에서 중요하게 다뤘던 대상으로 이해할 수 있다.

한편 넷마이너 프로그램에는 분석 결과를 그림으로 나타낼 수 있는 여러 가지 기능이 있다. 그중에서 PFNet은 네트워크의 링크 수가 너무 많아서 핵심 노드의 연결 관계가 보이지 않을 때 활용하면 쓸모가 있다(사이람, 2019). 문화콘텐츠 학위논문 분야의 중심 연구 대상으로 도출된 '문화', '활용', '콘텐츠', '사회', '가치', '한국', '방법', '사람', '산업'이 다른 대상과 어떻게 연결되어 연구 주제와 영역을 이루는지 살펴보기 위해서 PFNet 네트워크 맵을 그렸다.

본 논문에서 제시한 PFNet 네트워크 맵은 일종의 연구 지형도로 보고 해석할 수 있다. 지도에서 연구 대상과 연구 영역, 이들의 관계 따위를 알 수 있고 대상들이 어떻게 전체를 이루고 있는지를 한눈에 파악할 수 있다. 그림을 해석하는 방법을 간단히 설명하면, 그림에서 노드의 크기는 빈도수를 나타낸 것이다. 노드의 크기가 클수록 문화콘텐츠 학위논문 분야에서 논의가 많이 됐다는 뜻이다. 그리고 연결선 굵기는 노드 사이의 연결 강도를 나타낸 것이다. 연결선 굵기가 굵을수록 연결 강도가 세다는 뜻이다. 연결선 굵기가 굵은 노드들을 연결하면 연구 주제를 알 수 있다.

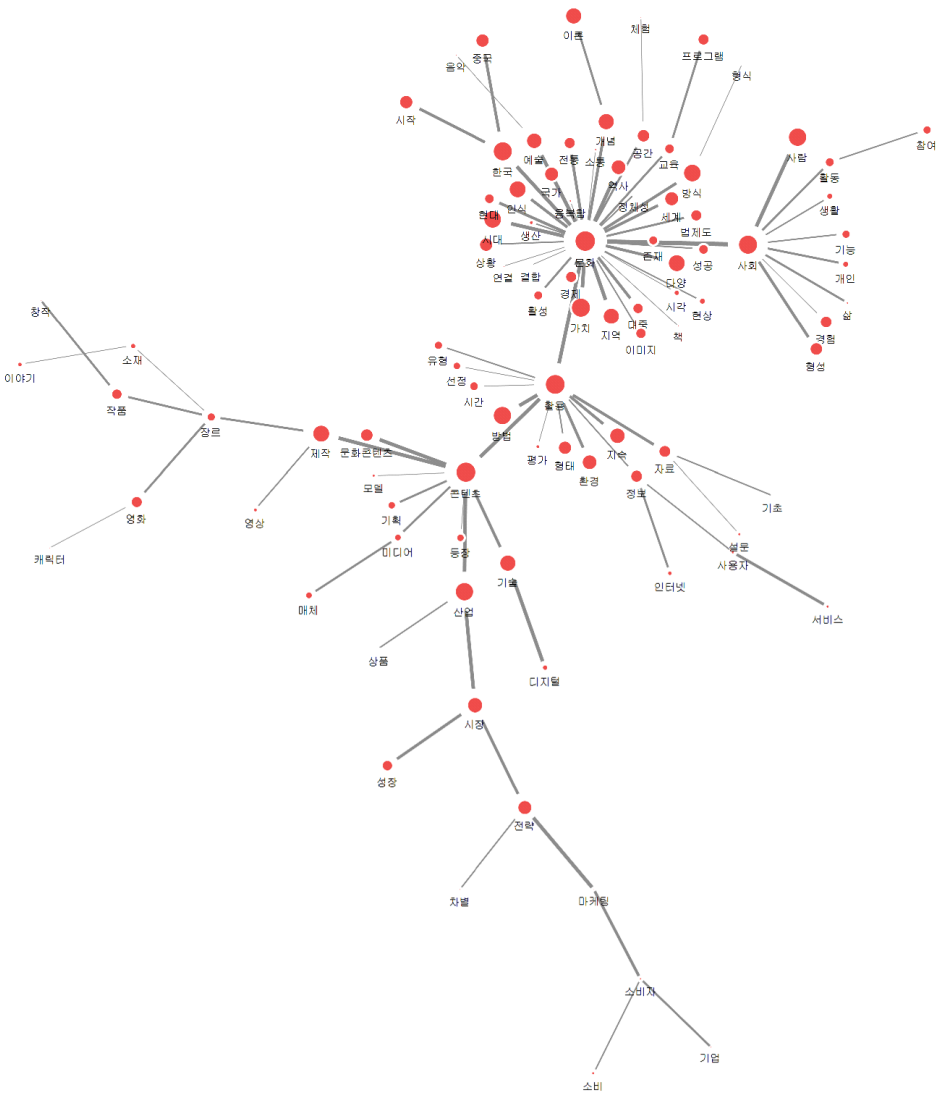
지금부터 [그림 3]을 보면서 2004년부터 2020년까지 문화콘텐츠 학위논문의 연구 지형도의 특징을 살펴보자. 가장 두드러지게 나타난 특징은 전체 지도가 4개 하위 연결망으로 구분된다는 것이다. 그리고 4개의 하위 연결망은 사회, 문화, 활용, 콘텐츠 노드를 중심으로 형성돼 있고 하위 연결망의 관계가 사회와 문화 사이에, 문화와 활용 사이에, 활용과 콘텐츠 사이에 형성돼 있다는 것이다. 이 내용을 풀어서 설명하면, 문화콘텐츠 학위논문의 연구 영역은 크게 '사회 영역', '문화 영역', '활용 영역', '콘텐츠 영역'으로 구성돼 있다. 영역 사이의 관계는 사회와 문화를 연구하고 그 결과물을 활용하고 콘텐츠로 제작하는 쪽으로 형성돼 있다.

4개의 연구 영역을 중심으로 연구 주제를 살펴보면 '사회 영역'에 사회와 사람, 사회 형성, 사회 경험, 사회 활동 참여와 같은 주제가 있다. '문화 영역'에 한국 문화, 세계 문화, 전통 문화 가치, 현대 문화, 문화 다양성, 문화 개념 및 이론과 같은 주제가 있다. '활용 영역'에 활용 방법, 정보 활용, 자료 활용, 활용 형태와 같은 주제가 있다. '콘텐츠 영역'에 디지털 기술, 미디어 매체, 제작 장르, 콘텐츠 기획, 산업 시장과 같은 주제가 있다.

'활용 영역'과 '콘텐츠 영역'을 구분하는 것이 모호했는데 연구 주제를 살펴보면서 분명해졌다. 정보와 자료 따위를 활용하거나 그 방법을 연구하는 것이 활용 영역이고 디지털 기술, 상품 기획, 시장 판매 전략 따위와 관련 있는 것이 콘텐츠 영역이다.

한편 시맨틱 네트워크 분석 방법은 텍스트에서 중요한 역할을 하는 단어들을 추출하고 단어의 연결 관계를 그림으로 나타내는 데 장점이 있지만 전체 문서의 내용을 파악하는 데 한계가 있다. 그래서 이러한 단점을 보완하기 위해 토픽 모델링을 함께 쓴다. 전체 문서에 어떤 토픽들이 있고 그 토픽을 표현하기 위해 어떤 단어들이 쓰였는지 알고 싶을 때 토픽 모델링을 활용하면 쓸모가 있다(오정심, 2020b).

[그림 3] PFNet 네트워크 맵(2004년~2020년)



3. 토픽 모델링 결과

문화콘텐츠 학위논문 1,812편의 국문초록을 LDA 토픽 모델링했다. 분석 결과, 40개 토픽이 추출됐다[5]. 2004년부터 2020년까지 발표된 문화콘텐츠 학위논문 1,812편

5) 샘플링 횟수를 1,000회로 설정하여 토픽 수를 2부터 50까지 변경하면서 토픽 모델링을 실시한 결과, 토픽 수를 40으로 했을 때 키워드 중복이 가장 적으면서 해석이 쉬었다.

을 40개 주제로 분류한 것이다. 토픽 모델링 결과를 〈표 8〉과 같이 정리했다. 보통 토픽명을 키워드1과 키워드2를 중심으로 정하며 연구자의 역량에 따라 바뀔 수 있다.

분석 결과, 목록에서 첫 번째 키워드 묶음이 '영화', '관객', '청소년', '다양'으로 나타났으며 토픽명을 '콘텐츠와 관객'으로 정했다. 관련 논문으로 86편(4.7%)이 있는 것으로 분석됐으며,[6] 논문 사례로 《4D 영화 관객의 만족에 영향을 미치는 요인(2015)》가 있다.

〈표 8〉 토픽 모델링 결과(2004년~2020년)

토픽명	키워드1	키워드2	키워드3	키워드4	키워드5	관련 논문 수(비율)
콘텐츠와 관객	영화	관객	청소년	다양	홍콩	86 (4.7%)
게임 콘텐츠	게임	플랫폼	사용자	모바일	온라인	79 (4.4%)
지역 역사 문화 자원 활용	지역	역사	문화	자원	활용	72 (4.0%)
교육 콘텐츠	교육	학습	학교	학생	프로그램	72 (4.0%)
서비스 (만족도) 조사	서비스	만족	의도	만족도	설문	71 (3.9%)
스토리텔링	드라마	스토리텔링	스토리	이야기	원작	69 (3.8%)
한류 시장	중국	한국	한류	국가	시장	64 (3.5%)
TV 방송 프로그램	프로그램	TV	방송	시청자	예능	62 (3.4%)
공간 콘텐츠	공간	박물관	체험	전시	융복합	58 (3.2%)
뮤지컬 작품	뮤지컬	작품	창작	장르	관객	54 (3.0%)
산업 시장 성장	산업	시장	성장	가치	모델	54 (3.0%)
공연예술 및 페스티벌	공연	페스티벌	전통	예술	극장	52 (2.9%)
미디어 및 매체	미디어	인터넷	뉴스	매체	보도	50 (2.8%)
디지털 기술	정보	디지털	저작	기술	사용자	49 (2.7%)
대중음악	음악	대중	창작	소재	시대	47 (2.6%)
사회 문화 프로그램	사회	노인	심리	프로그램	삶	46 (2.5%)
영상 제작 기술	제작	영상	테마파크	기술	방식	46 (2.5%)
상품 광고	광고	상품	애니메이션	태도	소비자	46 (2.5%)
브랜드 이미지	브랜드	이미지	책	인식	가치	44 (2.4%)
국가 문화 교류	문화	국가	교류	국제	민족	44 (2.4%)
문화유산 활용 및 관광	관광	문화유산	가치	무형	활용	44 (2.4%)
웹툰 및 출판만화	만화	웹툰	출판	콘텐츠	활용	43 (2.4%)

〈표 계속〉

6) 토픽 모델링은 전체 문서를 분석해 토픽(단어 묶음)으로 분류해 주는 방식이기 때문에 결과 목록에 논문 비율도 함께 계산돼 나온다.

토픽명	키워드1	키워드2	키워드3	키워드4	키워드5	관련 논문 수(비율)
아동 놀이 및 문학	아동	놀이	문학	작품	서예	43 (2.4%)
기업 마케팅 전략	전략	기업	마케팅	활동	소비자	41 (2.3%)
도시 및 사회 재생	사회	도시	경제	재생	여행	41 (2.3%)
세계 문화콘텐츠 산업	문화콘텐츠	세계	사람	소통	문화산업	39 (2.2%)
여성 및 가족 콘텐츠	여성	가족	일본	작가	사랑	39 (2.2%)
조선시대 차 문화, 생활 문화	차	조선	시대	생활	불교	39 (2.2%)
문화예술 지원 사업	예술	문화	지원	기관	사업	37 (2.0%)
전통 및 현대 음식문화	전통	현대	음식	사상	미	36 (2.0%)
신화 서사와 인물	서사	신화	인물	재현	영웅	34 (1.9%)
주민 참여 및 체험	참여	주민	체험	마을	농촌	31 (1.7%)
SNS 커뮤니케이션 시장	커뮤니케이션	시장	메시지	SNS	소셜미디어	30 (1.7%)
캐릭터 서사 및 정체성	캐릭터	주체	서사	정체성	성격	29 (1.6%)
문화원형 활용	음악	감정	춤	원형	문화원형	25 (1.4%)
법제도 도입	법제도	전문	방법	도입	자료	21 (1.2%)
수용자 경험 이론	수용자	유형	경험	이론	매체	21 (1.2%)
조직 역량 평가 활용	평가	활용	조직	디자인	역량	19 (1.0%)
다큐멘터리 연구	언어	시간	다큐멘터리	실험	사건	19 (1.0%)
시각 콘텐츠	미술	시각	소비	시대	가치	16 (0.9%)

두 번째 키워드 묶음이 '게임', '플랫폼', '사용자', '모바일', '온라인'으로 나타났으며 토픽명을 '게임 콘텐츠'으로 정했다. 관련 논문으로 79편(4.4%)이 있는 것으로 분석됐으며, 논문 사례로 《크로스 플랫폼 게임의 발전 방향에 대한 연구(2014)》가 있다.

세 번째 키워드 묶음이 '지역', '역사', '문화', '자원', '활용'으로 나타났으며, 토픽명을 '지역 역사 문화 자원 활용'으로 정했다. 관련 논문으로 72편(4.0%)이 있는 것으로 분석됐으며, 논문 사례로 《향토자원의 브랜드화 과정 연구(2012)》가 있다.

네 번째 키워드 묶음이 '교육', '학습', '학교', '학생', '프로그램'으로 나타났으며, 토픽명을 '교육 콘텐츠'으로 정했다. 관련 논문으로 72편(4.0%)이 있는 것으로 분석됐으며, 논문 사례로 《주한미군의 문화교육을 위한 콘텐츠 개발방안 연구(2011)》가 있다.

다섯 번째 키워드 묶음이 '서비스', '만족', '의도', '만족도', '설문'으로 나타났으며, 토픽명을 '서비스 (만족도) 조사'로 정했다. 관련 논문으로 71편(3.9%)이 있는 것으로 분석됐으며, 논문 사례로 《소셜미디어 이용 특성과 리조트 선택 및 여행만족의 관계

(2011)》가 있다. 이런 방식으로 토픽명을 정했으며, 나머지 토픽명과 논문 비율을 〈표 8〉에서 살펴보자.

4. 주요 연구 영역 및 주제 분류

지금까지 시맨틱 네트워크 분석과 토픽 모델링 결과를 살펴봤다. 분석 결과를 살펴보는 것에서 한걸음 더 나아가 분석 자료를 활용해 '문화콘텐츠 학위논문 분야의 주요 연구 영역과 연구 주제 분류 방안'을 제시해 보고자 한다.

2장에서 살펴본 시맨틱 네트워크 분석 결과를 기준으로 토픽 모델링 결과를 참고해 문화콘텐츠 학위논문의 주요 연구 영역을 '사회와 문화', '문화콘텐츠 활용', '문화콘텐츠 장르', '문화콘텐츠 제작 기술', '문화콘텐츠 산업' 5개로 구분했다. 그리고 토픽 모델링을 통해 도출된 40개 주제를 5개 영역에 맞게 분류했다. 〈표9〉는 그 내용을 정리한 것이다.

〈표 9〉에서 5개 영역별로 분류된 논문의 비율을 살펴보자. 문화콘텐츠 학위논문 분야에서 논문이 가장 많이 발표된 영역은 문화콘텐츠 장르(32.2%)였다. 이어서 문화콘텐츠 산업(31.6%), 활용(19.0%), 사회와 문화(9.1%), 문화콘텐츠 제작 기술(8.0%) 순이었다. 논문 비율을 통해 알 수 있는 사실은 문화콘텐츠 학위논문 저자들은 사회와 문화와 같은 인문학 기반의 연구(9.1%)보다 자원 활용, 콘텐츠 제작과 상품화 전략과 같이 실용성 목적의 연구(90.9%)를 압도적으로 많이 했다는 것이다.

필자는 선행 연구에서 문화콘텐츠 학술논문 3,685편을 분석해 학술논문의 주요 연구 영역을 '한국 사회와 문화콘텐츠', '문화콘텐츠 활용', '문화콘텐츠 장르', '문화콘텐츠 기술', '문화콘텐츠 산업', '문화콘텐츠 이론 체계' 6개로 도출했다. 선행 연구와 본 논문을 비교하면 연구 영역에서 '문화콘텐츠 이론 체계'만 빠지고 나머지는 같게 나타났다.

따라서 학술논문을 분석한 선행 연구와 학위논문을 분석한 본 연구의 내용을 종합해 문화콘텐츠 연구 분야의 주요 연구 영역을 '한국 사회와 문화', '문화콘텐츠 활용', '문화콘텐츠 장르', '문화콘텐츠 제작 기술', '문화콘텐츠 산업' 5개로 제시할 수 있다. 특히 '활용'은 역사학, 국문학, 신문방송학 등 다른 학문 분야에서 찾아볼 수 없는 문화콘텐츠학의 독특한 영역이라고 할 수 있다.

<표 9> 문화콘텐츠 학위논문의 주요 연구 영역 및 연구 주제 분류 방안

주요 연구 영역	연구 주제 예시	관련 논문 수(비율)
사회와 문화	사회문화 프로그램	165 (9.1%)
	국가 문화 교류	
	조선시대차문화, 생활문화	
	전통 및 현대 음식 문화	
문화콘텐츠 활용	지역 역사 문화자원 활용	345 (19.0%)
	스토리텔링	
	문화유산 활용과 관광	
	신화 서사와 인물 활용	
	캐릭터 서사 및 정체성	
	문화원형 활용	
	주민 참여와 체험	
	도시와 사회 재생	
문화콘텐츠 제작 기술	미디어 및 매체	145 (8.0%)
	디지털 정보 기술	
	영상 제작 기술	
문화콘텐츠 장르	게임 콘텐츠	584 (32.2%)
	TV 방송 프로그램	
	공간 콘텐츠	
	뮤지컬 작품	
	공연예술 및 페스티벌	
	대중음악	
	웹툰 및 출판만화	
	아동 놀이 및 문학	
	여성 및 가족 콘텐츠	
	시각 영상 콘텐츠	
	교육 콘텐츠	
문화콘텐츠 산업	콘텐츠와 관객	573 (31.6%)
	서비스 만족도	
	한류 시장	
	콘텐츠산업 시장 성장	
	상품 광고	
	브랜드 이미지	
	기업 마케팅 전략	
	세계 문화콘텐츠 산업	
	문화예술 지원 사업	
	SNS 커뮤니케이션 시장	
	법제도 도입	
	수용자 경험 이론	
	조직 역량 평가 활용	

4. 시기별 연구 흐름 비교

2004년부터 2020년까지 문화콘텐츠 학위논문 분야의 연구 영역과 연구 주제가 어떻게 변했는지 알아보기 위해 데이터를 시기별로 분류해 분석했다. 먼저 시기를 나누기 전에 2가지 가설을 세우고 가설에 따라 분석을 한 후에 의미가 있는 결과가 나온 방식을 최종 선택했다. 가설 내용을 첫째, '정권의 정책이 학계 연구에 영향을 미치기까지 약 2년 정도 걸릴 것이다.' 둘째, '문화콘텐츠 연구자는 정권의 정책과 사회 이슈의 변화에 민감하게 반응하여 연구에 반영할 것이다.'로 정했다.

2가지 가설에 따라 데이터를 분석한 결과 첫째 가설보다 둘째 가설에서 유의미한 특징이 나타났다. 그래서 정권 기간에 따라 2004년부터 2020년까지 기간을 1시기(2004~2007), 2시기(2008~2012), 3시기(2013~2016), 4시기(2017~2020)로 구분했다.

4개의 시기별로 문화콘텐츠 학위논문 1,812편의 국문초록을 넷마이너의 쿼리(Query) 기능을 활용해 분류했다. 그리고 분류한 데이터를 전처리 작업하여 명사형 형태소를 각각 추출했다. 〈표 10〉은 그 내용을 정리한 것이다.

〈표 10〉 시기별 문화콘텐츠 학위논문 분류와 추출된 단어

구분	1시기(2004~2007)	2시기(2008~2012)	3시기(2013~2016)	4시기(2017~2020)
발표 논문 수	70	362	587	801
추출 단어 수	1,820	6,499	8,577	10,338

이렇게 분류한 데이터를 가지고 시맨틱 네트워크 분석을 했으며 분석 결과를 〈표11〉과 같이 추려서 제시했다. 2장에서 문서 동시 출현 빈도수, 연결 중심성, 위세 중심성이 모두 크게 나온 노드를 해당 연구 분야에서 중요하게 다뤄진 대상(이하 중심 연구 대상)으로 해석한다고 설명한 바 있다. 문화콘텐츠 학위논문의 중심 연구 대상이 시기별로 어떻게 변했는지 살펴보기 위해서 3개의 지표 값이 모두 크게 나온 노드를 〈표 11〉에서 조사했다.

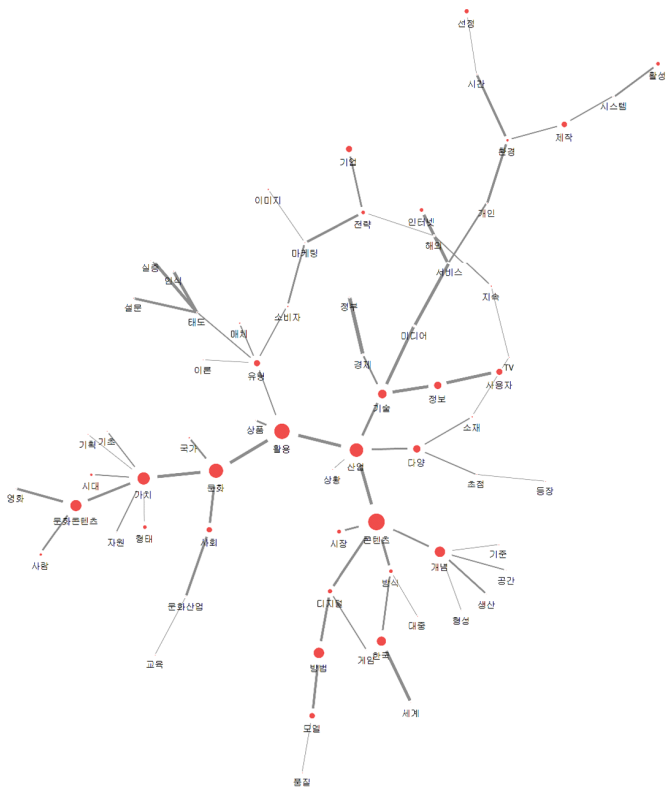
검토 결과, 중심 연구 대상은 1시기(2004~2007)에서 '콘텐츠, 활용, 문화, 산업, 가치, 기술, 개념', 2시기(2008~2012)에서 '문화, 콘텐츠, 사회, 활용, 가치, 한국, 산업', 3시기(2013~2016)에서 '문화, 활용, 사회, 가치, 콘텐츠, 지역', 4시기(2017~2020)에서 '문화, 활용, 사회, 콘텐츠, 가치, 한국'으로 나타났다. 눈에 띄는 점은 모든 시기에 걸쳐 '활용'이 중심 연구 대상으로 나타났다는 것이다. 즉 문화콘텐츠 학위논문 분야에서 '활용'은 시기와 상관없이 계속해서 중요하게 다뤘던 연구 대상이었다.

<표 11> 시기별 시맨틱 네트워크 분석 결과

구분	1시기(2004~2007)		2시기(2008~2012)		3시기(2013~2016)		4시기(2017~2020)	
	단어	값 (순위)	단어	값 (순위)	단어	값 (순위)	단어	값 (순위)
동시출현 빈도수	콘텐츠	37(1)	문화	193(1)	문화	300(1)	문화	394(1)
	활용	32(2)	콘텐츠	165(2)	활용	269(2)	활용	355(2)
	문화	31(3)	한국	157(3)	콘텐츠	215(3)	사회	322(3)
	산업	25(4)	활용	155(4)	한국	214(4)	콘텐츠	314(4)
	가치	24(5)	사회	140(5)	사회	212(5)	한국	301(5)
	문화콘텐츠	23(6)	방법	123(6)	가치	201(6)	방법	265(6)
	방법	22(7)	사람	118(7)	방법	182(7)	가치	262(7)
	개념	22(8)	산업	117(8)	사람	172(8)	사람	245(8)
	한국	21(9)	가치	115(9)	인식	165(9)	방식	233(9)
	기술	20(10)	형태	106(10)	예술	149(10)	중국	223(10)
	정보	19(11)	시대	98(11)	지역	141(14)	다양	193(15)
연결 중심성	콘텐츠	0.399180(1)	문화	0.416219(1)	문화	0.380704(1)	문화	0.369164(1)
	활용	0.397425(2)	콘텐츠	0.388746(2)	활용	0.354296(2)	활용	0.347153(2)
	문화	0.388644(3)	사회	0.358992(3)	사회	0.325216(3)	사회	0.331388(3)
	산업	0.380718(4)	활용	0.358449(4)	가치	0.322474(4)	콘텐츠	0.326436(4)
	가치	0.346152(5)	가치	0.353927(5)	콘텐츠	0.319922(5)	가치	0.318376(5)
	기술	0.344857(6)	한국	0.353224(6)	한국	0.312532(6)	한국	0.304692(6)
	정보	0.344630(7)	산업	0.350284(7)	인식	0.292723(7)	사람	0.297365(7)
	개념	0.343629(8)	사람	0.336923(8)	방법	0.283843(8)	방법	0.293898(8)
	시장	0.334795(9)	기술	0.319969(9)	다양	0.281264(9)	방식	0.292200(9)
	방법	0.331045(10)	형태	0.315557(10)	예술	0.279315(10)	시대	0.285252(10)
	다양	0.311041(11)	시대	0.313400(11)	지역	0.264873(15)	다양	0.280597(11)
위세 중심성	콘텐츠	0.176085(1)	문화	0.178345(1)	문화	0.174612(1)	문화	0.174004(1)
	활용	0.174340(2)	콘텐츠	0.166385(2)	활용	0.162140(2)	활용	0.163271(2)
	문화	0.170187(3)	사회	0.153514(3)	사회	0.149392(3)	사회	0.156282(3)
	산업	0.168248(4)	활용	0.153447(4)	가치	0.148839(4)	콘텐츠	0.153594(4)
	가치	0.152814(5)	가치	0.152734(5)	콘텐츠	0.146841(5)	가치	0.150485(5)
	기술	0.152388(6)	한국	0.152029(6)	한국	0.143799(6)	한국	0.144066(6)
	정보	0.151741(7)	산업	0.149847(7)	인식	0.134829(7)	사람	0.140086(7)
	개념	0.151264(8)	사람	0.144553(8)	방법	0.130830(8)	방법	0.138618(8)
	시장	0.147415(9)	기술	0.137066(9)	다양	0.129411(9)	방식	0.137447(9)
	방법	0.145461(10)	형태	0.135514(10)	예술	0.128959(10)	시대	0.134983(10)
	다양	0.136214(11)	시대	0.134822(12)	지역	0.123392(14)	다양	0.132552(11)

다음으로 중심 연구 대상이 다른 대상들과 어떻게 연결되어 영역을 이루고 이것은 시기별로 어떻게 달라졌는지 살펴보기 위해서 PFNet 네트워크 맵을 그려 비교했다.[7] 먼저 1시기(2004년~2007년) PFNet 네트워크 맵부터 살펴보자. [그림 4]에서 보듯이 노드와 링크의 수가 적고 하위 연결망도 발달돼지 않았다. 그런데 눈에 띄는 점은 다른 시기 맵에서 볼 수 없던 '개념' 노드가 나타났고 '가치' 노드에 연결된 노드의 수가 다른 시기보다 상대적으로 많다는 것이다. 이는 1시기(2004년~2007년)에 문화콘텐츠 개념과 가치 따위를 논의하는 연구가 활발하게 이뤄졌기 때문에 분석 결과에 나타난 것으

[그림4] PFNet 네트워크 맵(1시기(2004년~2007년))

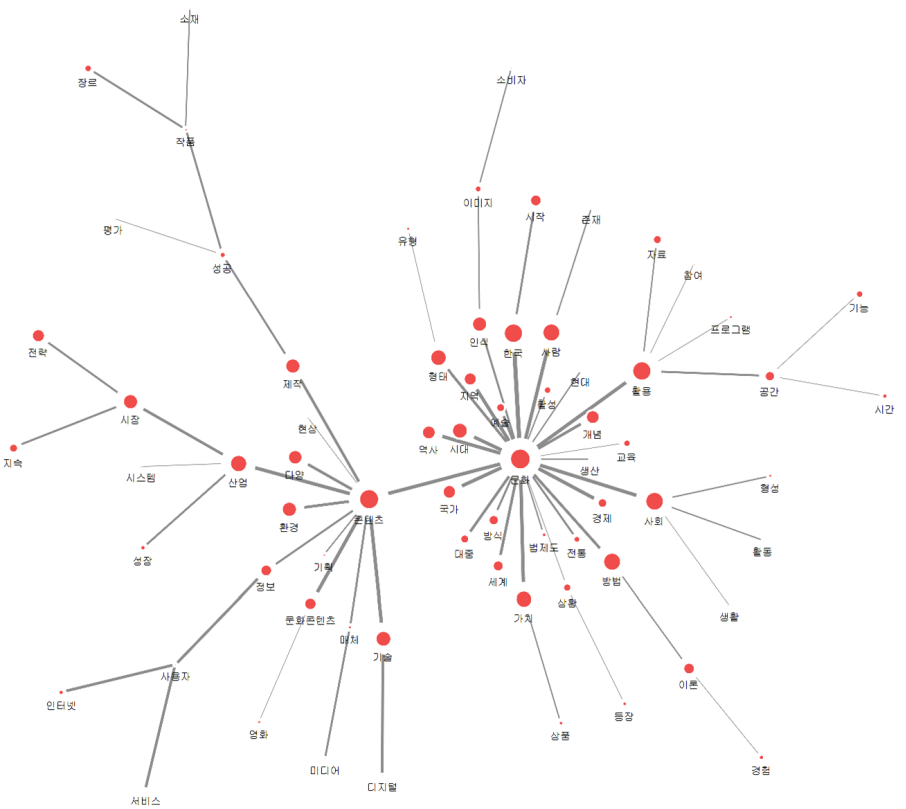


7) 필자는 선행 연구에서 본 논문과 같은 시기 구분 방법을 통해 문화콘텐츠 학술논문의 연구 흐름을 분석한 바 있다. 학술논문을 대상으로 했던 선행 연구의 결과를 요약해 설명하면 1시기(2003~2007)에 문화원형 소재 활용과 같은 연구가 활발하게 이뤄졌으며, 2시기(2008~2012)에 문화콘텐츠 산업을 확대하고 활성화하는 연구가 이뤄졌다. 3시기(2013~2016)에 활용과 관련한 연구가 하나의 연구 영역을 이루는 수준까지 발전했으며, 4시기(2017~2020)에 한류의 세계 확산과 같이 변화하는 세계 질서 속에서 한국 문화콘텐츠의 방향을 모색하는 연구가 이뤄졌다.

로 보인다. 2004년에 문화콘텐츠학 석사학위논문 6편이, 2006년에 박사학위논문 1편이 처음으로 발표됐던 점을 고려하면 2004년부터 2007년까지 문화콘텐츠 정체성 문제를 다룬 연구가 많았던 것으로 보인다.

[그림 5]는 2시기(2008~2012) PFNet 네트워크 맵을 그린 것이다. 1시기 때보다 노드와 링크의 수가 늘어났고 하위 연결망도 발달돼 있다. 연구 지형도로 해석할 수 있는 조건들이 나타나기 시작한 것이다. 그림에서 보듯이 전체 지도가 2개의 하위 연결망으로 구분된다. 하위 연결망은 '문화' 노드와 '콘텐츠' 노드를 중심으로 발달돼 있다. 이를 해석하면 2시기(2008~2012) 연구 지형은 문화 영역과 콘텐츠 영역을 중심으로 발달돼 있고 영역 관계가 문화를 연구하고 그 자료를 콘텐츠로 제작하는 쪽으로 형성돼 있다. 인문사회과학을 바탕으로 하는 문화 영역과 실용성을 바탕으로 하는 콘텐츠 영역을

[그림 5] PFNet 네트워크 맵(2시기(2008년~2012년))

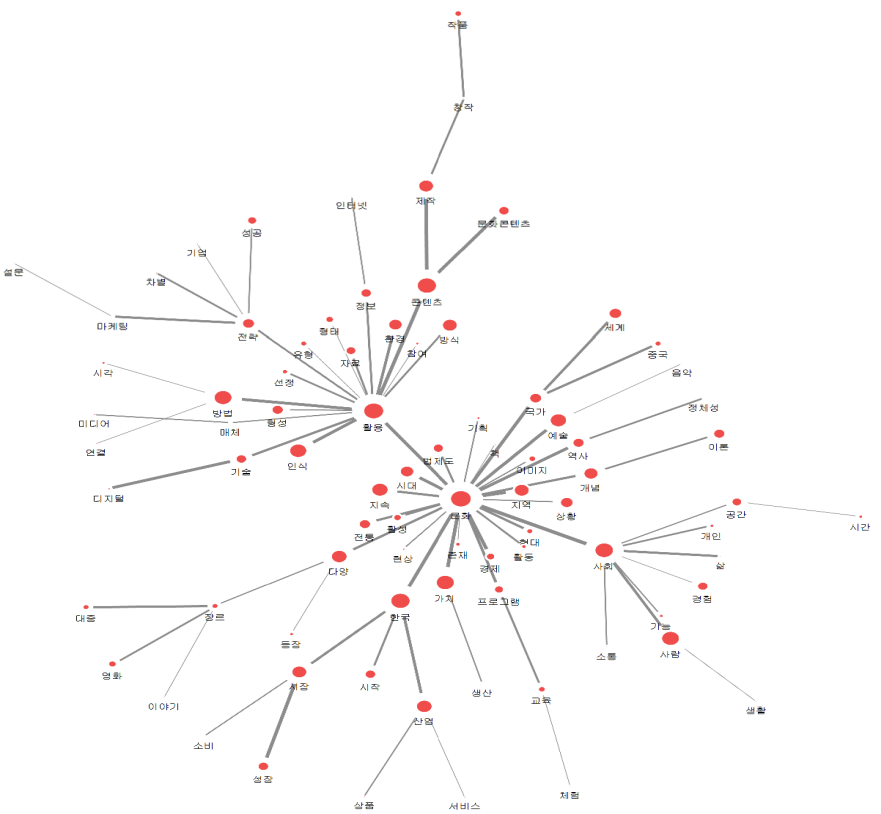


중심으로 연구 영역이 크게 구분되는 점이 흥미롭다.

2시기 맵에서 보이는 또 다른 특징은 '활용' 노드에 1시기에서 볼 수 없었던 자료, 프로그램, 공간 따위가 연결돼 있다는 것이다. 1시기 때 활용 노드에 상품, 산업, 소비자, 이론 노드가 연결돼 있었다. 이는 1시기(2004년~2007년)에 기존의 마케팅 이론을 바탕으로 활용 방법을 모색하는 연구가 주로 이뤄졌다면, 2시기(2008~2012)에 문화 자료 따위를 활용하는 연구가 주로 이뤄졌다고 해석할 수 있다.

[그림 6]은 3시기(2013~2016) PFNet 네트워크 맵을 그린 것이다. 2시기 맵과 비슷해 보이지만 2개의 하위 연결망이 '문화'와 '콘텐츠' 대신에 '문화'와 '활용' 노드 중심으로 발달돼 있다. 3시기(2013~2016)에 이르러 활용과 관련한 연구가 하나의 연구 영역을 이루는 수준까지 발전한 것이다. '활용' 영역에 나타난 연구 주제를 살펴보면 디지털 기술 활용, 마케팅 전략 활용, 자료 활용, 환경 활용, 인터넷 정보 활용, 콘텐츠 제작

[그림 6] PFNet 네트워크 맵(3시기(2013년~2017년))

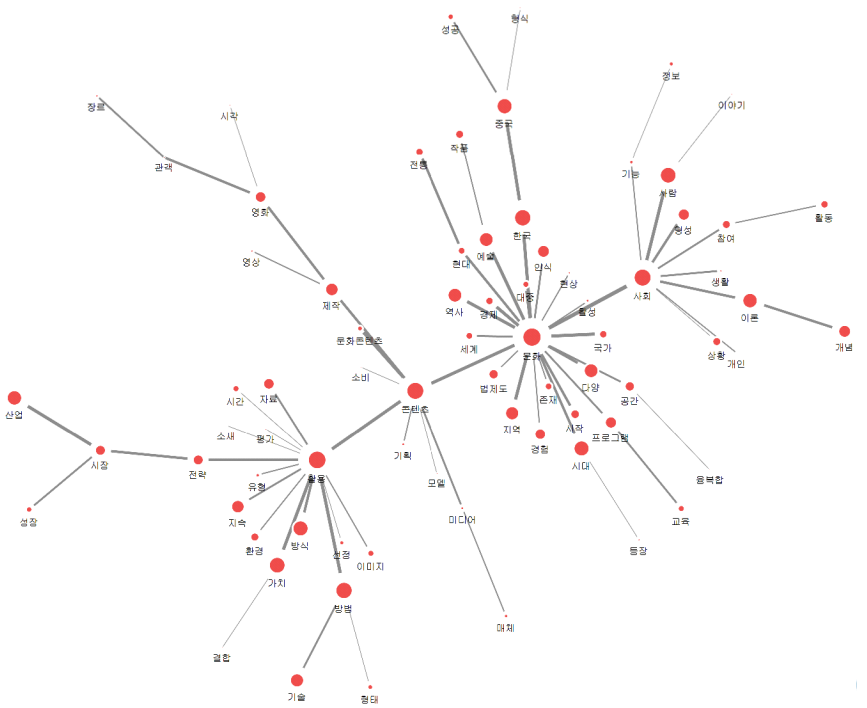


등이 있다. 2시기(2008~2012)에 자료와 프로그램 따위를 활용하는 연구가 이뤄졌다면, 3시기(2013~2016)에 활용 전략, 디지털 기술 활용, 환경 활용, 정보 활용과 같이 활용 방법에 관한 다양한 연구가 이뤄진 것이다.

3시기 맵에서 보이는 또 다른 특징은 '문화' 하위 연결망에서 노드들이 여러 단계에 걸쳐 연결돼 있다는 것이다. 예를 들어 2시기에서 한국-문화로 연결 관계가 끝났다면 3시기에서 한국-문화-산업, 한국-문화-시장-성장 등 여러 단계에 걸쳐 연결돼 있다. 이는 이 기간에 문화 관련 연구가 심화되고 범주화되었던 것으로 보인다.

마지막으로 4시기(2017~2020) PFNet 네트워크 맵을 살펴보자. [그림 7]에서 보듯이 지도가 4개 하위 연결망으로 구분되고 하위 연결망이 '문화', '사회', '콘텐츠', '활용' 노드를 중심으로 발달돼 있다. 이러한 특징은 앞에서 살펴본 전체 데이터를 대상으로 한 네트워크 맵(그림3)과 비슷하다. 이를 바탕으로 오늘날 문화콘텐츠 학위논문의 연구 지형은 4시기(2017~2020) 무렵에 완성된 것으로 생각할 수 있다. 다른 점이 있다면 4시기 맵에서 하위 연결망의 관계가 문화와 콘텐츠 사이에, 콘텐츠와 활용 사이에 형성

[그림 7] PFNet 네트워크 맵(4시기(2018년~2020년))



돼 있다는 것이다. 이는 4시기(2017~2020)에 문화 자료를 활용해 콘텐츠로 제작하는 연구보다 산업 시장에서 콘텐츠 활용 전략을 모색하는 연구가 주로 이뤄졌던 것으로 해석할 수 있다.

4시기 네트워크 맵에서 눈에 띄는 점은 중국 노드의 크기가 커졌다는 것과 연구 주제로 한국 문화와 중국 성공 등이 나타났다는 것이다. 이를 통해 4시기(2017~2020)에 한국 문화 콘텐츠의 중국 시장 전략과 같은 연구가 활발하게 이뤄졌음을 알 수 있다.

Ⅲ. 결론

지금까지 2004년부터 2020년까지 문화콘텐츠학 전공자들의 석·박사 학위논문 1,812편의 국문초록과 서지정보를 수집해 빅데이터 분석 방법을 통해 분석한 결과 내용을 살펴봤다. 서론에 제시한 연구 문제를 중심으로 연구 결과를 요약하면 다음과 같다.

첫째, 문화콘텐츠 학위논문 1,812편의 국문초록을 시맨틱 네트워크 분석의 문서 동시 출현 빈도수 지표로 분석했다. 분석 결과, 문화콘텐츠학 전공자들이 학위논문을 작성하면서 공통적으로 가장 많이 썼던 단어는 '문화(918)' '활용(811회)', '콘텐츠(731회)', '한국(693회)', '사회(691회)', '가치(602회)', '방법(592회)', '사람(549회)', '산업(497회)' 순으로 나타났다.

둘째, 문화콘텐츠 학위논문 분야에서 중요하게 다룬 연구 대상을 알아보기 위해서 문서 동시 출현 빈도수, 연결 중심성, 위세 중심성 3개 지표 값이 모두 크게 나온 노드를 조사했다. 조사 결과, '문화', '사회', '한국', '사람', '가치', '활용', '콘텐츠', '방법', '산업'이 중심 연구 대상으로 나타났다.

셋째, 중심 연구 대상이 다른 대상과 어떻게 연결되어 연구 영역과 주제를 이루는지 살펴보기 위해서 PFNet 네트워크 맵을 그려서 검토했다. 검토 결과, 문화콘텐츠 학위논문의 연구 영역은 크게 '사회 영역', '문화 영역', '활용 영역', '콘텐츠 영역'으로 구성돼 있다. 그리고 영역 사이의 관계는 사회와 문화를 연구하고 그 결과물을 활용하고 콘텐츠로 제작하는 쪽으로 형성돼 있다.

넷째, LDA 토픽 모델링을 통해 문화콘텐츠 학위논문 1,812편을 40개 주제로 요약, 분류했다. 토픽 모델링으로 분류된 주제를 검토한 결과, 문화콘텐츠 학위논문 분야에서

가장 많이 다뤄진 주제는 '콘텐츠와 관객(4.7%)'으로 나타났다. 이어서 '게임 콘텐츠(4.4%)', '지역 역사 문화 자원 활용(4.0%)', '교육 콘텐츠(4.0%)', '서비스 (만족도) 조사(3.9%)', '스토리텔링(3.8%)', '한류 시장(3.5%)'과 같은 순서로 나타났다.

다섯째, 시맨틱 네트워크 분석과 토픽 모델링의 결과를 활용해 '문화콘텐츠 학위논문의 주요 연구 영역과 주제 분류 방안'을 제시했다. 주요 연구 영역을 '사회와 문화', '문화콘텐츠 활용', '문화콘텐츠 장르', '문화콘텐츠 산업', '문화콘텐츠 제작 기술' 5개로 구분했다.

여섯째, 5개 영역별로 분류된 논문의 비율을 조사했다. 조사 결과 문화콘텐츠학 전공자들은 사회와 문화(9.1%)와 같이 인문・사회과학 기반의 연구보다 자원 활용, 콘텐츠 제작과 같이 실용성 목적의 연구(90.1%)를 압도적으로 많이 했던 것으로 나타났다.

일곱째, 2004년부터 2020년까지 문화콘텐츠 학위논문 분야의 연구 흐름을 살펴보기 위해서 시기를 1시기(2004~2007)와 2시기(2008~2012), 3시기(2013~2016), 4시기(2017~2020)로 구분하고, 시기별로 시맨틱 네트워크 분석을 했다. 분석 결과에서 두드러지게 나타난 특징은 모든 시기에서 '활용'의 분석 값이 모두 크게 나왔다는 것이다. 활용은 시기와 상관없이 계속해서 문화콘텐츠 학위논문 분야의 중요한 연구 대상이었던 것으로 파악됐다.

여덟째, 시기별로 연구 영역과 주제의 변화 흐름을 살펴보기 위해서 PFNet 네트워크 맵을 시기별로 비교했다. 검토 결과, 2시기 연구 영역이 '문화 영역'과 '콘텐츠 영역'을 중심으로 발달돼 있다. 영역 사이의 관계는 문화를 연구하고 그 자료를 콘텐츠로 제작하는 쪽으로 형성돼 있다. 연구 영역이 인문・사회과학을 바탕으로 하는 문화 연구 영역과 실용성을 목적으로 하는 콘텐츠 제작 영역으로 구분된다는 점은 주의 깊게 살펴볼 필요가 있다. 그리고 3시기(2013~2016)에 이르러 활용과 관련한 연구가 하나의 영역을 이루는 수준까지 발전했다. 4시기(2017~2020)에 연구 영역은 '사회', '문화', '활용', '콘텐츠'로 확장됐으며 네트워크 맵의 특징이 전체 데이터를 대상으로 그렸던 네트워크 맵과 비슷하게 나타났다. 이를 통해 4시기 무렵에 오늘날의 연구 지형도가 완성된 것으로 이해할 수 있다.

아홉째, 활용에 관한 연구 변화 흐름도 발견되었다. 1시기(2004~2007)에 기존의 마케팅 이론을 바탕으로 활용 방법을 모색하는 연구가 많았다. 그러나 2시기(2008~2012)에 문화 자료, 프로그램 따위를 활용하는 연구가 많았다. 3시기(2013~2016)에

디지털 기술 활용 등 방법에 관한 구체적이고 다양한 연구가 이뤄졌다. 그리고 관련 연구가 하나의 영역을 이루는 수준까지 발전할 만큼 활발하게 이뤄졌다. 4시기(2017~2020)에 문화 자료를 활용해 콘텐츠로 제작하는 연구보다 산업 시장에서 콘텐츠 활용 전략을 모색하는 연구가 주로 이뤄졌다.

요컨대 문화콘텐츠 학위논문의 연구 영역을 크게 '사회와 문화', '문화콘텐츠 활용', '문화콘텐츠 장르', '문화콘텐츠 산업', '문화콘텐츠 제작 기술' 5개로 구분할 수 있으며 이중에서 활용은 인문학 등 다른 학문 분야에서 찾아볼 수 없는 문화콘텐츠의 독특한 부분으로 내세울 수 있다. 그리고 문화콘텐츠학 전공자들은 사회와 문화와 같은 인문・사회과학 기반의 연구보다 자원 활용, 콘텐츠 제작과 같이 실용성 목적의 연구를 압도적으로 많이 했다. 이러한 점은 학문으로서 문화콘텐츠의 특징을 밝히는 일에 시사하는 바가 크다고 할 수 있다.

본 논문의 가치는 지금까지 쌓여 있기만 했던 문화콘텐츠 학위논문의 데이터를 분석해 연구 영역, 연구 주제, 연구 흐름을 도출했다는 것과 이를 통해 학문으로서 문화콘텐츠의 특징을 밝히는 일에 기초를 제공했다는 것이다. 그리고 연구 영역 관계, 연구 흐름과 같이 추상적 내용을 단어 클라우드와 지식 지도 형태로 제시했다는 것이다. 본 논문의 연구방법론은 문화콘텐츠 뿐만 아니라, 문화산업 정책과 같은 유사 분야에서 빅데이터 활용이 활성화되는 일에 도움을 줄 것이다.

[참고문헌]

김용학 · 김영진(2016),「사회 연결망 분석」, 서울:박영사.

문화관광부(2001),「콘텐츠 코리아 비전 21- 문화콘텐츠산업 발전 추진계획」.

문화관광부(2004),「참여정부 문화산업 정책 비전」.

문화체육관광부 (2012),「한국 콘텐츠정책 진흥체계 개선방안 연구」.

민요한 · 김지영 · 박옥남(2021), 토픽모델링과 키워드 네트워크 분석을 활용한 '문화콘텐츠' 연구 경향 분석,「사회과학연구」, 32권 2호, pp.113p-131.

박상천(2007), 문화콘텐츠 개념 정립을 위한 시론,「한국언어문화」, 제33집, pp.179-210.

박치완 · 유제상(2015), 문화콘텐츠학의 정체성 확립을 위한 필독서 선정 방안,「인문콘텐츠」, 제39호, pp.9-32.

사이람(2019),「'SNA를 활용한 연구동향 분석' 교육 자료」.

신광철(2014), 문화콘텐츠학 연구사 정리의 방향과 과제,「인문콘텐츠」, 제38호, pp.9-15.

오정심(2020a), 빅데이터 텍스트 마이닝을 통한 무형문화유산 분야 연구동향 및 지식체계 분석,「무형유산」, 제8호, pp.93-127.

오정심(2020b), 빅데이터 토픽 모델링 및 네트워크 분석을 통한 문화콘텐츠학 지식구조 연구,「문화정책논총, 제34권 2호, pp.35-69.

윤효준 · 박재현 · 윤지운(2019), 비정형 텍스트 자료에서 잠재정보 추출을 위한 토픽모델링 소개,「체육과학연구」, 제30호, pp.501-512.

이기창(2019),「한국어 임베딩」, 서울:에이콘.

임영상(2017),「역사와 문화콘텐츠」, 서울:신서원.

정창권(2007), 문화콘텐츠학, 어떻게 연구하고 가르칠 것인가,「동양한문학연구」. 24권 24호, pp.25-48.

조성준(2019),「세상을 읽는 새로운 언어, 빅데이터」, 파주:21세기북스.

한국정보화진흥원 (2015),「IT & Future Strategy 보고서」.

황동열 · 황고은(2016), 빅데이터 기술을 활용한 인문콘텐츠 분야의 의미연결망 분석,「인문콘텐츠」, 제43호, pp.229-255.

황서이 · 박정배 · 김문기(2020), 인문콘텐츠분야 연구의 경향 분석: 토픽모델링과 의미연결망분석을 중심으로,「인문콘텐츠」, 제56호, pp.123-138.

KOCCA(2017),「2017 콘텐츠 교육기관 및 인력수급 현황조사 보고서」.

Evelyn L.(2018.10.16.), Studying the stars with machine learning. symmetry, Available https://www.symmetrymagazine.org/article/studying-the-stars-with-machine-learning

KONKUK UNIVERSITY(2021.8.30.), Available: https://culturecontents.konkuk.ac.kr/html.do?siteId=CULTURECONTENTS&menuSeq=7453

Hankuk University of Foreign Studies(2021.08.30.), Available: http://www.hufs.ac.kr/user/hufsmuncon/

Chung Ang University(2021.8.30.) Available: http://www.gsa.cau.ac.kr/2017/intro03.php

Dooperida(2021.8.30.), Available: https://terms.naver.com/entry.naver?docId=3347329&cid=40942&categoryId=32845

Netminer(2021.8.30.), Available: http://www.netminer.com/main/main-read.do

[Abstract]

An Analysis of Research Trends in Theses on Cultural Contents using Big Data from 2004-2020

Oh, Jung Shim

This study aims to collect data on theses related to cultural contents published from 2004 through 2020, and analyze the main research subjects, topics, and research trends, using big data analysis. Furthermore, this study will establish the academic system in the field of cultural contents research.

For data collection, this study gathered and analyzed abstracts and bibliographical information of 1,812 theses from 2004 to 2020, using the keyword "cultural contents" in the department information search engine. This study was conducted broadly in four steps, including "data collection," "data pre-processing," "data analysis," and "synthesis and interpretation." Big data analyses were conducted using NetMiner.

The results of the study can be summarized as follows, focusing on the research questions presented in Chapter 1. First, due to the co-occurrence frequency analysis, "culture" had the highest frequency, followed by "utilization," "contents," "Korea," "society," "value," etc. These words were commonly used when the authors prepared theses.

Second, this study identified and examined the words that had high co-occurrence frequency, degree centrality, and eigenvector centrality. The words that had high values of the three indices can be the main research subjects in the field of studies of cultural contents or the keywords that influence the preparation of related texts. The examination revealed that "culture," "society," "Korea," "people," "value," "utilization," "contents" etc, were keywords or the main research subjects.

Third, keyword network analysis of data was conducted by classifying the

timeline into Period 1 (2004–2007), Period 2 (2008–2012), Period 3 (2013–2016), and Period 4 (2017–2020), to examine the change and flow of research in the field of theses on cultural contents, and the result was compared. According to the co-occurrence frequency and centrality analysis, the value of utilization was high in all periods. The analysis showed that utilization was an important research subject in theses on cultural contents throughout the periods.

The significance of this paper was that it enabled us to understand research subjects, the result of big data analysis, and the change and flow in the studies of cultural contents by collecting and analyzing academic data accumulated to date in the cultural contents research field. Additionally, the contents of the abstracts were intuitively comprehensible because the research results were presented in the form of knowledge maps.

[Keywords] cultural contents, research trend, big data analysis, semantic network analysis, topic modeling

Oh, Jung Shim_Ph. D. in Cultural Contents / First Author (ruaths0802@naver.com)