KOREA INSTITUTE FOR HEALTH AND SOCIAL AFFAIRS

사람을
생각하는
사람들

정책자료 2020-08

2020년 소셜 빅데이터 기반 보건복지 이슈 동향 분석

오매애
최호식·유한별

한국보건사회연구원
KOREA INSTITUTE FOR HEALTH AND SOCIAL AFFAIRS

【연구책임자】
오미애 한국보건사회연구원 연구위원

【공동연구진】
최호식 서울시립대학교 일반대학원 도시빅데이터융합학과 부교수
유한별 한국보건사회연구원 연구원

정책자료 2020-08

2020년 소셜 빅데이터 기반
보건복지 이슈 동향 분석

발 행 일 2020년 12월
발 행 인 조 흥 식
발 행 처 한국보건사회연구원
주 소 [30147]세종특별자치시 시청대로 370
세종국책연구단지 사회정책동(1~5층)
전 화 대표전화: 044)287-8000
홈페이지 http://www.kihasa.re.kr
등 록 1994년 7월 1일(제8-142호)
인 쇄 처 (사)아름다운사람들복지회

ISBN 978-89-6827-753-5 93510

발|간|사

보건복지 분야는 공급자 중심에서 수요자 중심의 맞춤형 서비스 체계로 변화하고 있다. 이러한 여건 변화를 빠르게 인지하고 보건복지 분야의 이슈를 파악하기 위해서는 소셜 빅데이터 활용이 중요하다. 빅데이터 분석 환경이 갖추어지고 기하급수적으로 늘어나는 비정형 빅데이터를 수집·분석할 수 있게 되면서 소셜 빅데이터 활용 수요는 증가하고 있다.

비정형 빅데이터는 모바일, 소셜네트워크서비스(SNS), 센서 등의 결합을 통한 새로운 형태의 데이터 생성으로 기존의 정형 데이터로는 파악하기 어려운 변화를 감지하고 정책 욕구를 즉시 확인하고 활용할 수 있는 근거를 제공한다. 비정형 빅데이터는 데이터에서 얼마나 많은 부가가치를 창출할 수 있는가의 관점에서 소셜 빅데이터 분석으로부터 새롭게 얻을 수 있는 지식 또는 부가가치의 양과 차이는 크지 않지만, '왜', '어떻게'에 대한 방향성을 제시해 줄 수 있다는 면에서 가치가 있다.

이 연구에서는 보건·복지 분야의 주요 키워드인 '보건', '복지', '사회보장' 관련 문서를 수집하고 다양한 분석결과를 살펴보았다. 그리고 비정형 빅데이터 활용성 확장을 위한 방법론에 대해서도 자세히 설명하였다. 본 연구를 위해 조언을 해 주신 많은 전문가들과 원고 집필에 참여해 주신 최호식 교수님께 감사드린다.

끝으로 본 보고서에 수록된 모든 내용은 우리 연구원의 공식적인 견해가 아니며 연구에 참여한 연구진의 의견임을 밝힌다.

2020년 12월

한국보건사회연구원 원장

조 흥 식

목 차

Abstract ····· 1

요 약 ····· 3

제1장 서 론 ····· 7

제2장 보건·복지·사회보장 키워드 빅데이터 분석 ····· 9

제1절 보건·복지·사회보장 상위 20개 키워드 분석 ····· 11

제2절 보건·복지·사회보장 주요 키워드 월별 트렌드 분석 ····· 31

제3절 보건·복지·사회보장 키워드 간 순위 비교 ····· 41

제3장 보건·복지·사회보장 키워드 클러스터링 분석 ····· 47

제4장 비정형빅데이터 활용성 확장을 위한 방법론 연구 ····· 61

제1절 비정형데이터와 정형데이터 ····· 63

제2절 임베딩방법론 ····· 70

제3절 임베딩에 기반한 연계방법론 ····· 85

제4절 딥러닝에 기반한 연계방법론 고도화 방안 ····· 95

제5장 결론 ····· 125

참고문헌 ····· 129

부록 ····· 133

표 목차

〈표 2-1〉 보건·복지·사회보장 상위 20개 키워드('20.01. ~ '20.11.) ····· 13
〈표 2-2〉 '코로나', '확진' 관련 뉴스기사의 예 ····· 17
〈표 2-3〉 '대구', '이태원', '클럽', '집회' 관련 뉴스기사의 예 ····· 22
〈표 2-4〉 '접종', '백신', '독감' 관련 뉴스기사의 예 ····· 26
〈표 2-5〉 '경제', '지원', '재난' 관련 뉴스기사의 예 ····· 29
〈표 2-6〉 '코로나' 관련 뉴스기사 예 ····· 32
〈표 2-7〉 '마스크' 관련 뉴스기사 예 ····· 33
〈표 2-8〉 '백신' 관련 뉴스기사 예 ····· 34
〈표 2-9〉 '접종' 관련 뉴스기사 예 ····· 35
〈표 2-10〉 '예방' 관련 뉴스기사 예 ····· 36
〈표 2-11〉 '지원금' 관련 뉴스기사 예 ····· 37
〈표 2-12〉 '지급' 관련 뉴스기사 예 ····· 38
〈표 2-13〉 '재난' 관련 뉴스기사 예 ····· 39
〈표 2-14〉 월별 키워드 목록(빈도수 2,500이하 기준) ····· 40
〈표 2-15〉 kendall-tau 기준 상위 20여개 키워드 ····· 42
〈표 2-16〉 kendall-tau 기준 하위 20개 키워드 ····· 44
〈표 4-1〉 자료의 설명 ····· 110
〈표 4-2〉 협동지식그래프 추천시스템방법의 입력과 출력 ····· 119

그림 목차

[그림 2-1] 월별 '코로나' 추출 건 수 ····· 31
[그림 2-2] 월별 '마스크' 추출 건 수 ····· 32
[그림 2-3] 월별 '백신' 추출 건 수 ····· 33
[그림 2-4] 월별 '접종' 추출 건 수 ····· 34
[그림 2-5] 월별 '예방' 추출 건 수 ····· 35
[그림 2-6] 월별 '지원금' 추출 건 수 ····· 37
[그림 2-7] 월별 '지급' 추출 건 수 ····· 38
[그림 2-8] 월별 '재난' 추출 건 수 ····· 39
[그림 2-9] kendall-tau 기준 상위 20여개 키워드 ····· 43
[그림 2-10] kendall-tau 기준 하위 20개 키워드 ····· 45
[그림 3-1] 1월 보건·복지·사회보장 키워드 클러스터링 결과 ····· 49
[그림 3-2] 2월 보건·복지·사회보장 키워드 클러스터링 결과 ····· 50
[그림 3-3] 3월 보건·복지·사회보장 키워드 클러스터링 결과 ····· 51
[그림 3-4] 4월 보건·복지·사회보장 키워드 클러스터링 결과 ····· 52
[그림 3-5] 5월 보건·복지·사회보장 키워드 클러스터링 결과 ····· 53
[그림 3-6] 6월 보건·복지·사회보장 키워드 클러스터링 결과 ····· 54
[그림 3-7] 7월 보건·복지·사회보장 키워드 클러스터링 결과 ····· 55
[그림 3-8] 8월 보건·복지·사회보장 키워드 클러스터링 결과 ····· 56
[그림 3-9] 9월 보건·복지·사회보장 키워드 클러스터링 결과 ····· 57
[그림 3-10] 10월 보건·복지·사회보장 키워드 클러스터링 결과 ····· 58
[그림 3-11] 11월 보건·복지·사회보장 키워드 클러스터링 결과 ····· 59
[그림 4-1] XML 자료 예시 ····· 67
[그림 4-2] JSON 자료 예시 ····· 68
[그림 4-3] 1차원 합성망 ····· 71
[그림 4-4] 오토인코더의 구조 ····· 74
[그림 4-5] Mnist 숫자 이미지에 대한 오토인코더와 입력과 복원된 출력 이미지 ····· 75
[그림 4-6] seq2seq 모형에서 어텐션의 활용 ····· 83

[그림 4-7] RBM의 모형구조 ····· 86
[그림 4-8] 요인모형 ····· 90
[그림 4-9] 요인모형 ····· 90
[그림 4-10] 심층 정준상관분석 ····· 93
[그림 4-11] LSTM 오토인코더 모형 ····· 99
[그림 4-12] LeNet-5 모형구조 ····· 100
[그림 4-13] 합성곱층(convolution layer) ····· 101
[그림 4-14] 풀링층(pooling layer) ····· 101
[그림 4-15] U-net의 구조 ····· 104
[그림 4-16] W-net의 구조 ····· 105
[그림 4-17] CAM 방법의 모형구조 ····· 107
[그림 4-18] 2020년 3월 1일 서울시 관내의 생활인구(20대) ····· 109
[그림 4-19] 2020년 3월 1일 서울시 관내의 생활인구(20대) ····· 111
[그림 4-20] 합성망모형 ····· 112
[그림 4-21] 25개 구역의 어텐션 벡터 ····· 113
[그림 4-22] 사용자 임베딩과 아이템 임베딩의 작용 모델 ····· 114
[그림 4-23] 지식그래프(knowledge graph)의 예시(Wang 등, 2019) ····· 115
[그림 4-24] 지식그래프에서의 도달 경로(path)의 예시 ····· 116
[그림 4-25] 심층정준상관분석 모형 구조들 ····· 121

Abstract

Social big data trend analysis based on health and welfare issues in 2020

Project Head: Oh, Miae

The health and social welfare sector is changing from a provider-oriented to a consumer-oriented customized service system. It is important to use social big data to readily recognize these changes and to identify issues in the health and welfare sector. The use of social big data is on increasing demand as big data analytics has improved rapidly and an ever-growing amount of unstructured big data is available for collection and analysis.

In this study, documents were collected monthly and looked at various analysis results to find out what topics are being issued in documents related to "health," "welfare," and "social security," the main keywords in the health and welfare sectors. We also elaborate on the relevant methodologies to extend the utilization of unstructured big data.

In this work, we conducted a theoretical review for expanding the utilization of unstructured big data, focusing on embedding methodologies. We examine linkage methods based on embedding methodology, and describe the methodologies of canonical correlation analysis and representation learning and

Co-Researchers: Choi, Hosik · Yu, Hanbyeol

upgrading that are utilized in linkage methodology.

Social big data analysis can serve as an important competitive edge in understanding the current situation of issues of national and social interest in the area of health and welfare policy.

요약

1. 연구의 배경 및 목적

보건복지 분야는 공급자 중심에서 수요자 중심의 맞춤형 서비스 체계로 변화하고 있다. 이러한 여건의 변화를 빠르게 인지하고 보건복지 분야의 이슈를 파악하기 위해서는 소셜 빅데이터 활용이 중요하다. 빅데이터 분석 환경이 갖추어지고 기하급수적으로 늘어나는 비정형 빅데이터를 수집·분석할 수 있게 되면서 텍스트 자료 활용 수요는 증가하고 있다.

이 연구에서는 보건·복지 분야의 주요 키워드인 '보건', '복지', '사회보장' 관련 문서에서 어떤 주제가 이슈화되고 있는지를 알아보기 위해 각 월별로 문서를 수집하고 다양한 분석결과를 살펴보았다. 그리고 비정형빅데이터의 활용성을 확장시키고자 관련 방법론에 대해서도 자세히 설명하였다.

2. 주요 연구결과

2장에서는 각 월별 상위 20개 키워드를 분석하였고, 2020년의 주요 키워드로 도출된 '코로나, 확진', '대구, 이태원, 클럽, 집회', '접종, 백신, 독감', '경제, 지원, 재난'과 관련된 기사를 구체적으로 살펴보았다. 그리고 각 월별로 연구진에서 선정한 주요 키워드인 '코로나', '마스크', '백신', '접종', '예방', '지원금', '지급', '재난' 키워드 추출 건수에 대한 추이를 살펴보았다. 각 월별 새로운 이슈로 나타난 키워드들도 비교해보기 위해 빈도수 2,500이하인 키워드들도 제시하였다. 시간의 흐름과 키워드 빈도수 간의 상관관계도 산출하여 하반기로 갈수록 어떤 키워드가 이슈화되었는지도 살펴보았다. 3장에서는 각 월별로 보건, 복지, 사회보장 관

련 키워드의 클러스터링 결과를 제시하였다. 클러스터링 수는 분석결과가 가장 해석가능하다고 판단되는 4개로 정하였다. 분석결과, 각 월별 클러스터링 결과에 큰 차이가 없었는데 이는 2020년이 코로나19 관련 이슈로 특수한 상황이었기 때문이다.

4장에서는 임베딩 방법론 위주로 비정형 빅데이터의 활용성 확장을 위한 이론적 검토를 하였다. 우선 임베딩 방법론에 기반한 연계 방안에 대해 살펴보고, 연계방법론에 활용되는 정준상관분석 및 표현학습 및 고도화 방법론을 기술하였다. 그리고 최근 딥러닝 모형 학습방법을 살펴보고 보건복지분야의 비정형 빅데이터 활용성 확장을 위해 추천시스템 등을 포함한 딥러닝 방법 고도화 부분을 설명하였다.

3. 결론 및 시사점

보건복지분야 주요 키워드로 살펴본 위 분석결과는 당연한 결과일 수 있으며 코로나 19 상황으로 인해 2020년 각 월별 주요 이슈 변화가 크지 않았다. 그럼에도 불구하고, 소셜 빅데이터 분석은 보건복지 정책 영역에서 국가적·사회적으로 관심이 있는 이슈에 대해 현 상황을 파악하는 데 중요한 경쟁력으로 작용할 수 있으며, 앞으로의 정책 관련 이슈를 도출하고 연구 전략을 세우는 데 근거자료로 활용될 수 있다. 비정형 빅데이터는 비정형데이터 및 정형데이터와 연계될 수 있는 가능성이 있다. 앞서 살펴본 임베딩 방법론에 기반한 연계 분석 기술을 바탕으로 주요 보건복지 정책에 관한 사회적 관심도, 영향력 등을 분석하고 그 변화 과정을 살펴본다면 시의성 높은 보건복지 정책 연구의 기반을 마련할 수 있을 것이다.

*주요 용어: 소셜 빅데이터, 보건, 복지, 사회보장, 키워드, 클러스터링, 임베딩 방법론

제1장

서론

제 1 장 서론

2020년도는 COVID-19의 이슈로 인해 모든 국내외 언론과 정책이 이에 대한 현황파악과 지원, 예방 대책으로 집중된 한 해이다. 보건복지 분야의 이슈 역시 코로나 19 관련 주제로 집중될 수밖에 없다. 그럼에도 불구하고 보건복지 분야의 이슈를 파악하기 위해서는 텍스트 분석도 중요하기에, 보건복지 분야의 주요 키워드인 '보건', '복지', '사회보장' 관련 문서를 2020년 1월부터 11월까지 수집하였고, 각 월별 주요 이슈 트렌트를 살펴보고자 하였다. 또한, 연구진에서 정한 '마스크', '지원', '백신' 등의 주요 키워드의 월별 트렌드도 분석해보고자 하였다. 그리고 클러스터링 분석을 통해 각 월별 주요 이슈를 도출하였다.

비정형 빅데이터의 중요한 부분 중 하나인 텍스트 분석은 저장하고, 표시, 출력하는 기록물 보관의 수단에서 발전하여 주어진 텍스트를 표현하는 효율적인 수리적 방법이 개발이 지속적으로 이루어지고 있다. 텍스트 분석은 자연어 처리, 텍스트 임베딩, 텍스트 주제추출, 텍스트 생성, 대화형 텍스트 처리 방법 등 수많은 관련 연구들이 존재하는데, 이번 연구에서는 텍스트 분석의 확장 가능성을 검토해보고자 다른 형식의 데이터를 연계하는 방법론을 살펴보았다. 그리고 보건복자분야에 적용할 수 있는 방안을 제시하고자 하였다.

본 보고서의 구성은 2장에서 보건, 복지, 사회보장 키워드 분석, 3장에서는 클러스터링 분석 결과를 제시하고자 한다. 4장에서는 비정형빅데이터 활용성 확장을 위한 방법론을 기술하고 5장은 결론으로 마무리하고자 한다.

제2장

보건·복지·사회보장 키워드 빅데이터 분석

제1절 보건·복지·사회보장 상위 20개 키워드 분석
제2절 보건·복지·사회보장 주요 키워드 월별 트렌드 분석
제3절 보건·복지·사회보장 키워드 간 순위 비교

제 2 장 보건·복지·사회보장 키워드 빅데이터 분석

제1절 보건·복지·사회보장 상위 20개 키워드 분석

2020년 1월부터 2020년 11월까지 보건·복지·사회보장 상위 20개 키워드를 살펴보면, 〈표 2-1〉과 같이 대부분이 COVID-19과 관련되어 있는 것을 알 수 있다. 구체적인 분석 결과는 다음과 같다.

빈도수를 기준으로 전체 기간 동안 코로나, 확진, 지역, 환자, 바이러스, 지원, 감염, 신종, 검사, 정부, 서울, 방역, 보건, 발생, 마스크, 중국, 사회, 격리, 병원, 확산 순으로 많이 나타났다. 1월의 경우, 중국, 코로나, 바이러스, 신종, 환자, 폐렴, 확진, 지역, 정부, 지원, 증상, 감염, 보건, 관리, 발생, 확산, 감염증, 격리, 서울, 복지 순으로, 2월은 코로나, 확진, 환자, 신종, 바이러스, 중국, 지역, 감염, 격리, 대구, 병원, 검사, 정부, 감염증, 방역, 마스크, 보건, 확인, 발생, 지원 순으로 나타났다.

3월의 상위 20개 키워드는 코로나, 확진, 마스크, 지역, 지원, 환자, 검사, 대구, 바이러스, 감염, 정부, 격리, 병원, 방역, 확산, 판정, 보건, 서울, 신종, 발생순으로, 4월은 코로나, 확진, 지원, 지역, 정부, 바이러스, 격리, 사회, 검사, 미국, 경제, 방역, 환자, 보건, 마스크, 감염, 서울, 복지, 신종, 재난 순으로 나타났다. 5월은 코로나, 확진, 지역, 검사, 지원, 서울, 방역, 정부, 클럽, 감염, 바이러스, 사회, 이태원, 경제, 센터, 보건, 사업, 미국, 환자, 발생순으로, 6월은 코로나, 확진, 지원, 지역, 서울, 검사, 감염, 보건, 방역, 복지, 사회, 정부, 발생, 사업, 센터, 경제, 바이러스, 마스크, 환자, 판정 순으로 나타났다.

7월은 코로나, 확진, 지역, 지원, 사업, 서울, 감염, 방역, 복지, 검사,

사회, 보건, 정부, 마스크, 발생, 바이러스, 환자, 시설, 센터, 경제 순으로, 8월은 코로나, 확진, 서울, 검사, 지역, 정부, 감염, 방역, 의료, 발생, 판정, 지원, 사회, 보건, 환자, 바이러스, 확산, 접촉, 집단, 집회 순으로 빈번히 나타났다. 9월의 상위 20개 키워드는 코로나, 확진, 지원, 서울, 지역, 감염, 방역, 정부, 사회, 사업, 검사, 의료, 발생, 복지, 보건, 금지, 추석, 환자, 바이러스, 센터 순으로, 10월은 코로나, 확진, 접종, 백신, 서울, 보건, 지원, 지역, 병원, 복지, 감염, 사회, 독감, 정부, 발생, 환자, 사업, 방역, 검사, 의원 순으로 나타났다. 마지막으로, 11월은 코로나, 확진, 지역, 지원, 사회, 서울, 검사, 감염, 복지, 사업, 발생, 방역, 정부, 보건, 시설, 판정, 경제, 금지, 센터, 단계 순으로 나타났다.

〈표 2-1〉 보건·복지·사회보장 상위 20개 키워드('20.01. ~ '20.11.)

순위	total	'20.01.	'20.02.	'20.03.	'20.04.	'20.05.	'20.06.	'20.07.	'20.08.	'20.09.	'20.10.	'20.11.
1	코로나	중국	코로나	코로나	코로나	코로나	코로나	코로나	코로나	코로나	코로나	코로나
2	확진	코로나	확진	확진	확진	확진	확진	확진	확진	확진	확진	확진
3	지역	바이러스	환자	마스크	지원	지역	지원	지역	서울	지원	접종	지역
4	환자	신종	신종	지역	지역	검사	지역	지원	검사	서울	백신	지원
5	바이러스	환자	바이러스	지원	정부	지원	서울	사업	지역	지역	서울	사회
6	지원	폐렴	중국	환자	바이러스	서울	검사	서울	정부	감염	보건	서울
7	감염	확진	지역	검사	격리	방역	감염	감염	감염	방역	지원	검사
8	신종	지역	감염	대구	사회	정부	보건	방역	방역	정부	지역	감염
9	검사	정부	격리	바이러스	검사	클럽	방역	복지	의료	사회	병원	복지
10	정부	지원	대구	감염	미국	감염	복지	검사	발생	사업	복지	사업
11	서울	증상	병원	정부	경제	바이러스	사회	사회	판정	검사	감염	발생
12	방역	감염	검사	격리	방역	사회	정부	보건	지원	의료	사회	방역
13	보건	보건	정부	병원	환자	이태원	발생	정부	사회	발생	독감	정부
14	발생	관리	감염증	방역	보건	경제	사업	마스크	보건	복지	정부	보건
15	마스크	발생	방역	확산	마스크	센터	센터	발생	환자	보건	발생	시설
16	중국	확산	마스크	판정	감염	보건	경제	바이러스	바이러스	금지	환자	판정
17	사회	감염증	보건	보건	서울	사업	바이러스	환자	확산	추석	사업	경제
18	격리	격리	확인	서울	복지	미국	마스크	시설	접촉	환자	방역	금지
19	병원	서울	발생	신종	신종	환자	환자	센터	집단	바이러스	검사	센터
20	확산	복지	지원	발생	재난	발생	판정	경제	집회	센터	의원	단계

1. 코로나, 확진

1월을 제외하고 2월~11월 모두 '코로나'가 제일 빈번히 나타났으며, 그 다음으로 '확진'이 1월을 제외하고 2~11월 모두 빈번히 나타났다. 11개월 간 전체 키워드 중 제일 많이 나타난 만큼 '코로나'에 대한 기사는 손쉽게 찾아볼 수 있다. 해당 뉴스기사의 예는 아래와 같다.

> *중국 중부 후베이(湖北)성 우한(武漢)에서 집단 발생한 원인 불명의 바이러스성 폐렴이 초보 단계 조사 결과 신종 코로나 바이러스로 판정됐다고 중국중앙방송(CCTV)이 9일 보도했다. 이 바이러스는 사스(SARS·중증급성호흡기증후군)를 일으키는 코로나 바이러스를 포함해 이미 발견된 것들과 다르며 추가 과학연구가 필요하다고 CCTV는 전했다. 전문가들은 병의 원인을 찾기 위해 전장 유전체 분석, 핵산 검사, 바이러스 분리 등을 실시했다. 환자 15명에서 신종 코로나바이러스 양성 결과가 나왔다.*
>
> *… 보도에 따르면 코로나바이러스는 호흡기와 장의 질환을 일으키는 병원체다. 인간 외에 소, 고양이, 개, 낙타, 박쥐, 쥐, 고슴도치 등의 포유류와 여러 종의 조류가 감염될 수 있다. 지금까지 확인된 코로나바이러스는 6종이다.*

자료: 김윤구(2020.01.09.). 중국 원인불명 폐렴, '신종 코로나바이러스' 잠정 판정. 연합뉴스. https://www.yna.co.kr/view/AKR20200109072900083 2020.12.22.인출.

> *신종 코로나바이러스 감염증(신종 코로나) 확산 우려로 학교 봄방학도 당초 일정보다 앞당겨지고 있다. 휴업에 들어간 학교의 경우 휴업 종료 후 곧바로 봄방학에 들어가는 곳도 있어 3주 이상의 '긴' 봄방학을 맞는 학교도 나올 것으로 전망된다. … 서울시교*

육청 관계자는 "신종 코로나 확산 위험이 있는 경우 수업일수 단축이나 조정을 통해 조기 봄방학이 가능하다고 일선 학교에 안내했다"며 "학교 운영위원회 논의를 거쳐 며칠 정도 봄방학을 앞당기는 학교들이 있다"고 밝혔다. 이날 기준 신종 코로나로 휴업 중인 학교는 전국 365곳으로 집계돼 지난 주말(647곳)보다 크게 줄었다. 휴업 중인 초·중·고교는 179곳으로 지난 주말(188곳)과 큰 변화가 없었지만 휴업에 들어간 유치원 상당수가 학사일정이 종료된 데 따른 것이다. 중국 후베이성을 방문한 뒤 자율격리 중인 학생·교직원은 총 7명으로 모두 건강상태는 양호하다고 교육부는 밝혔다.

자료: 송진식(2020.02.10.) ['신종 코로나' 확산불안한 학교들 '봄방학'도 앞당긴다. 경향신문. http://news.khan.co.kr/kh_news/khan_art_view.html?artid=202002101743001&code=940401&utm_campaign=zum_news&utm_source=zum&utm_medium=related_news 2020.12.22.인출.

'코로나' 다음으로 빈번하게 나타난 '확진'에 대한 기사의 예시는 아래와 같다.

경기 용인시 처인구 공무원 1명이 신종 코로나바이러스 감염증(코로나19) 확진 판정을 받았다. 용인시는 구청 건물을 소독하고 임시 폐쇄했다. 확진자 동료의 배우자가 근무하는 용인동부경찰서 사이버수사팀 사무실도 문을 닫았다. 7일 용인시에 따르면 A 씨(41·여)는 이날 오전 6시 코로나19 확진 판정을 받았다. 용인시는 구청 직원 등 400여 명의 출근을 금지하고 역학조사관의 조사 결과가 나올 때까지 자가 격리하도록 했다. A 씨의 동료 직원 25명에 대해서 코로나19 검사를 진행했다. … 용인시 관계자는 "A 씨는 기저질환이 없었고 최근 해외여행을 다녀오지도 않았다"며 "조사

결과가 나오는 대로 접촉자를 추가 격리할 것"이라고 말했다.

자료: 이경진(2020.04.08.). 용인 처인구청 폐쇄… 400여명 자가격리. 동아일보. http://www.donga.com/news/article/all/20200408/100548101/1 2020.12.22.인출.

7일 교육부에 따르면 다음날부터 초등학교 5, 6학년, 중학교 1학년 약 135만명이 올해 들어 처음으로 등교수업을 실시한다. 지난달 20일 고3을 시작으로 이어진 순차 등교 중 마지막 차례다. 이로써 전국 유치원과 초중고 학생 약 595만명이 모두 학교에 나가 수업을 받게 됐다.

… 확진자 발생 여파로 등교를 중단하는 학교도 속출하고 있다. 이날 서울 중랑구에 위치한 원묵고 3학년 학생 1명이 확진돼, 오는 8일부터 3일간 등교가 중지됐다. A(19)양은 지난 5일 친구 3명과 함께 롯데월드를 방문했는데 인근 롯데월드몰에서 확진자가 발생했다는 사실을 듣고 전날 진단검사를 받은 결과 이날 오전 10시 확진 통보를 받았다. 원묵고는 A양과 같은 반 학생, 선택교과 학생과 교사 등 150명을 우선 검사하고, 이후 전교생과 교직원에 대한 전수검사를 실시할 예정이다. …

자료: 송옥진(2020.06.07.). 8일 전 학년 등교인데…코로나 일일 확진자 50명대 어쩌나. 한국일보. https://www.hankookilbo.com/News/Read/202006071692373878 2020.12.22.인출.

이외 관련 뉴스기사는 〈표 2-2〉에 제시하였다.

〈표 2-2〉 '코로나', '확진' 관련 뉴스기사의 예

키워드	관련 뉴스기사
코로나	· 김윤구(2020.01.09.). 중국 원인불명 폐렴, '신종 코로나바이러스' 잠정 판정. 연합뉴스. https://www.yna.co.kr/view/AKR20200109072900083 2020.12.22.인출. · 송진식(2020.02.10.) ['신종 코로나' 확산]불안한 학교들 '봄방학'도 앞당긴다. 경향신문. http://news.khan.co.kr/kh_news/khan_art_view.html?artid=202002101743001&code=940401&utm_campaign=zum_news&utm_source=zum&utm_medium=related_news 2020.12.22.인출. · 이유범(2020.03.17.) 유·초·중·고 개학 2주 더 미뤄… 긴급돌봄·온라인 수업 강화. 파이낸셜 뉴스. https://www.fnnews.com/news/202003171811278620 2020.12.22.인출. · 김동일(2020.08.06.). 코로나로 의정부 위기가구 지원대상 급증… 지난해 2배 예산확보 비상. 경기일보. https://www.kyeonggi.com/news/articleView.html?idxno=2309636 2020.12.22.인출. · 박현영(2020.08.07.). 美, 한국 여행금지 5개월만에 풀었다…1단계 내려 여행재고. 중앙일보. https://news.joins.com/article/23843226 2020.12.22.인출.
확진	· 이경진(2020.04.08.). 용인 처인구청 폐쇄… 400여명 자가격리. 동아일보. http://www.donga.com/news/article/all/20200408/100548101/1 2020.12.22.인출. · 최예린(2020.05.20.). 충남 서산서 코로나19 확진자 발생…삼성서울병원 확진자 접촉. 한겨레. http://www.hani.co.kr/arti/area/chungcheong/945664.html 2020.12.22.인출. · 송옥진(2020.06.07.). 8일 전 학년 등교인데…코로나 일일 확진자 50명대 어쩌나. 한국일보. https://www.hankookilbo.com/News/Read/202006071692373878 2020.12.22.인출. · 오재용(2020.07.17.). '마스크 안쓰고, 해열제' 제주방문, 결국 4명 확진. 조선일보. https://www.chosun.com/site/data/html_dir/2020/07/17/2020071700766.html?utm_source=bigkinds&utm_medium=original&utm_campaign=news 2020.12.22.인출.

2. 대구, 이태원, 클럽, 집회

2, 3월의 경우 '대구'에서 집단감염이 발생함에 따라 이슈가 된 것으로 예상된다. 5월의 경우 이태원 클럽 관련 감염이 확산됨에 따라 '이태원', '클럽'이 이슈가 된 것으로 판단되며, 8월의 경우, 여러 집회로 인해 '집회'가 이슈로 떠오른 것으로 사료된다. 먼저, 2, 3월에 특히 빈번히 나타난 '대구'에 대한 기사의 예시는 아래와 같다.

정부가 21일 대구·경북을 중심으로 지역사회 전염 확산이 시작단계에 접어든 신종 코로나바이러스 감염증(코로나19) 방역을 위해 대구·청도 지역을 '감염병 특별관리지역'으로 지정하는 특단의 조치를 하기로 했다.

… 특히 대구·청도지역은 이번에 감염병 특별관리구역으로 지정되면서 예산 투입 등 재정적 지원과 중앙 정부 차원에서 병상·인력·장비가 전폭 지원되는 것은 물론, 군 인력을 포함해 공공인력까지 투입된다. 정부는 코로나19가 전국적인 확산은 아니지만 적어도 대구·청도 지역에서 만큼은 광범위한 지역 전파로 판단해 위기경보 '심각' 단계에 준하는 것으로 인식하고, 국가 차원의 비상한 대책을 통해 피해 사례가 집중된 지역을 중심으로 철저하게 차단하겠다는 것이다.

자료: 김대우(2020.02.22.) 대구·청도 군인력 투입…국가역량 총동원. 헤럴드경제. http://biz.heraldcorp.com/view.php?ud=20200221000477 2020.12.22.인출.

권영진 대구시장이 이재명 경기지사에게 코로나19(신종 코로나바이러스 감염증) 확진환자의 병상 제공을 요청했으나 이 지사가 수용하기 어렵다는 입장을 밝혔다. 대안으로 다른 일반 환자를 받겠다는 절충안을 내놨으나 권 시장은 이에 대한 반응을 내놓지 않았다. 이 경기지사는 26일 페이스북에 "대구의 어려움을 모르는 바 아니지만 대구의 코로나 확진자를 경기도의료원 등에 수용하는 문제는 정말로 어려운 주제"라고 밝혔다. … 앞서 권 시장은 이날 오전 이 지사에게 전화를 걸어 코로나19 확진환자를 위한 병상 확보에 비상이 걸렸다면서 대구 환자를 경기도 소재 병원에 입원할 수 있도록 도와 달라고 협조를 요청했다.

자료: 김병철, 한찬규(2020.02.27.). 권영진 "확진자 병상 제공" 요청에 이재명 "어렵다". 서울신문. http://go.seoul.co.kr/news/newsView.php?id=20200227012047 2020.12.22.인출.

한편, 5월에 빈번히 나타난 '이태원', '클럽'에 관한 기사의 예시는 아래와 같다.

방역당국에 따르면 A씨는 지난 2일 오전 12시 20분~오전 3시 이태원동 B클럽을 방문한 것을 비롯해 1~2일 이틀 동안 이태원에서 3곳 이상 클럽에 다녀간 것으로 알려졌다. 용산구는 A씨의 동선을 조사하고 있다. … 지금까지 용인시가 파악한 A씨의 접촉자는 냉면집 종업원, 주류전문점 사장, 보험사 직원 등 5명이다. A씨의 동거인은 6일 진단 검사에서 음성 판정을 받았다. 방역당국은 타 지역 동선을 확인하는 대로 발표하겠다고 밝혔다. 한 관계자는 "이태원 클럽들의 명부를 확보했지만 정확성을 따져봐야 한다"며 "A씨가 다녀간 클럽은 방역하고 일시 폐쇄했다"고 말했다.

자료: 최은경(2020.05.07.). 용인 20대 확진자, 이태원 클럽 3곳 이상 방문…"명부 있다". 중앙일보. https://news.joins.com/article/23770781 2020.12.22.인출.

12일 교육당국에 따르면 박백범 교육부 차관은 전날 "교육부는 방역당국과 협의를 거쳐 학생의 안전을 보장하기 위해 고3 학생의 등교 수업을 20일로 1주일 연기하는 것이 불가피하다고 결정했다"고 밝혔다. 이 때문에 고3을 포함한 다른 학생들의 등교도 모두 일주일씩 미뤄졌다. 지난 6일 첫 확진자가 발생한 이태원 클럽 관련 감염이 확산되면서 학교도 안전할 수 없다는 판단이다. … 급식 업체가 마주한 현실은 참담하다. 업계에 따르면 각 업체들은 등교일을 앞두고 수·목·금 3일치 점심 급식을 준비해놓은 상태였다. 등교 여부가 불투명하지만 등교할 경우를 대비한 것이다.

자료: 윤홍집(2020.05.13.). "수백명분 음식 버려야 하나"… 허탈한 급식업체. 파이낸셜 뉴스. https://www.fnnews.com/news/202005121747015472 2020.12.22.인출.

경기도가 이태원 일대 유흥시설 방문자에게 '대인접촉 금지 명령'을 내린 것과 관련해 정부가 방역에 도움이 되는 방안이라고 판단한다며 전국 확대 여부를 검토하겠다고 밝혔다. 박능후 중앙재난안전대책본부 1차장(보건복지부 장관)은 10일 정부서울청사에서 열린 정례 브리핑에서 "대인접촉 금지 명령의 실효성이 담보된다면, 이태원 클럽 집단감염 사태와 관련해 방역에 많은 도움이 될 것"이라고 말했다. … 앞서 정부는 이태원 클럽발 집단감염이 확산하자 8일 보건복지부 장관 명의로 전국 유흥시설에 대해 운영 자제를 권고하는 행정명령을 내렸다.

자료: 여승구(2020.05.11.). 박능후 "이재명 경기도지사의 대인접촉 금지 명령, 전국 확대 여부 검토". 경기일보.http://www.kyeonggi.com/news/articleView.html?idxno=2281716 2020.12.22.인출.

8월의 경우, 빈번히 나타난 단어 중 '집회'와 관련된 기사의 예시는 아래와 같다.

청주시의 안일한 방역 대처가 코로나19 바이러스 지역 감염을 확산시켰다는 지적이 일고 있다. … 이슬람교의 2대 축제 중 하나인 '이드알아드하'로 불리는 행사로 당시 아랍권 등 외국인 341명이 참석한 것으로 확인됐다. 참석자들은 행사 후 마스크를 벗고 빵과 우유를 나눠 먹었다는 진술도 전해지고 있다. 문제는 청주시가 이 종교집회를 사전에 알고 있으면서도 제대로 대처하지 않았다는 점이다.

청주흥덕경찰서는 지난 7월 29일 흥덕보건소에 "외국인 400명이 참석하는 행사 있으니 방역조치가 필요하다"고 요청했다. 그러나 해당 보건소는 미숙한 정보 파악으로 엉뚱한 곳에 소독작업

만 했다. … 풋살장 정도 크기의 공원에서 외국인 수백 명이 밀접 접촉하면서 집회를 한 현장에는 방역 손길이 미치지 않은 것이다.

자료: 박재원(2020.08.05.). 종교행사 알고도 '수수방관'… 청주시, 코로나 확산 키웠다. 중부매일. http://www.jbnews.com/news/articleView.html?idxno=1301412 2020.12.22.인출.

정부가 수도권 코로나19의 가파른 확산세가 대구·경북 집단감염 때보다 더 위험한 상황이라고 경고했다. 수도권을 중심으로 다양한 지역과 시설에서 코로나19 집단감염이 발생하고 감염자의 교회 예배와 집회를 통한 불특정 다수 접촉 등으로 신규 확진자 급증 우려가 높아져서다. 17일 정부세종청사에서 열린 정례브리핑에서 김강립 중앙사고수습본부 1총괄조정관(보건복지부 차관)은 "현재 서울·경기 상황은 지난 2~3월 대구·경북의 집단감염 사태를 떠올리게 하지만, 감염양상이나 방역 대응 측면에서는 그때보다 더 위험한 요소를 지니고 있다"고 밝혔다.

… 실제 질병관리본부 중앙방역대책본부는 17일 오전 0시 기준 국내 코로나19 하루 신규 확진자를 197명으로 집계했다. 지난 13일 신규확진자가 세자릿수로 재진입한 이후 16일까지 나흘간 745명에 이른다. 하루평균 186명에 달해 재확산 우려가 고조되고 있다.

자료: 정명진(2020.08.18.). 나흘간 745명… "대구·경북때보다 위험". 파이낸셜 뉴스. https://www.fnnews.com/news/202008171727290316 2020.12.22.인출.

이외 관련 뉴스기사의 예는 〈표 2-3〉과 같다.

〈표 2-3〉 '대구', '이태원', '클럽', '집회' 관련 뉴스기사의 예

키워드	관련 뉴스기사
대구	· 최태범(2020.02.02.). 또 '3차 감염' 가능성…국내 신종코로나 환자 총 15명(종합). 머니투데이. https://news.mt.co.kr/mtview.php?no=2020020214230312945&outlink=1&ref=%3A%2F%2F 2020.12.22.인출. · 김대우(2020.02.22.) 대구·청도 군인력 투입…국가역량 총동원. 헤럴드경제. http://biz.heraldcorp.com/view.php?ud=20200221000477 2020.12.22.인출. · 김병철, 한찬규(2020.02.27.). 권영진 "확진자 병상 제공" 요청에 이재명 "어렵다". 서울신문. http://go.seoul.co.kr/news/newsView.php?id=20200227012047 2020.12.22.인출. · 최종필(2020.03.05.). 전남, 경북에 '사랑의 도시락' 9000개 지원. 서울신문. http://go.seoul.co.kr/news/newsView.php?id=20200305012004 2020.12.22.인출. · 구자윤(2020.03.26.). 지역특산물 '가치삽시다'… 중기부 온라인 기획전. 파이낸셜뉴스. https://www.fnnews.com/news/202003251742095244 2020.12.22.인출.
이태원	· 최은경(2020.05.07.). 용인 20대 확진자, 이태원 클럽 3곳 이상 방문…"명부 있다". 중앙일보. https://news.joins.com/article/23770781 2020.12.22.인출. · 윤홍집(2020.05.13.). "수백명분 음식 버려야 하나"… 허탈한 급식업체. 파이낸셜 뉴스. https://www.fnnews.com/news/202005121747015472 2020.12.22.인출. · 이정민(2020.05.18.). 성남시, 개학 앞둔 모든 학생 교직원에 마스크 지원. 경기일보. https://www.bigkinds.or.kr/v2/news/search.do;Bigkinds=32761FEF4DF02D06E9BFDBA98695430D 2020.12.22.인출. · 이유범(2020.05.18.). 丁총리 "고3 예정대로 20일 등교개학… 지역감염 통제 가능한 수준". 파이낸셜 뉴스. https://www.fnnews.com/news/202005171742580258 2020.12.22.인출. · 신혜연(2020.05.30.). 수도권 방역 강화 조치 후 첫 주말… 박능후 "외출 자제 부탁드린다". 중앙일보. https://news.joins.com/article/23789471 2020.12.22.인출.
클럽	· 황경근(2020.05.06.). 황금연휴 19만명 다녀간 제주… '조용한 전파자' 우려에 초긴장. 서울신문. http://go.seoul.co.kr/news/newsView.php?id=20200506012024 2020.12.22.인출. · 여승구(2020.05.11.). 박능후 "이재명 경기도지사의 대인접촉 금지 명령, 전국 확대 여부 검토". 경기일보. http://www.kyeonggi.com/news/articleView.html?idxno=2281716 2020.12.22.인출. · 오세광(2020.05.19.). 부천시, 코로나19 대응 '워크스루' 선별진료소 운영. 경기일보. http://www.kyeonggi.com/news/articleView.html?idxno=2285273 2020.12.22.인출. · 강국진(2020.05.25.). 클럽·노래방 갈 때 'QR코드' 찍는다. http://www.seoul.co.kr/news/newsView.php?id=20200525002006 서울신문. 2020.12.22.인출. · 이상규(2020.05.31.). 확진자 증가세 속에도 정부 3차 개학 진행…178만명 학교로. 매일경제.https://www.mk.co.kr/news/society/view/2020/05/554891/2020.12.22.인출.

키워드	관련 뉴스기사
집회	· 박재원(2020.08.05.). 종교행사 알고도 '수수방관'… 청주시, 코로나 확산 키웠다. 중부매일. http://www.jbnews.com/news/articleView.html?idxno=1301412 2020.12.22.인출. · 백경서(2020.08.11.). 지진 70% 보상에 뿔난 포항시민들 "100% 구제하라"…300명 청와대 상경시위. 중앙일보. https://news.joins.com/article/23845793 2020.12.22.인출. · 이승륜(2020.08.13.). 의협 "14일 총파업 강행"…부산 50~60% 참여 전망. 국제신문. http://www.kookje.co.kr/news2011/asp/newsbody.asp?code=0300&key=20200813.22004004037 2020.12.22.인출. · 정명진(2020.08.18.). 나흘간 745명… "대구·경북때보다 위험". 파이낸셜 뉴스. https://www.fnnews.com/news/202008171727290316 2020.12.22.인출. · 이연우(2020.08.24.). 정은경 "거리두기 3단계 필요성 검토"…수도권 신규 확진자 급증, 병상 확보 '비상'. 경기일보. http://www.kyeonggi.com/news/articleView.html?idxno=2312749 2020.12.22.인출.

3. 접종, 백신, 독감

10월의 경우, 이전과 달리 '접종', '백신', '독감'이 상위 20개 키워드 목록에 포함되었으며, 이는 10월에 실시된 독감 접종 및 COVID-19 백신 개발 관련 이슈로 인한 것으로 판단된다. 이와 관련된 기사의 예시는 아래와 같다.

상온 노출이 의심되는 인플루엔자(독감) 백신을 접종한 사람들이 1300여명으로 늘어난 가운데 부작용에 대한 우려도 커지고 있다. 1일 질병관리청에 따르면 현재 상온 노출 여부를 조사 중인 정부조달 백신 물량을 접종한 건수는 지난달 28일 기준 1362명이다. 질병관리청에 따르면 전날 기준 백신 이상반응 신고사례는 총 4건이다. 중등도 이상의 부작용은 나타나지 않는 상황이다. 사례별로는 접종부위 통증 1건, 발열 1건, 오한·근육통 1건, 접종부위 멍 1건이다. 이 중 접종부위 통증 사례는 증상이 호전됐다.

… 보건당국은 현재 접종자의 건강 상태를 매일 확인하고 있다.

지방자치단체가 유선 전화 또는 문자 메시지를 통해 접종일 1주일간 모니터링하는 방식이다. 정은경 질병관리청장은 "백신의 안정성에서 가장 문제가 되는 것은 접종 후 알레르기 초기나 중증 부작용이 생길 수 있고 접종 후 2~3일 안에 발열이나 발작 같은 이상반응이 보통 보고된다"고 설명했다.

자료: 조현의(2020.10.01.). '상온노출 백신' 부작용 있나 없나. 아시아경제. https://www.asiae.co.kr/article/2020092912125804231 2020.12.22.인출.

신종 코로나바이러스 감염증(코로나19) 대유행이 해를 넘길 가능성이 제기되면서 세계인들의 관심은 온통 백신에 쏠리고 있다. 국가임상시험지원재단에 따르면 지난 9월 15일 기준으로 미국 국립보건원(NIH)에 등록된 코로나19 백신 임상시험은 83건이다. 다수 제약기업들이 개발하고 있는 이들 백신은 모두 코로나19를 예방하기 위한 목적이긴 하지만, 작용 방식이 각기 다르다. … 대다수 코로나19 백신은 바로 이 스파이크 단백질이 체내에서 생성되도록 하는 게 목적이다. 실제 코로나19 바이러스가 몸 속에 들어오지 않았어도 면역체계가 스파이크 단백질을 인식하고 마치 바이러스가 들어온 것처럼 착각하게 만들어 대응 훈련을 시키는 것이다.

자료: 임소형(2020.10.01.). "코로나19 백신, 만드는 방식 이렇게 다르네" 따져봐야 안전하다. 한국일보. https://www.hankookilbo.com/News/Read/A2020092317320000134 2020.12.22.인출.

전세계 주요 제약사들이 코로나19 백신 개발 막바지에 돌입하면서 내년에 최대 160억회분이 풀릴 것으로 보인다. 하지만 이를 어떻게 공급할지가 문제로 떠오른다는 지적이 나온다. 28일 홍콩 사우스차이나모닝포스트(SCMP)는 영국 의약시장 조사업체 에어피니티(Airfinity) 조사 결과를 인용해 현재 글로벌 제약사들이

내년 생산할 것으로 예상되는 백신은 총 163억2201만3000회분이라고 보도했다. 하지만 백신의 초기 공급은 양극화할 것으로 보인다. 이미 선진국들이 상당한 물량을 확보해 개발도상국들의 백신 공급 시기가 늦어질 뿐만 아니라, 공급 되더라도 물량을 소화하지 못할 가능성이 큰 상황이기 때문이다.

… 이렇게 되면 결국 선진국들의 선구매 물량에만 공급이 집중될 수 밖에 없어 다른 국가들은 소외될 수 밖에 없다는 지적이다. 이를 방지하기 위해 전세계에 동등한 백신 접근권을 주기 위해 '코백스 퍼실리티' 프로젝트가 운영되고 가입국도 184개국이 됐지만 여전히 문제는 있다. … 호주 라트로브대 보건전문가인 데보라 글리슨은 선진국에 기반한 제약사들에게 백신 공급을 의존하는 것은 비현실적이라고 지적한다. 가격 책정 문제나 생산할 수 있는 분량을 생각하면 그렇다는 얘기다. 글리슨 교수는 "코백스도 결국 각국 인구의 20%만 접종받을 수 있는 백신을 확보하는데 그친다"면서 "이같은 공급 문제를 해결하려면, 지적재산권을 공유하는 등 다른 해결책이 필요하다"고 말했다.

자료: 강기준(2020.10.29.). 내년에 풀릴 백신 최대 160억회분…문제는 공급. 머니투데이. https://news.mt.co.kr/mtview.php?no=2020102812012731774&outlink=1&ref=%3A%2F%2F 2020.12.22.인출.

인플루엔자(독감)백신 접종 후 사망하는 사고가 잇따르면서 불안감이 증폭되고 있다. 고령층뿐 아니라 10대에서도 사망자가 발생하고, 지역이 다양해 '독감백신 포비아(공포증)'로 확산되는 분위기이다. … 현재 백신 접종과 사망 간의 직접적인 연관성은 아직 확인되지 않았다. 하지만 인천, 전북, 대전, 대구, 제주 등 각 지역에서 독감백신 접종 이후 숨지는 일이 잇따라 발생하면서 불안감이 커지고 있다. 감염병 전문가들은 건강한 일반인의 경우

독감 백신으로 인한 사망이 드물어 과도한 공포감을 가질 필요는 없다고 조언한다. 하지만 고령자의 경우에는 독감백신을 맞은 후 사망한 사례가 있어 주의할 필요가 있다.

자료: 정명진(2020.10.22.). 6일새 9명 사망… 독감백신 공포. 파이낸셜뉴스. https://www.fnnews.com/news/202010211827413130 2020.12.22.인출.

이외 관련 뉴스기사의 예는 〈표 2-4〉와 같다.

〈표 2-4〉 '접종', '백신', '독감' 관련 뉴스기사의 예

키워드	관련 뉴스기사
접종	· 조현의(2020.10.01.). '상온노출 백신' 부작용 있나 없나. 아시아경제. https://www.asiae.co.kr/article/2020092912125804231 2020.12.22.인출. · 이무헌, 김천열(2020.10.06.). 청소년〈만 13~18세〉 8만3천명 독감백신 못 맞아…'골든타임 놓칠라' 전전긍긍. 강원일보. http://www.kwnews.co.kr/nView.asp?s=501&aid=220100500070 2020.12.22.인출. · 홍수민(2020.10.15.). 러시아, 두번째 코로나 백신도 공식 승인···두달만에 자체 개발. 중앙일보. https://news.joins.com/article/23894704 2020.12.22.인출. · 정자연(2020.10.24.). 독감백신 접종 후 사망자 일주일 동안 32명…경기도 3명. 경기일보. http://www.kyeonggi.com/news/articleView.html?idxno=2324270 2020.12.22.인출. · 강기준(2020.10.29.). 내년에 풀릴 백신 최대 160억회분...문제는 공급. 머니투데이. https://news.mt.co.kr/mtview.php?no=2020102812012731774&outlink=1&ref=%3A%2F%2F 2020.12.22.인출.
백신	· 임소형(2020.10.01.). "코로나19 백신, 만드는 방식 이렇게 다르네" 따져봐야 안전하다. 한국일보. https://www.hankookilbo.com/News/Read/A2020092317320000134 2020.12.22.인출. · 강국진(2020.10.06.). 국가예방접종 사업 백신 3년간 4만명분 폐기. 서울신문. http://www.seoul.co.kr/news/newsView.php?id=20201006012022 2020.12.22.인출. · 임보미(2020.10.15.). 일라이릴리도 항체치료제 3상 시험 중단. 동아일보. https://www.donga.com/news/article/all/20201015/103426142/1 2020.12.22.인출. · 최선을(2020.10.24.). "사망자 36명"…'독감접종 계속' 방침에도 곳곳서 혼선(종합). 서울신문. http://www.seoul.co.kr/news/newsView.php?id=20201024500002 2020.12.22.인출. · 박경은(2020.10.31.). 독감백신 접종 후 사망자 83명 중 72명 '무관'…"접종 계속". 아주경제. https://www.ajunews.com/view/20201031161141137 2020.12.22.인출.

키워드	관련 뉴스기사
독감	· 김명호(2020.10.05.). 경인지역, '상온노출 의심' 독감백신 접종 887명. 경인일보. http://www.kyeongin.com/main/view.php?key=20201004010000256 2020.12.22.인출. · 이범수(2020.10.13.). 백신보다 효과 좋은 '개인위생'… 독감 환자 절반으로 줄었다. 서울신문. http://www.seoul.co.kr/news/newsView.php?id=20201013008016 2020.12.22.인출. · 전주영(2020.10.19.). 19일부터 70세이상 무료 독감접종 "컨디션 좋을때 마스크 쓰고가세요". 동아일보. https://www.donga.com/news/article/all/20201019/103501634/1 2020.12.22.인출. · 정명진(2020.10.22.). 6일새 9명 사망… 독감백신 공포. 파이낸셜뉴스. https://www.fnnews.com/news/202010211827413130 2020.12.22.인출. · 신성식, 김민욱(2020.10.28.). 부처 칸막이에 방치된 백신 유통…질병청은 단속권도 없다. 중앙일보. https://news.joins.com/article/23905194 2020.12.22.인출.

4. 경제, 지원, 재난

한편, COVID-19 확산 등으로 인해 '경제' 여건이 악화되고 정부 및 지방자치단체에서 재난 지원금을 지급하여 '지원', '재난' 등의 단어가 자주 나타난 것으로 판단된다. 해당 뉴스기사의 예시는 아래와 같다.

> *코로나19 확산에 따른 서비스 수요 위축과 유가 하락 등이 겹치며 지난달 소비자물가 상승률이 올해 처음 0%대로 떨어졌다. 전세계적인 코로나19 충격으로 경기 불확실성이 커져 저물가 추세가 계속될 수 있다는 전망이 나오는 가운데 통계당국은 디플레이션 가능성에는 선을 그었다. … 최근 경제여건을 감안할 때 당분간 0%대 저물가 추세가 계속될 수 있다는 분석이 나온다. 지난달 물가상승률이 마찬가지로 코로나19 영향이 있었던 지난 3월(1.0%)보다 크게 떨어지고, 근원물가도 20여년만에 최저치를 나타내는 등 물가를 끌어올릴 수요 자체가 위축됐다는 것이다.*

자료: 박광연(2020.05.04.). 4월 소비자물가 0.1% 상승···코로나19·유가 하락으로 올해 첫 0%대. 경향신문. http://biz.khan.co.kr/khan_art_view.html?artid=202005040929001&code=920100 2020.12.22.인출.

정세균 국무총리는 8일 논란이 계속되는 코로나19 긴급재난지원금과 관련, "고소득자 환급을 전제로 모든 국민에게 보편적으로 지급할 수 있다"고 밝혔다. 정 총리는 이날 정부세종청사에서 열린 기자간담회에서"국민 모두에게 지원금을 지급하는 것은 고소득자에게 다시 환수한다는 전제조건이 충족되면 못할 일도 없다"며 이같이 말했다.

… 정 총리는 "신속성, 행정편의 차원에서 100% 다 지급하는 것이 쉽고 논란의 소지도 없다. 그러나 지원금액이 상당히 커서 필요한 사람에게 선별적 지원이 원칙"이라고 말했다. 다만 정 총리는 "급하고 속도전이 필요한 상황에서는 타협할 수 있다. 고소득자에 대해 환수장치가 마련된다는 전제조건이 있어야 한다"고 했다.

자료: 정상균(2020.04.09.). 정세균 "고소득자 환급 전제, 재난지원금 전국민 지급 가능". 파이낸셜뉴스. https://www.fnnews.com/news/202004081805296491 2020.12.22.인출.

경남도는 정부의 긴급재난지원금 지급계획이 확정되면 '경남형 긴급재난지원금'추진계획을 보완해 지급할 것이며 정부 지원금과 중복되지 않도록 하겠다는 입장을 밝혔다. 김경수 도지사는 1일 긴급재난지원금과 관련해 "정부가 소득하위 70% 이내 전국 1400만 가구에 최대 100만원(4인 가구 기준)을 지원하는 긴급재난지원금을 지급키로 한 것은 코로나19로 어려움을 겪고 있는 지역과 현장의 목소리를 반영한 결과로 환영하지만 고소득층을 제외한 보편적 재난기본소득인 긴급재난소득 도입이 실현되지 않은 것은 아쉬움이 남는다"고 말했다.

… 도는 중앙정부가 지급 예정인 긴급재난지원금에 앞서 도 예산으로 중위소득 100%이하 48만 3000가구에 최대 50만원(5인

가구 이상)까지 긴급재난소득을 우선 지원한다. 중위소득 50% 이하는 정부의 3월 추경으로 대상과 지원액이 이미 확정돼 이달 중 지급이 시작될 예정이다. 결과적으로 1차 차상위 계층 이하는 정부의 1차 추경으로 우선 지원하고, 2차 중위소득 100% 이하는 경남형 지급재난지원금, 3차는 중앙정부의 긴급재난지원금이 정부 2차 추경을 거쳐 지급될 전망이다.

자료: 이준희(2020.04.02.). "긴급재난지원금, 정부-경남 중복 안 되게 할 것". 경남신문. http://www.knnews.co.kr/news/articleView.php?idxno=1322568 2020.12.22.인출.

이외 관련 뉴스기사의 예는 〈표 2-5〉와 같다.

〈표 2-5〉 '경제', '지원', '재난' 관련 뉴스기사의 예

키워드	관련 뉴스기사
경제	· 이환주(2020.04.17.). 코로나發 실업대란… 고용노동 전문가 "무급휴직 근로자에 실업급여 지원 검토해야". 파이낸셜뉴스. https://www.fnnews.com/news/202004161823248365 2020.12.22.인출. · 박광연(2020.05.04.). 4월 소비자물가 0.1% 상승…코로나19·유가 하락으로 올해 첫 0%대. 경향신문. http://biz.khan.co.kr/khan_art_view.html?artid=202005040929001&code=920100 2020.12.22.인출. · 신주희(2020.06.24.). "졸업해도 취직을 못해요"… 특성화고 학생들 코로나 때문에 울상. 헤럴드경제. http://biz.heraldcorp.com/view.php?ud=20200611000102 2020.12.22.인출. · 이종규(2020.07.20.). "기본법은 사회적 경제의 든든한 밑돌…올해 안 제정돼야". http://www.hani.co.kr/arti/economy/economy_general/954318.html 한겨레. 2020.12.22. 인출. · 한영준(2020.11.10.). "퇴직시 실업급여 주듯 가게 문 닫았을 때 '폐업급여' 줘야" [인터뷰]. 파이낸셜뉴스. https://www.fnnews.com/news/202011091737017961 2020.12.22.인출.

키워드	관련 뉴스기사
지원	· 정상균(2020.04.09.). 정세균 “고소득자 환급 전제, 재난지원금 전국민 지급 가능”. 파이낸셜뉴스. https://www.fnnews.com/news/202004081805296491 2020.12.22.인출. · 임진흥(2020.05.22.), 의왕시, 코로나로 지친 정신건강복지센터 회원들의 마음 치유. 경기일보. http://www.kyeonggi.com/news/articleView.html?idxno=2287056 2020.12.22.인출. · 김창배(2020.06.30.). 울산시, 취약계층 맞춤형 ‘폭염’ 지원책 추진. 한국일보. https://www.hankookilbo.com/News/Read/A202006300823000028 2020.12.22.인출. · 이민영(2020.07.28.). 용산, 이태원·소상공인 경영안전 자금 50억원 지원. 서울신문. http://go.seoul.co.kr/news/newsView.php?id=20200728014002 2020.12.22.인출. · 이유범(2020.09.30.). 아동돌봄지원금 20만원 지급 완료.. 추석연휴 끝나면 중학생 지원 시작. 파이낸셜뉴스. https://www.fnnews.com/news/202009291503210348 2020.12.22.인출.
재난	· 이준희(2020.04.02.). “긴급재난지원금, 정부-경남 중복 안 되게 할 것”. 경남신문. http://www.knnews.co.kr/news/articleView.php?idxno=1322568 2020.12.22.인출. · 원선영(2020.04.10.). “재난지원금 하위 70%” 30세 미만 55.8% `최다'. 강원일보. http://www.kwnews.co.kr/nView.asp?s=101&aid=220040900146 2020.12.22.인출. · 이석주(2020.04.16.). 태풍 땐 어쩌나…부산 소상공인 지원금 70%가 재난재해기금. 국제신문. http://www.kookje.co.kr/news2011/asp/newsbody.asp?code=0200&key=20200416.22013006157 2020.12.22.인출. · 민현배(2020.04.22.). 재난지원금 선불카드 한도 50만→300만 원…지급 신속 지원. 경기일보. http://www.kyeonggi.com/news/articleView.html?idxno=2273679 2020.12.22.인출. · 임동우(2020.04.28.). “이주민도 노숙인도 차별마라” 보편적 재난지원금 촉구. 국제신문. http://www.kookje.co.kr/news2011/asp/newsbody.asp?code=0300&key=20200428.22010010863 2020.12.22.인출.

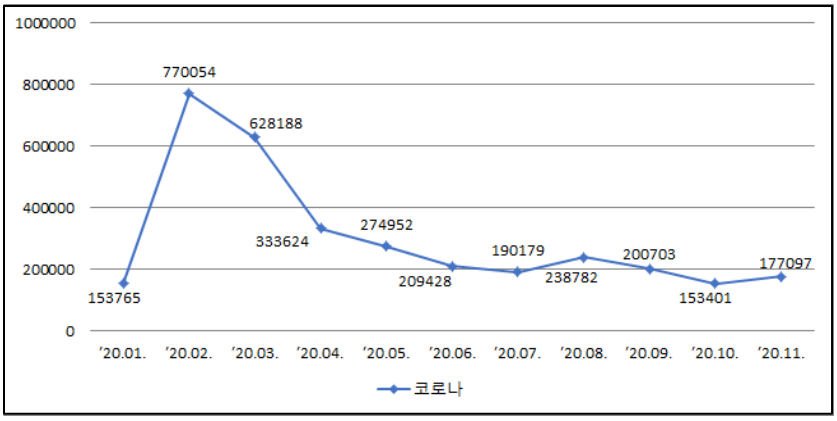
제2절 보건·복지·사회보장 주요 키워드 월별 트렌드 분석

1. 보건·복지·사회보장 주요 키워드 월별 분석

가. 월별 '코로나' 추출 건 수

〔그림 2-1〕, 〔그림 2-2〕에서 알 수 있듯이 '마스크'와 함께 2, 3월에 최대치로 나타났으며, 3월을 기점으로 감소하는 추세이다.

〔그림 2-1〕 월별 '코로나' 추출 건 수



'코로나' 관련 뉴스기사는 〈표 2-6〉과 같다.

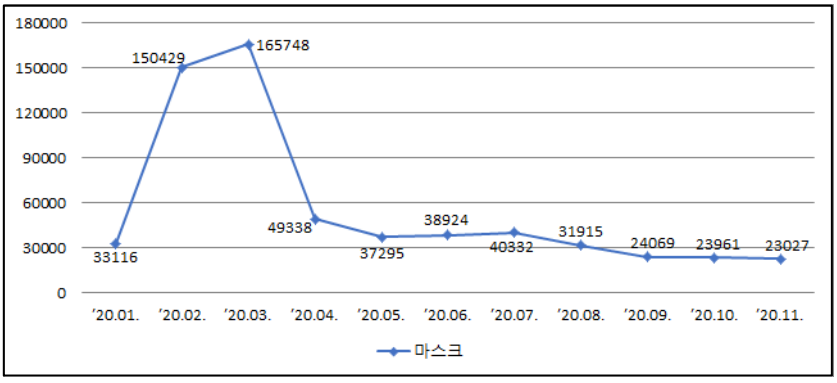
〈표 2-6〉 '코로나' 관련 뉴스기사 예

키워드	관련 뉴스기사
코로나	· 김윤구(2020.01.09.). 중국 원인불명 폐렴, '신종 코로나바이러스' 잠정 판정. 연합뉴스. https://www.yna.co.kr/view/AKR20200109072900083 2020.12.22.인출. · 송진식(2020.02.10.) ['신종 코로나' 확산불안한 학교들 '봄방학'도 앞당긴다. 경향신문. http://news.khan.co.kr/kh_news/khan_art_view.html?artid=202002101743001&code=940401&utm_campaign=zum_news&utm_source=zum&utm_medium=related_news 2020.12.22.인출. · 이유범(2020.03.17.) 유·초·중·고 개학 2주 더 미뤄… 긴급돌봄·온라인 수업 강화. 파이낸셜 뉴스. https://www.fnnews.com/news/202003171811278620 2020.12.22.인출. · 김동일(2020.08.06.). 코로나로 의정부 위기가구 지원대상 급증… 지난해 2배 예산확보 비상. 경기일보. https://www.kyeonggi.com/news/articleView.html?idxno=2309636 2020.12.22.인출. · 박현영(2020.08.07.). 美, 한국 여행금지 5개월만에 풀었다…1단계 내려 여행재고. 중앙일보. https://news.joins.com/article/23843226 2020.12.22.인출.

나. 월별 '마스크' 추출 건 수

〔그림 2-1〕, 〔그림 2-2〕에서 알 수 있듯이 '코로나'와 함께 2, 3월에 최대치로 나타났으며, 역시 3월을 기점으로 급격히 감소하였다.

〔그림 2-2〕 월별 '마스크' 추출 건 수



'마스크' 관련 뉴스기사는 〈표 2-7〉과 같다.

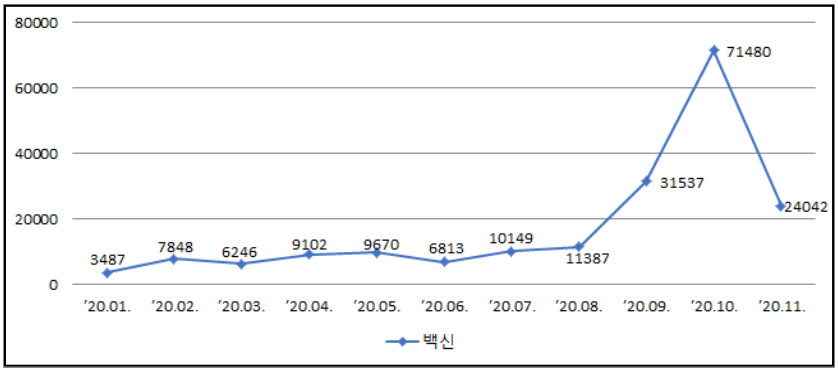
〈표 2-7〉 '마스크' 관련 뉴스기사 예

키워드	관련 뉴스기사
마스크	· 김용현(2020.02.05.). 마스크·손소독제 매점매석 형사처벌 받는다. 제민일보. http://www.jemin.com/news/articleView.html?idxno=640402 2020.12.22.인출. · 한상훈(2020.02.17.). 광주시, 신종 코로나 확산 방지 마스크 배부. 경기일보. http://www.kyeonggi.com/news/articleView.html?idxno=2241579 2020.12.22.인출. · 김성호(2020.02.29.). 정부, 마스크 매일 500만개씩 공적판매. 파이낸셜뉴스. https://www.fnnews.com/news/202002281810002388 2020.12.22.인출. · 윤홍집(2020.03.16.). "마스크 써도 불안, 안써도 불안"… 사각지대에 놓인 돌봄 노동자. 파이낸셜뉴스. https://www.fnnews.com/news/202003151733149169 2020.12.22.인출. · 임병선(2020.03.27.). 미국 코로나19 환자 세계 1위, 곳곳에서 의료장비 대란. 서울신문. http://www.seoul.co.kr/news/newsView.php?id=20200327500006 2020.12.22.인출.

다. 월별 '백신' 추출 건 수

〔그림 2-3〕 과 〔그림 2-4〕에 제시된 바와 같이 '접종'과 더불어 10월에 최대치로 나타났다.

〔그림 2-3〕 월별 '백신' 추출 건 수



'백신' 관련 뉴스기사는 〈표 2-8〉과 같다.

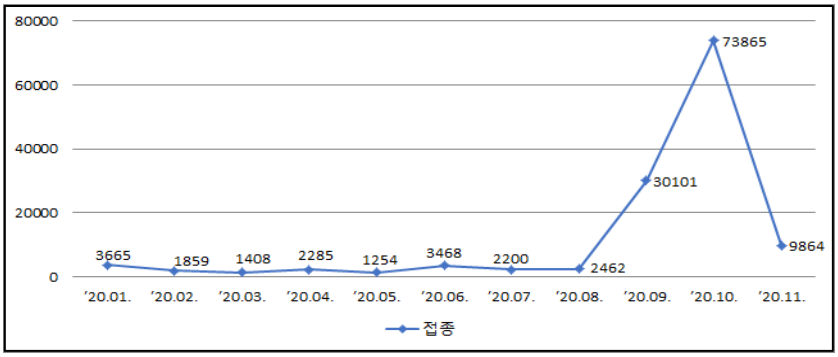
〈표 2-8〉 '백신' 관련 뉴스기사 예

키워드	관련 뉴스기사
백신	· 임소형(2020.10.01.). "코로나19 백신, 만드는 방식 이렇게 다르네" 따져봐야 안전하다. 한국일보. https://www.hankookilbo.com/News/Read/A2020092317320000134 2020.12.22.인출. · 강국진(2020.10.06.). 국가예방접종 사업 백신 3년간 4만명분 폐기. 서울신문. http://www.seoul.co.kr/news/newsView.php?id=20201006012022 2020.12.22.인출. · 임보미(2020.10.15.). 일라이릴리도 항체치료제 3상 시험 중단. 동아일보. https://www.donga.com/news/article/all/20201015/103426142/1 2020.12.22.인출. · 최선을(2020.10.24.). "사망자 36명"…'독감접종 계속' 방침에도 곳곳서 혼선(종합). 서울신문. http://www.seoul.co.kr/news/newsView.php?id=20201024500002 2020.12.22.인출. · 박경은(2020.10.31.). 독감백신 접종 후 사망자 83명 중 72명 '무관'…"접종 계속". 아주경제. https://www.ajunews.com/view/20201031161141137 2020.12.22.인출.

라. 월별 '접종' 키워드의 추출 건 수

〔그림 2-3〕 과 〔그림 2-4〕에 제시된 바와 같이 '백신'과 더불어 10월에 최대치로 나타났다.

〔그림 2-4〕 월별 '접종' 추출 건 수



‘접종’ 관련 뉴스기사는 〈표 2-9〉와 같다.

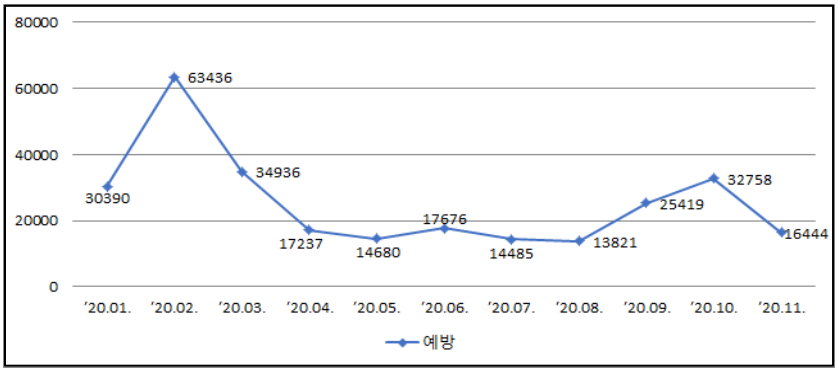
〈표 2-9〉 ‘접종’ 관련 뉴스기사 예

키워드	관련 뉴스기사
접종	· 조현의(2020.10.01.). ‘상온노출 백신’ 부작용 있나 없나. 아시아경제. https://www.asiae.co.kr/article/2020092912125804231 2020.12.22.인출. · 이무헌, 김천열(2020.10.06.). 청소년〈만 13~18세〉 8만3천명 독감백신 못 맞아…‘골든타임 놓칠라’ 전전긍긍. 강원일보. http://www.kwnews.co.kr/nView.asp?s=501&aid=220100500070 2020.12.22.인출. · 홍수민(2020.10.15.). 러시아, 두번째 코로나 백신도 공식 승인…두달만에 자체 개발. 중앙일보. https://news.joins.com/article/23894704 2020.12.22.인출. · 정자연(2020.10.24.). 독감백신 접종 후 사망자 일주일 동안 32명…경기도 3명. 경기일보. http://www.kyeonggi.com/news/articleView.html?idxno=2324270 2020.12.22.인출. · 강기준(2020.10.29.). 내년에 풀릴 백신 최대 160억회분…문제는 공급. 머니투데이 https://news.mt.co.kr/mtview.php?no=2020102812012731774&outlink=1&ref=%3A%2F%2F 2020.12.22.인출.

마. 월별 ‘예방’ 키워드의 추출 건 수

〔그림 2-5〕와 같이 2월에 최대치로 나타났으며, 이후 감소하다가 10월에 다시 증가하였다.

〔그림 2-5〕 월별 ‘예방’ 추출 건 수



'예방' 관련 뉴스기사는 〈표 -10〉과 같다.

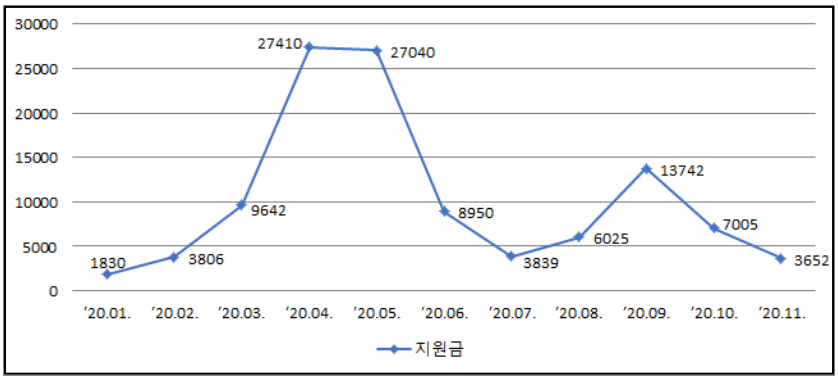
〈표 2-10〉 '예방' 관련 뉴스기사 예

키워드	관련 뉴스기사
예방	· 안태호(2020.01.29.). 안전관리 사각지대 '다중이용업소' 미리 관리한다. 파이낸셜뉴스. https://www.fnnews.com/news/202001281904524182 2020.12.22.인출. · 조건희, 이소정(2020.02.14.). 위기아동 3개기준 해당땐 이르면 4월 '우선 방문조사'. 동아일보. http://www.donga.com/news/article/all/20200214/96684144/1 2020.12.22.인출. · 박수진(2020.02.26.). 전남 지자체, 코로나19 차단·지역경제 활성화 '안간힘'. 전남일보. https://jnilbo.com/view/media/view?code=2020022519080495613 2020.12.22.인출. · 이춘봉(2020.03.04.). 보건·위생교육이 이손요양병원 2차 감염 막았다. 경상일보. http://www.ksilbo.co.kr/news/articleView.html?idxno=755095 2020.12.22.인출. · 배준용(2020.10.16.). 코로나, 손만 잘 씻어도 감염 확률 20% 줄어요. 조선일보. https://www.chosun.com/culture-life/health/2020/10/16/7AP76H6YEFB7LBMKZ6DKLWMX2E/?utm_source=bigkinds&utm_medium=original&utm_campaign=news 2020.12.22.인출.

바. 월별 '지원금' 키워드의 추출 건 수

〔그림 2-6〕, 〔그림 2-7〕, 〔그림 2-8〕에 나타난 바와 같이 '지원금', '지급', '재난' 모두 공통적으로 4월에 높은 편이며, '지원금'은 9월에 한 번 더 증가한 것으로 나타났다.

〔그림 2-6〕 월별 '지원금' 추출 건 수



'지원금' 관련 뉴스기사는 〈표 2-11〉과 같다.

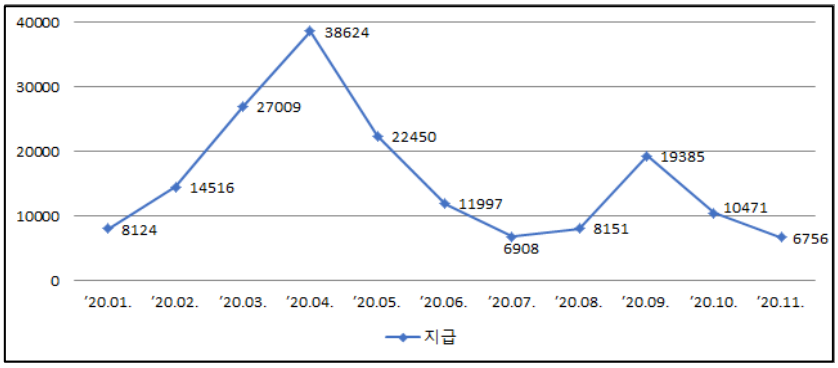
〈표 2-11〉 '지원금' 관련 뉴스기사 예

키워드	관련 뉴스기사
지원금	· 윤종현(2020.03.09.). 고용유지지원금 2주 만에 신청 13배 급증. 강원일보. http://www.kwnews.co.kr/nView.asp?s=401&aid=220030800034 2020.12.22.인출. · 김병철(2020.04.09.). 이재명 "경기도 재난기본소득, 18개 시군과 동시 지급"… 오늘부터 접수. 서울신문. http://go.seoul.co.kr/news/newsView.php?id=20200409014014 2020.12.22.인출. · 김태준(2020.04.09.). "애 맡길 곳 없다" 아우성에…가족돌봄휴가 5일→10일로 확대하고 지원금 50만원까지. 매일경제. https://www.mk.co.kr/news/economy/view/2020/04/371319/ 2020.12.22.인출. · 이현(2020.05.17.). 충주시 긴급재난지원금 집중 신청…지역경제 활성화 박차. 충청일보. http://www.ccdailynews.com/news/articleView.html?idxno=1066827 2020.12.22.인출. · 안중현(2020.09.17.). 3연타 맞아도 재난지원금 소외… "농민은 국민 아입니까". 조선일보. https://www.chosun.com/economy/2020/09/17/2QCJYPDW4RANTFTE5HMK7LEWEE/?utm_source=bigkinds&utm_medium=original&utm_campaign=news 2020.12.22.인출.

사. 월별 '지급' 키워드의 추출 건 수

[그림 2-6], [그림 2-7], [그림 2-8]에 나타난 바와 같이 '지원금', '지급', '재난' 모두 공통적으로 4월에 높은 편이며, '지급' 역시 9월에 한 번 더 증가한 것으로 나타났다.

[그림 2-7] 월별 '지급' 추출 건 수



'지급' 관련 뉴스기사는 〈표 2-12〉와 같다.

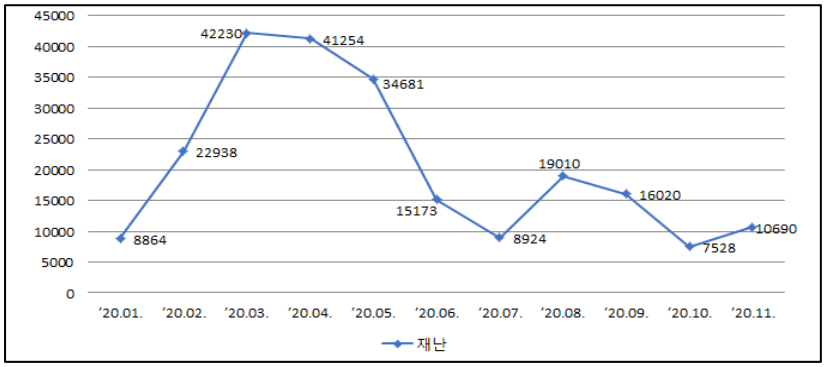
〈표 2-12〉 '지급' 관련 뉴스기사 예

키워드	관련 뉴스기사
지급	· 이환주(2020.03.25.). 문 대통령 "4대보험료·전기료 감면" 지시… 고민 깊어진 기재부·한전 [코로나 직격탄 맞은 경제]. 파이낸셜뉴스. https://www.fnnews.com/news/202003241729092167 2020.12.22.인출. · 남건우(2020.04.09.). 정세균 총리 "재난지원금 모두 준 뒤 고소득자 환수". 동아일보. http://www.donga.com/news/article/all/20200409/100567262/1 2020.12.22.인출. · 이권영(2020.04.21.). 충남 복지혜택 사각지대 없앤다. 충청투데이. http://www.cctoday.co.kr/news/articleView.html?idxno=2065161 2020.12.22.인출. · 김창수(2020.05.04.). 강화군, 전국 최초 소상공인 임차료 지원 확대. 경기일보. http://www.kyeonggi.com/news/articleView.html?idxno=2278806 2020.12.22.인출. · 조용휘(2020.09.18.). 부산 기장군 주민들 "해안지역을 특별재난지역으로 선포해달라". 동아일보. https://www.donga.com/news/article/all/20200917/102991291/1 2020.12.22.인출.

아. 월별 '재난' 키워드의 추출 건 수

[그림 2-6], [그림 2-7], [그림 2-8]에 나타난 바와 같이 '지원금', '지급', '재난' 모두 공통적으로 4월에 높은 편이며, '재난'의 경우 3월에 최대치로 나타났고 8월에 한 번 더 증가한 것으로 나타났다.

[그림 2-8] 월별 '재난' 추출 건 수



'재난' 관련 뉴스기사는 〈표 2-13〉과 같다.

〈표 2-13〉 '재난' 관련 뉴스기사 예

키워드	관련 뉴스기사
재난	· 이준희(2020.04.02.). "긴급재난지원금, 정부-경남 중복 안 되게 할 것". 경남신문. http://www.knnews.co.kr/news/articleView.php?idxno=1322568 2020.12.22.인출. · 원선영(2020.04.10.). "재난지원금 하위 70%" 30세 미만 55.8% `최다'. 강원일보. http://www.kwnews.co.kr/nView.asp?s=101&aid=220040900146 2020.12.22.인출. · 이석주(2020.04.16.). 태풍 땐 어쩌나…부산 소상공인 지원금 70%가 재난재해기금. 국제신문. http://www.kookje.co.kr/news2011/asp/newsbody.asp?code=0200&key=20200416.22013006157 2020.12.22.인출. · 민현배(2020.04.22.). 재난지원금 선불카드 한도 50만→300만 원…지급 신속 지원. 경기일보. http://www.kyeonggi.com/news/articleView.html?idxno=2273679 2020.12.22.인출. · 임동우(2020.04.28.). "이주민도 노숙인도 차별마라" 보편적 재난지원금 촉구. 국제신문. http://www.kookje.co.kr/news2011/asp/newsbody.asp?code=0300&key=20200428.22010010863 2020.12.22.인출.

2. 보건·복지·사회보장 월별 키워드 분석

월별 새로운 이슈로 나타난 키워드들을 살펴보기기 위해 보건복지 및 사회보장 관련 키워드를 월별 빈도수 2,500 이하 순으로 정리한 결과는 〈표 2-14〉와 같다.

〈표 2-14〉 월별 키워드 목록(빈도수 2,500이하 기준)

월	키워드
1	이국종, 성형외과, 쪽방촌, 영아, 의료원장, 청각, 길병원, 보라매병원, 발생지, 상해, 목감기, 열병, 소아마비, 임금체계, 처방전, 수어
2	수본, 안암병원, 항바이러스제, 조선대병원, 엑스레이, 고대안암병원, 유행병, 간병, 국군수도병원, 무사증, 상담원, 치료법, 보건성, 이비인후과, 휴무, 한림대, 전염력, 징후, 구로병원, 긴급회의, 중증도, 이식, 완쾌
3	결식, 비타민, 약사, 파티마병원
4	안내견, 흡연자, 실직자, 잠실종합운동장, 자선, 공황, 체류자, 형평
5	소수자, 보험법, 구직자, 양성자, 수술실, 생존자, 수혜, 사용처, 모금액, 강남병원, 입양인, 인권침해, 양극화, 부담액, 보건의, 특수학교
6	출혈, 장출혈, 깁스, 구강, 합병증, 복통, 구균, 서울아산병원, 일본뇌염
7	흑사병, 수돗물, 실버, 페스트, 탈북민, 디프테리아, 양육비, 투약, 교보생명, 병변, 악성, 피로, 복지과
8	수재민, 주택자, 변이, 인권위, 분만실, 보험금
9	건강식품, 병가, 의대생, 세브란스, 우울감, 기독병원, 후유증
10	미혼모, 한국건강관리협회, 항원, 건보, 생리대, 발암, 독성, 주사기, 효능, 심혈관, 주치의, 자궁, 수액, 한국보건산업진흥원, 의학상, 메디컬
11	겨울나기, 항생제, 의약국, 미혼, 점자, 한의원, 뇌혈관, 난방비, 신경외과, 근로기준법, 보험사

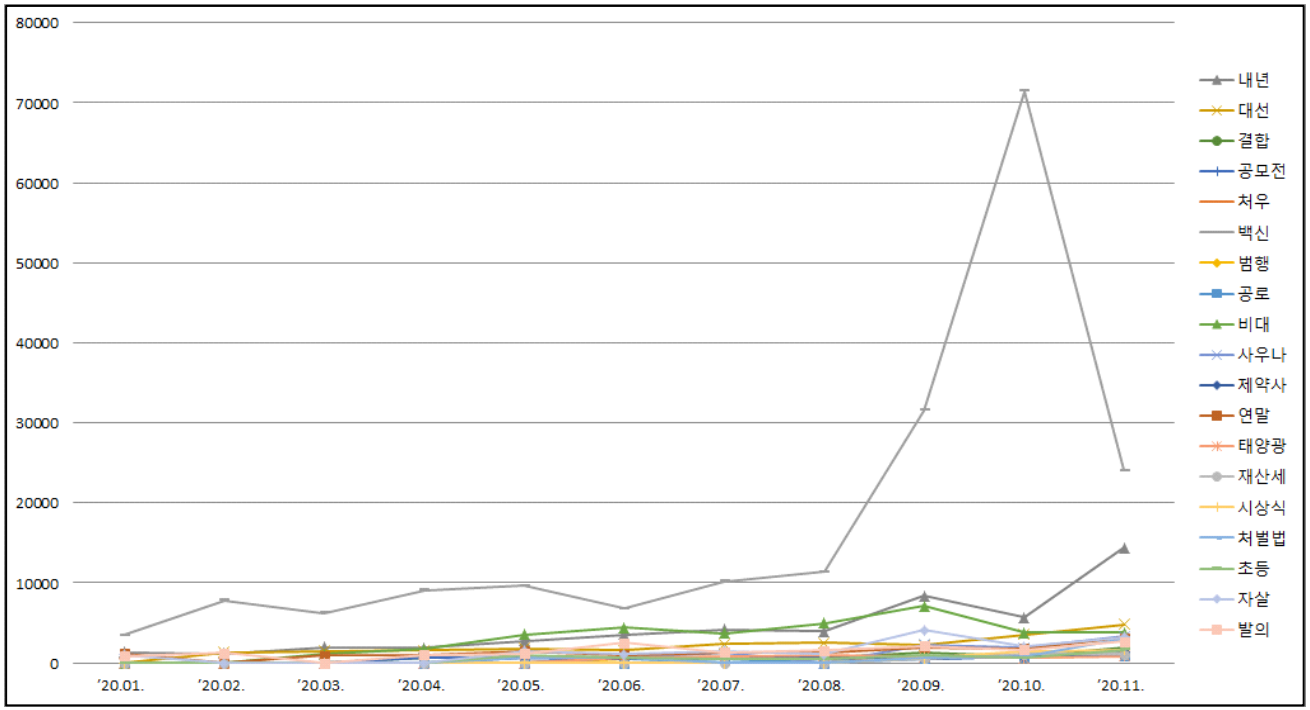
제3절 보건·복지·사회보장 키워드 간 순위 비교

여기에서는 상반기에 집중된 이슈, 하반가에 집중된 이슈를 살펴보기 위해, 시간의 흐름과 키워드 빈도수 간 상관계수를 구하고 그에 따른 키워드 간 순위를 비교해 보았다. 하반기에 집중된 이슈를 알아보기 위해 상관관계는 비모수 검정인 kendall-tau 기준으로 상위 20여개 키워드를 제시하였고 그에 따른 그래프는 〈표 2-15〉와 〔그림 2-9〕에 제시하였다. 구체적으로, '내년, 대선, 결합, 공모전, 처우, 백신, 범행, 공로, 비대, 사우나, 제약사, 연말, 태양광, 재산세, 시상식, 처벌법, 초등, 자살, 발의' 순으로 나타났다. 한편, 상반기에 집중된 이슈로는 kendall-tau 계수 기준 하위 20개 키워드와 그에 따른 그래프는 〈표 2-16〉과 〔그림 2-10〕에 제시하였다. 구체적으로, '발병, 체온, 유증, 소독제, 본부장, 감시, 전염병, 증상, 메인, 양호, 초기, 전염, 지방자치단체, 조기, 검역, 박종일, 발열, 지목, 지정, 체제' 순으로 나타났다.

〈표 2-15〉 kendall-tau 기준 상위 20여개 키워드

순위	키워드	kendalltau	total	'20.01.	'20.02.	'20.03.	'20.04.	'20.05.	'20.06.	'20.07.	'20.08.	'20.09.	'20.10.	'20.11.
1	내년	0.854545	49103	1303	1186	1951	1904	2759	3558	4140	3960	8339	5637	14366
2	대선	0.854545	23151	0	1353	1501	1563	1771	1528	2383	2487	2216	3547	4802
3	결합	0.785355	7597	0	0	0	606	806	651	688	664	1357	856	1969
4	공모전	0.785355	6177	0	0	0	625	670	555	621	876	873	901	1056
5	처우	0.783929	4365	0	0	0	0	0	580	610	523	601	956	1095
6	백신	0.781818	191761	3487	7848	6246	9102	9670	6813	10149	11387	31537	71480	24042
7	범행	0.743728	3985	0	0	0	0	0	535	688	581	637	587	957
8	공로	0.739996	2614	0	0	0	0	0	0	494	0	521	752	847
9	비대	0.733976	34130	0	0	1090	1728	3500	4420	3707	4961	7126	3792	3806
10	사우나	0.724882	9292	0	0	0	0	0	614	1239	0	2274	1849	3316
11	제약사	0.712726	4961	0	0	0	0	573	556	695	583	512	749	1293
12	연말	0.709091	15152	1179	0	943	1011	1496	1037	1272	1305	1953	1822	3134
13	태양광	0.703526	4439	0	0	0	0	0	571	751	962	759	578	818
14	재산세	0.700649	2714	0	0	0	0	0	0	0	0	513	1078	1123
15	시상식	0.700649	3697	0	0	0	0	0	0	0	0	620	1416	1661
16	처벌법	0.682242	5688	0	0	0	0	572	574	0	0	658	846	3038
17	초등	0.6742	5722	0	0	0	0	892	626	483	528	895	628	1670
18	자살	0.673162	15484	988	0	0	0	1457	1090	1474	1121	4057	2043	3254
19	발의	0.672727	16109	823	1272	0	1039	1168	2518	1330	1543	2102	1653	2661

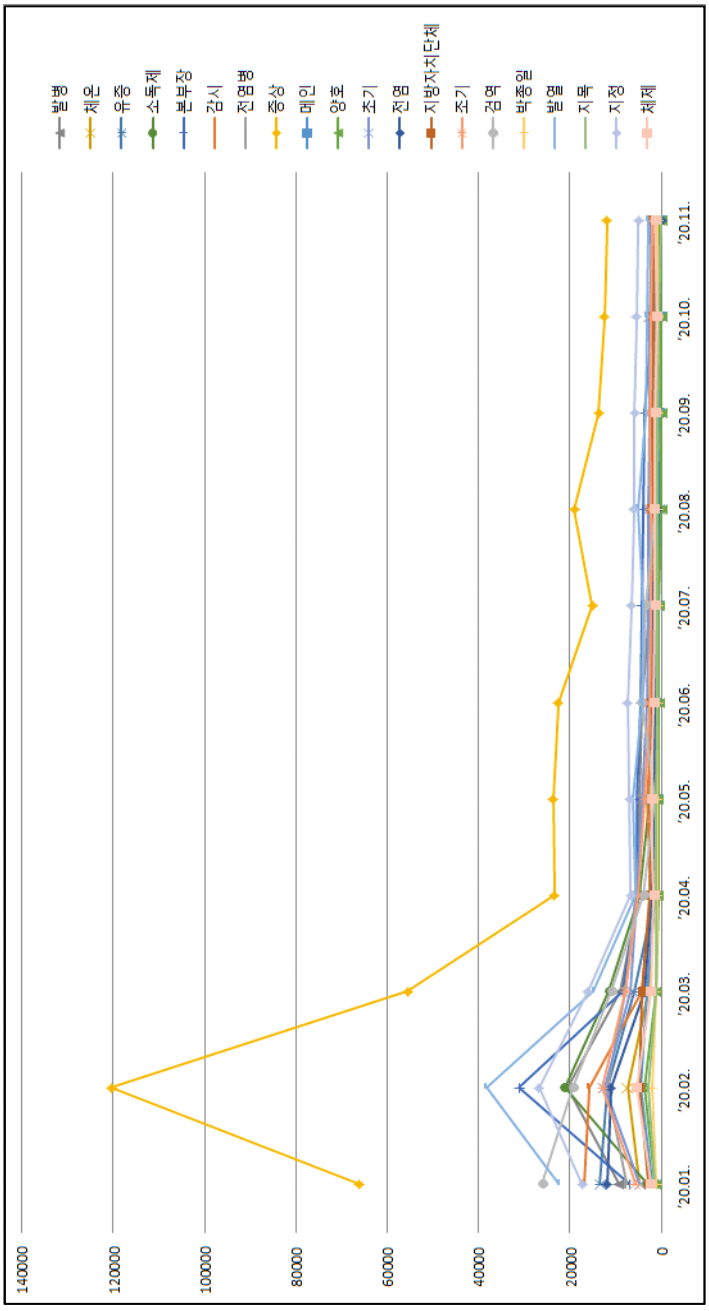
[그림 2-9] kendall-tau 기준 상위 20여개 키워드



〈표 2-16〉 kendall-tau 기준 하위 20개 키워드

순위	키워드	kendalltau	total	'20.01.	'20.02.	'20.03.	'20.04.	'20.05.	'20.06.	'20.07.	'20.08.	'20.09.	'20.10.	'20.11.
1	발병	-0.92727	63325	9329	20629	9042	4903	4765	4168	2535	2953	1803	1690	1508
2	체온	-0.92727	22164	4684	7421	2837	1336	1180	1028	810	1021	710	642	495
3	유증	-0.89755	36711	13471	11761	5969	1565	906	1768	722	549	0	0	0
4	소독제	-0.89091	51263	3220	21195	11418	4865	2504	2447	1692	1357	1197	758	610
5	본부장	-0.89091	77770	6917	31159	8332	5297	5235	4514	4265	4051	3680	2047	2273
6	감시	-0.89091	50358	16798	15776	4176	2244	3175	1834	1578	1219	1703	1074	781
7	전염병	-0.89091	48307	7738	11774	8013	5559	3834	2032	2944	1808	1731	1469	1405
8	증상	-0.89091	384536	66197	120538	55642	23430	23732	22544	15147	19126	13767	12515	11898
9	메인	-0.88077	16694	1900	4499	3185	1462	1367	1305	1125	1302	549	0	0
10	양호	-0.86683	8820	1400	3773	1261	612	681	569	524	0	0	0	0
11	초기	-0.85455	49071	4689	12417	6664	5226	4375	2991	2805	2769	2484	2667	1984
12	전염	-0.85455	35985	11849	11113	3660	1900	1557	1549	1955	863	899	640	0
13	지방자치단체	-0.85455	28304	2723	4809	4135	2602	2243	2056	2143	2066	1974	1827	1726
14	조기	-0.85455	49567	5752	12867	7864	5031	3628	2667	2397	2326	2645	2294	2096
15	검역	-0.85455	73915	26114	19352	10923	3932	1538	4333	3379	1485	1133	855	871
16	박종일	-0.85455	10815	1047	1995	1272	951	879	858	833	775	827	604	774
17	발열	-0.85455	110438	22402	38344	14869	5598	6204	4357	3955	5216	3398	3243	2852
18	지목	-0.85455	11848	1293	2725	1285	1056	1221	914	993	918	815	628	0
19	지정	-0.85455	110133	17421	26740	16170	6667	6947	7352	6555	5966	5884	5399	5032
20	체제	-0.85455	22224	2395	5413	2373	1624	2189	1585	1426	1436	1422	989	1372

[그림 2-10] kendall-tau 기준 하위 20개 키워드



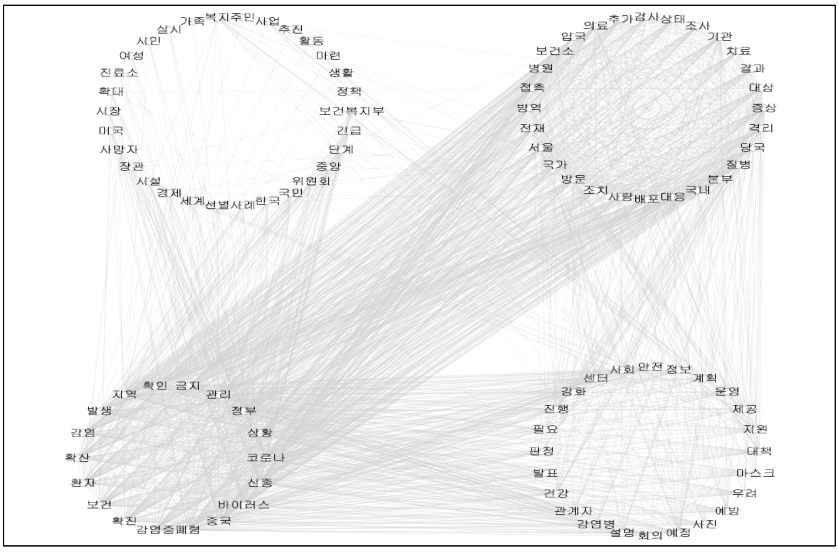
제3장

보건·복지·사회보장 키워드 클러스터링 분석

제 3 장 보건·복지·사회보장 키워드 클러스터링 분석

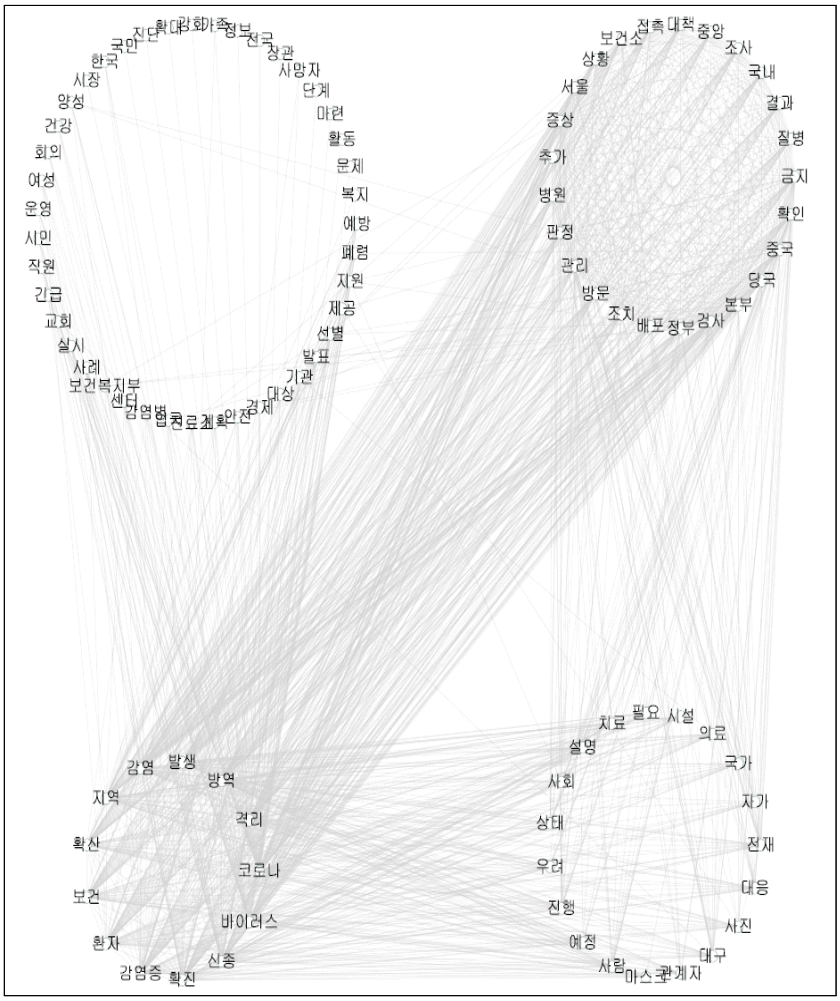
본 장에서는 보건·복지·사회보장 관련 키워드의 월별 클러스터링 분석을 실시하였다. K-means 방법을 활용하였고, 제일 적합하다고 판단되는 4개의 클러스터를 만들어 시각화하였다. 각 월별 클러스터 명칭은 연구진이 세부 키워드를 검토하고 정하였다.

〔그림 3-1〕 1월 보건·복지·사회보장 키워드 클러스터링 결과



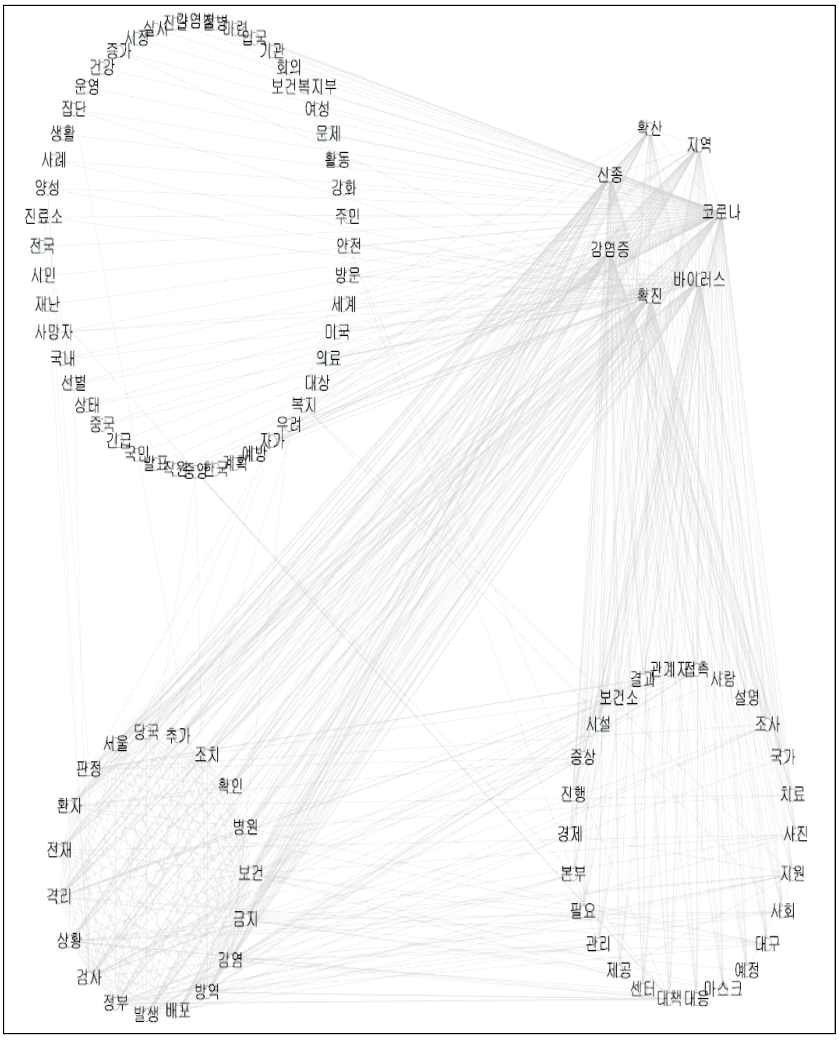
1월 보건·복지·사회보장 키워드 클러스터의 명칭은 좌측 상단부터 시계방향으로 코로나 관련 국내외 정세, 코로나 확진 관련 조치, 코로나 지원 대책, 코로나 환자 확진이다.

[그림 3-2] 2월 보건·복지·사회보장 키워드 클러스터링 결과



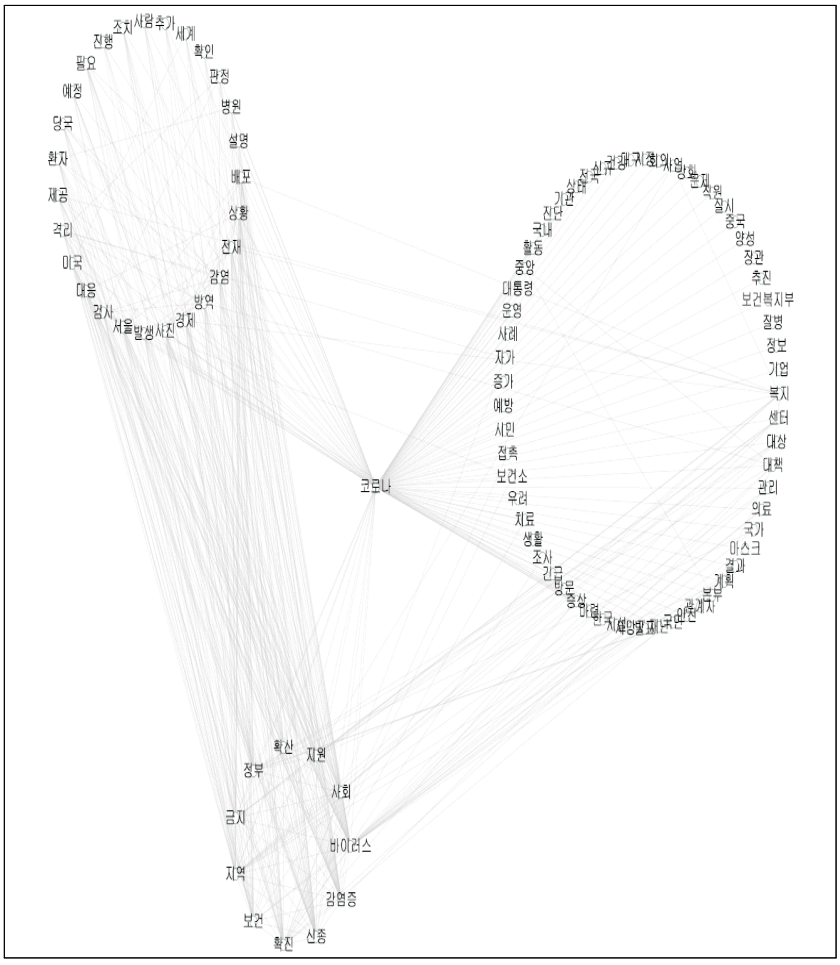
2월 보건·복지·사회보장 키워드 클러스터의 명칭은 좌측 상단부터 시계방향으로 코로나에 대한 정부의 대처, 코로나 확진 관련 조치, 코로나 대응, 코로나 환자 확진이다.

[그림 3-3] 3월 보건·복지·사회보장 키워드 클러스터링 결과



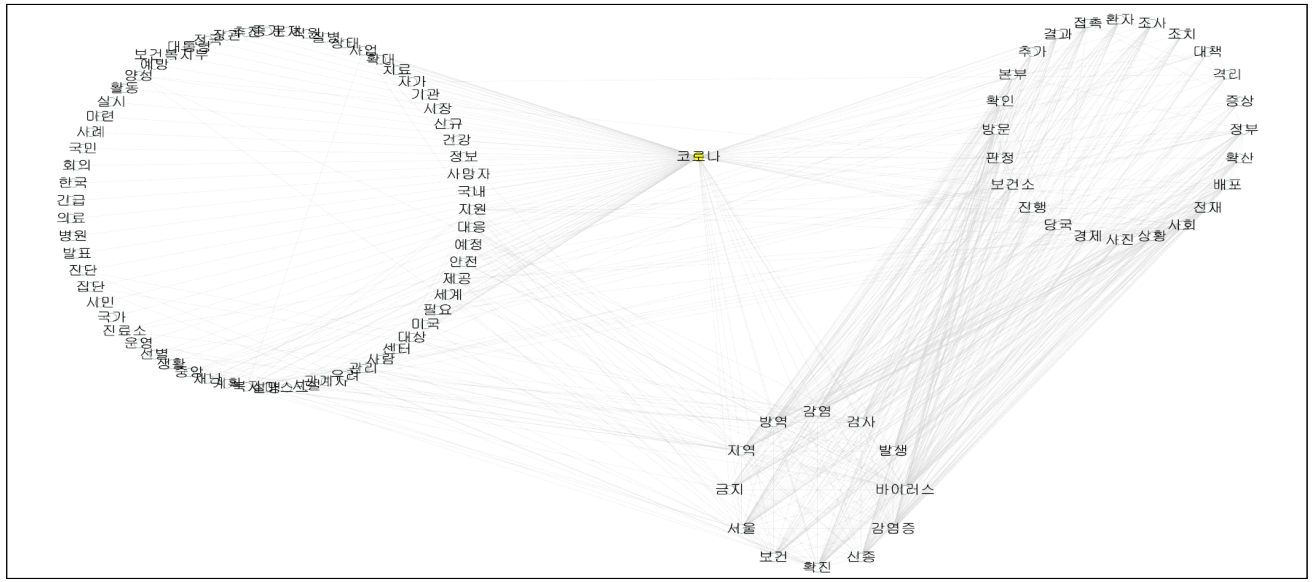
3월 보건·복지·사회보장 키워드 클러스터의 명칭은 좌측 상단부터 시계방향으로 국내외 코로나 관련 이슈, 코로나 지역사회 확산, 코로나 대응, 정부 방역 조치이다.

[그림 3-4] 4월 보건·복지·사회보장 키워드 클러스터링 결과



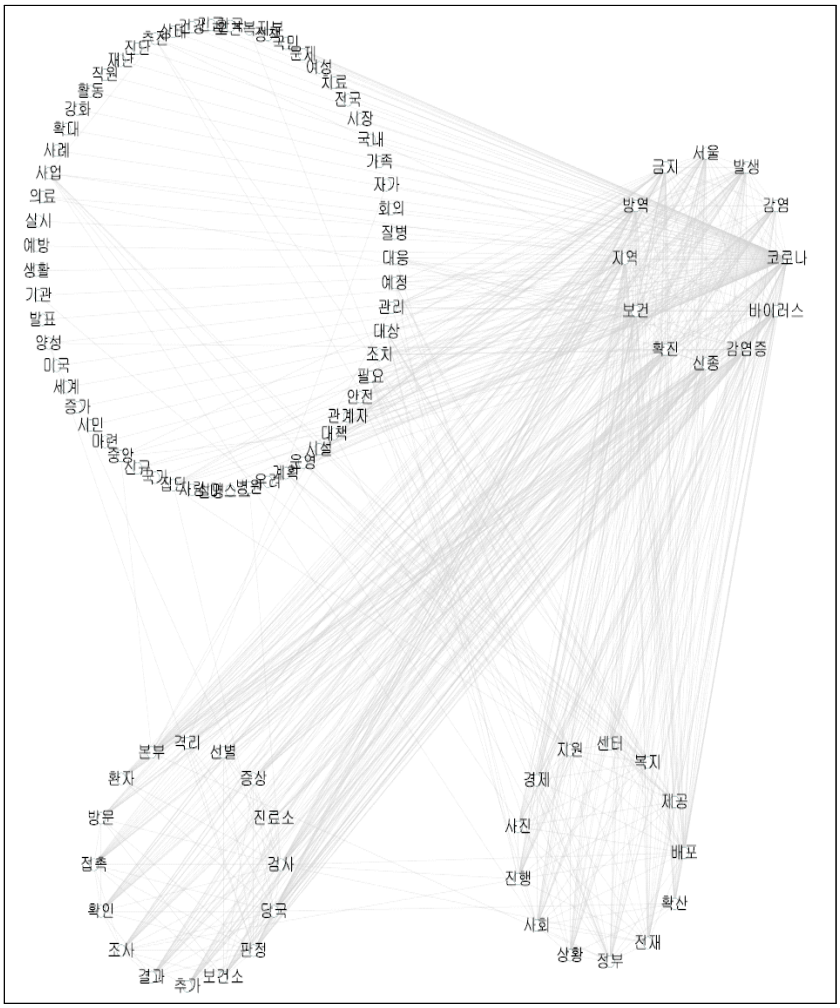
4월 보건·복지·사회보장 키워드 클러스터의 명칭은 다음과 같다. 중심에 위치한 클러스터는 코로나이며, 좌측 상단부터 시계방향으로 코로나 확진 관련 조치, 정부 대응 및 국내외 상황, 코로나 지역사회 확산 및 확진이다.

[그림 3-5] 5월 보건·복지·사회보장 키워드 클러스터링 결과



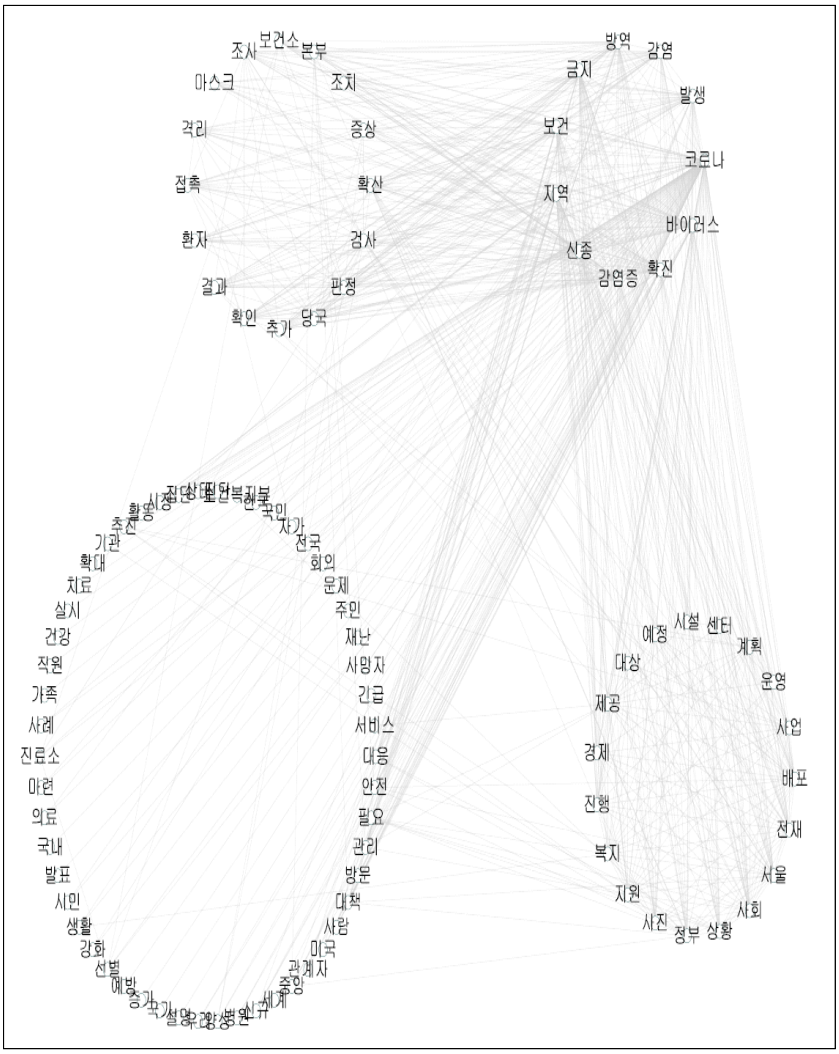
5월 보건·복지·사회보장 키워드 클러스터의 명칭은 다음과 같다. 중심에 위치한 클러스터는 코로나이며, 좌측 상단부터 시계방향으로 코로나 관련 기타 상황, 코로나 확진자에 대한 보건당국의 조치, 코로나 지역사회 확산이다.

[그림 3-6] 6월 보건·복지·사회보장 키워드 클러스터링 결과



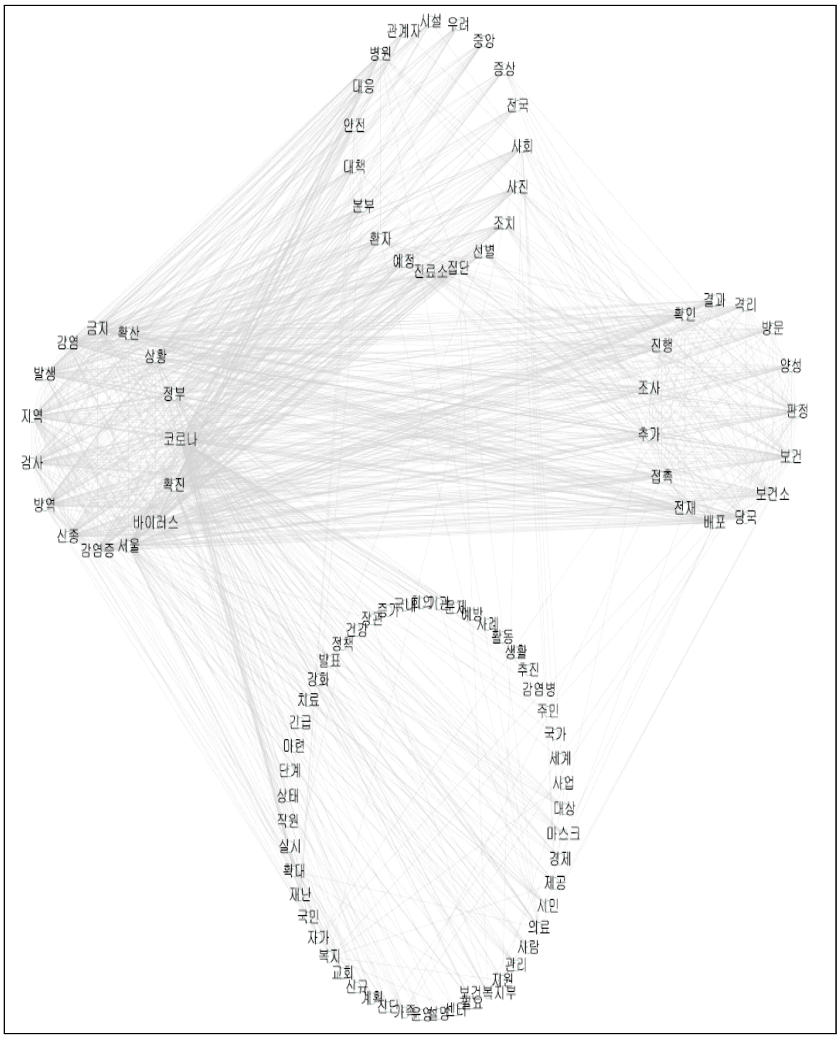
6월 보건·복지·사회보장 키워드 클러스터의 명칭은 좌측 상단부터 시계방향으로 코로나 관련 국내외 상황, 코로나 확진 관련 이슈, 정부 지원 상황, 코로나 확진자 관련 대응이다.

〔그림 3-7〕 7월 보건·복지·사회보장 키워드 클러스터링 결과



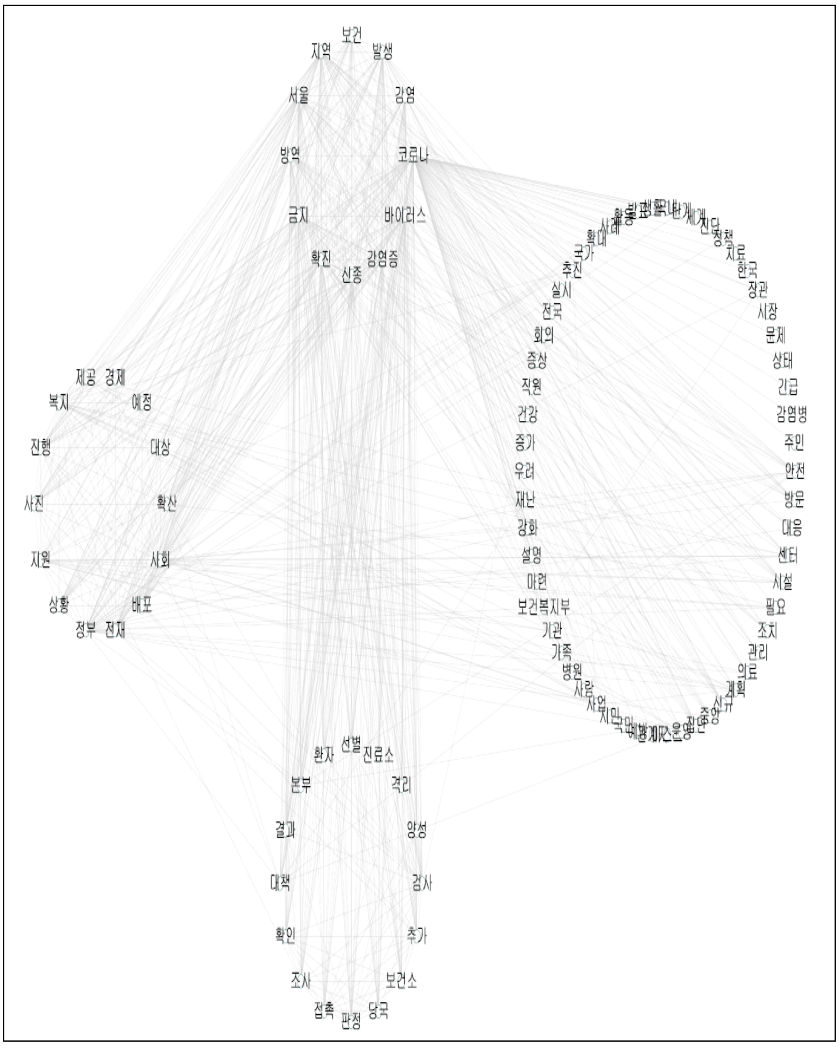
7월 보건·복지·사회보장 키워드 클러스터의 명칭은 좌측 상단부터 시계방향으로 코로나 확산 대응 관련 이슈, 코로나 확진 관련 이슈, 코로나 관련 정부 지원, 국내외 정세이다.

[그림 3-8] 8월 보건·복지·사회보장 키워드 클러스터링 결과



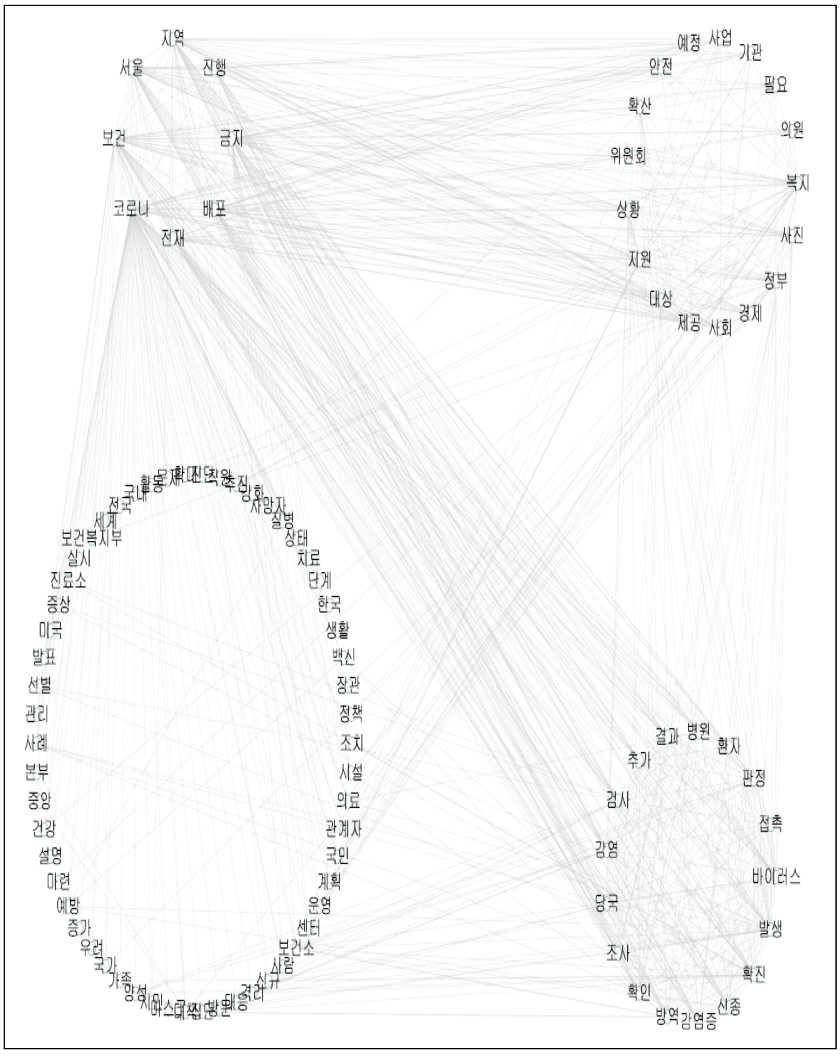
8월 보건·복지·사회보장 키워드 클러스터의 명칭은 좌측부터 시계방향으로 코로나 지역사회 확산, 코로나 사전안전 대책, 코로나 확진자 판정 이슈, 그 외 국내외 이슈이다.

〔그림 3-9〕 9월 보건·복지·사회보장 키워드 클러스터링 결과



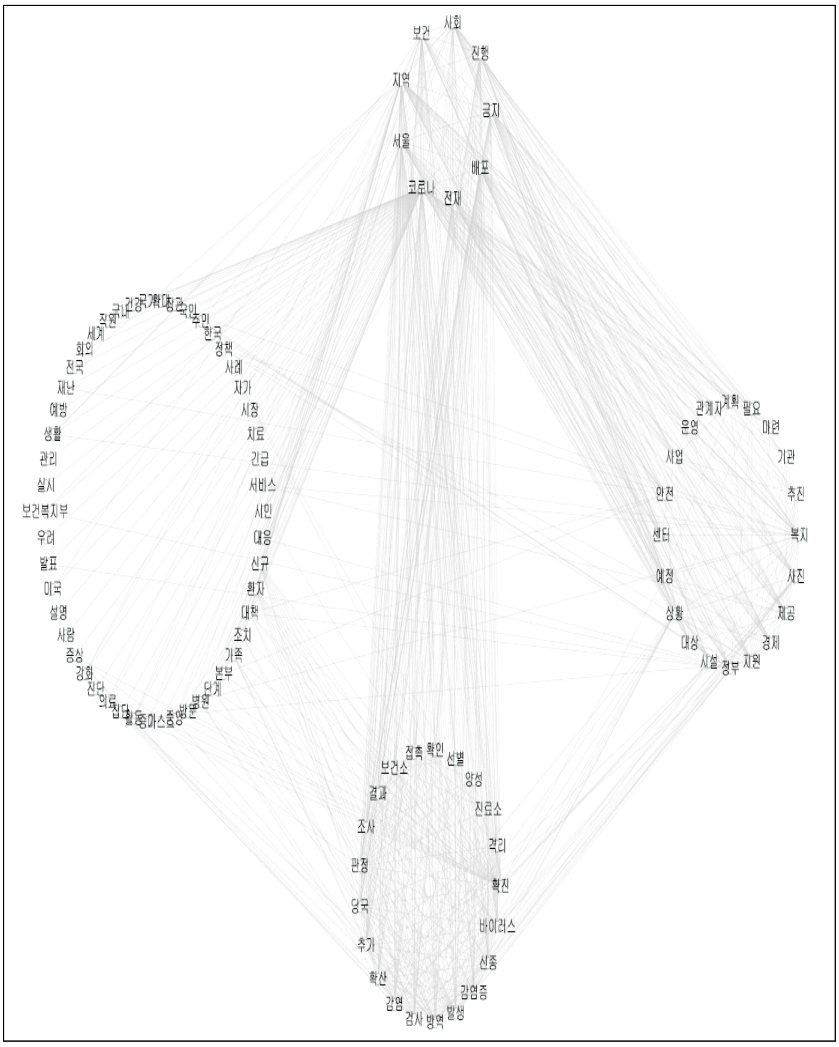
9월 보건·복지·사회보장 키워드 클러스터의 명칭은 좌측부터 시계방향으로 정부 지원 상황, 코로나 지역사회 확산 방지, 기타, 코로나 확진자 관련 이슈이다.

〔그림 3-10〕 10월 보건·복지·사회보장 키워드 클러스터링 결과



10월 보건·복지·사회보장 키워드 클러스터의 명칭은 좌측 상단부터 시계방향으로 코로나 지역사회 확산 방지, 코로나 대응 정부 지원, 코로나 확진 관련 이슈, 그 외 이슈이다.

〔그림 3-11〕 11월 보건·복지·사회보장 키워드 클러스터링 결과



11월 보건·복지·사회보장 키워드 클러스터의 명칭은 상단부터 시계방향으로 코로나 지역사회 확산 방지, 코로나 관련 정부 지원 상황 및 대책, 코로나 확진자 이슈, 기타이다.

제4장

비정형빅데이터 활용성 확장을 위한 방법론 연구

제1절 비정형데이터와 정형데이터
제2절 임베딩방법론
제3절 임베딩에 기반한 연계방법론
제4절 딥러닝에 기반한 연계방법론 고도화 방안

제 4 장 비정형빅데이터 활용성 확장을 위한 방법론 연구

제1절 비정형데이터와 정형데이터

1. 비정형데이터의 특징

최근 IoT 저장 기술의 범용적인 활용성 확대로 인해 비정형 데이터가 보편적인 분석대상으로 인식되고 있다. 더불어 개개인이 만들어 내는 데이터의 연결 정도가 높아지는 초연결화 사회로 진입함에 따라, 정형화되지 않은 개인의 데이터로부터 의미를 찾아내기 위한 연구가 활발해지고 있다. 이러한 개인 정보에 대한 분석 기술의 발달은 기업이나 정부기관들의 경향 분석에 대한 수요를 증대시키는 주요한 요인이 되고 있다.

정형자료와 더불어 폭발적으로 증가하고 있는 비정형 데이터는 대부분의 인터넷 등 온라인 활동을 통해 획득되는 자료로 저장주기가 비교적 짧거나 자료 규격이 표준화되어 있지 않아 활용성에 제약이 따르는 자료라 정의할 수 있다. 이 절에서는 정형과 비정형자료의 구조를 대표적인 유형을 예로 들어 살펴보기로 한다.

가. 정형데이터(structured data)

비정형 데이터와 달리 정형 데이터는 규격화된 데이터 필드에 저장된 데이터를 말하며 일반적으로 그 값이 의미를 파악하기 쉽고 규칙적인 값을 가진다. 관계형 데이터베이스(relational database)나 스프레드시트(spreadsheet) 등을 대표적인 예로 들 수 있다. 데이터베이스의 설계안

에 따라 수집되는 정보의 형태가 정해지게 되며 한정된 정보들 속에서 분석을 진행할 수 있다.

이러한 방식은 데이터 구조화 정도가 높고 데이터 갱신 주기가 빠르지 않을 경우 관리 및 분석에 유용할 수 있다. 비정형 데이터와 비교한 정형 데이터의 특징들을 살펴보면 내부 시스템인 경우가 대부분이라 형식을 가지고 있어 수집 및 처리가 쉽다. 데이터의 스키마(database schema)를 지원하기 때문에 데이터를 탐색하는 과정이 테이블 탐색, 열 구조 탐색, 행 탐색 순으로 정형화되어 있다(Data on-air, 2020a).

스키마는 데이터베이스의 구조와 제약조건에 관해 전반적인 명세를 기술한 것이다. 즉, 개체의 특성을 나타내는 속성(attribute)과 속성들의 집합으로 이루어진 개체(entity), 개체 사이에 존재하는 관계(relation)에 대한 정의와 이들이 유지해야 할 제약조건들을 기술한 것이라고 볼 수 있다. 정리하면 스키마는 데이터베이스 내에 데이터가 저장된 구조를 나타낸다. 스키마는 데이터 사전에 저장되며, 현실 세계의 어떤 특정 부분을 표현하기 위하여 특정 데이터 모델을 이용하여 만들어진다. 따라서 스키마는 시간에 따라 변하지 않는 특성을 가지고 특정 시점에 데이터베이스에 들어있는 데이터인 인스턴스에 의해 규정된다. 스키마의 구조는 사용자의 어플리케이션과 논리적인 데이터베이스의 기술, 물리적인 데이터베이스 구조의 기술에 이르는 3단계로 구분하여 명명한다. 그 목적은 사용자의 응용과 물리적 데이터베이스를 분리하는 것이며 외부 단계(external level), 개념 단계(conceptual level), 내부 단계(internal level)로 나뉜다(인코돔, 2016).

나. 비정형데이터(unstructured data)

비정형 데이터는 규격화된 데이터 필드에 저장되지 않은 데이터로, 텍스트 분석이 가능한 텍스트 문서 외 이미지/동영상/음성 데이터 등을 지칭한다. 따로 정해진 규칙이 없기 때문에 값의 의미를 파악하기 어렵다. 텍스트의 경우 데이터 형태로 파싱(parsing)해야 할 필요성이 있기 때문에 수집된 데이터를 처리하기가 어려운 편이다. 텍스트나 파일을 파싱해 메타구조를 갖는 데이터 셋 형태로 바꾸고 정형 데이터 형태의 구조로 만들 수 있도록 아키텍처 구조를 수정해야 하여 향후 분석에 높은 잠재적 가치를 제공할 수 있게 가공하여야 한다.

현재 전 세계 데이터 생성량은 전년도에 비해 매년 2배가량 성장하고 있다고 알려져 있으며 시간이 갈수록 데이터 증가속도는 가속화되고 있다. 기존 데이터 관리 체계로는 데이터를 감당하기 힘든 수준에 이르게 되었다. 비정형 데이터의 생산량이 기업 데이터의 80% 이상이라는 조사 결과를 감안할 때 기존 정형데이터의 관계형 데이터베이스(related DBMS, RDBMS)에서 규칙적으로 데이터를 저장한 후 필요에 따라 SQL을 활용하여 데이터를 추출하는 방식은 유용성이 떨어진다고 할 수 있다(박대현, 송동현, 2014).

최근 SNS 개인화된 소셜 웹 콘텐츠가 쏟아지면서 실시간으로 수집하고 분석해 의미 있는 결과를 도출하려는 시도가 늘고 있다. 즉, 텍스트 위주의 소셜 미디어에서 감성분석(sentiment analysis)을 실시해 목표 집단의 성향을 분석하는 것이다. 감성분석은 텍스트 하나하나의 의미를 세밀하게 분석해 결과를 도출하거나 해당 주제와 연관된 정보를 수집해 여론을 파악하는데 활용된다. 그리고 텍스트 비정형 데이터를 기반으로 특정 대상에 대한 대중의 감성을 분석하기 위한 도구로 자연어 처리와 웹

크롤링(web-crawling)과 같은 검색 기술을 필요로 한다. 또한 단순한 긍정, 부정의 양극화에서 벗어나 감각, 태도, 의견 등의 연관 단어들도 분석 대상으로 하여 다양한 감성(sentiment)을 분류할 수 있는 여러 가지 기계학습/딥러닝 방법론이 개발되어 활용되고 있다(DATANET, 2011).

텍스트분석에서 기본적으로 요구되는 자연어 처리기술(NLP; natural language processing)은 인간의 언어를 컴퓨터가 이해할 수 있도록 하는 기술로, 이용자의 의도를 컴퓨터가 파악해 보다 정확한 정보를 취합해 제공한다. 정보의 행간을 읽어 해당 데이터가 가진 뜻을 파악하는데 사용하고 있으며 형태소분석, 미등록어(은어, 비속어, 신조어 등) 추정, 구문분석, 의미 분석, 담화분석 등의 요소기술이 필요하다. 자연어의 분석은 자연어 문장으로부터 형태소 분석(morphological analysis), 구문 분석(syntax analysis), 의미 분석(semantic analysis), 화용 분석(pragmatic analysis) 등을 여러 단계의 분석과정을 거치게 된다.

자연어 처리를 쓰는 분야는 텍스트 분석, 기계 번역과 언어 모델, 질의응답 시스템은 물론이거니와 최근의 음성인식, 영상처리 기술과 결합되어 개인비서나 이미지 캡션 생성 기술로까지 발전되고 있는 추세이다. 딥러닝을 활용하기 전 자연어 처리 기술은 대부분 규칙 기반(rule based)의 알고리즘으로 상용화하기 어려웠고 그 성능도 인간과 비교했을 때 월등히 떨어졌다. 하지만 컴퓨터 하드웨어와 기계학습(machine learning) 기술의 급격한 발전으로 자연어 처리기술 역시 이전과 비교하면 뛰어난 성능을 보이고 있으며, 최근에는 GPU와 클라우드 컴퓨팅을 활용한 대량의 빅데이터와 매우 복잡한 모델을 이용해 패턴 학습이 가능한 딥러닝(deep learning)과 같은 기법을 자연어 처리에 적용하고 있다(SD아카데미, 2018).

다. 반정형데이터(semi-structured data)

반정형 데이터는 데이터가 부분적으로 정형구조를 가진 데이터를 의미한다. 대표적으로 HTML, JSON, XML등과 같은 포맷을 반정형 데이터의 범위에 넣을 수 있다. 일반적으로 데이터베이스는 데이터를 저장하는 장소와 스키마가 분리되어 있어서 테이블을 생성하고, 저장하는 구조로 구성되어 있다. 그러나 JSON이나 XML와 같은 데이터의 구조를 가진 반정형 데이터는 한 텍스트파일에 변수명과 값을 모두 가지고 있다. 다음의 예를 통해 살펴보기로 한다(Data on-air, 2020b).

XML(extensible markup language)은 HTML과 비슷한 태그 등의 문자로 기반된 마크업 언어(markup language)이다. 사람과 기계가 동시에 읽기 편한 구조로 되어있고 데이터를 보여주는 목적이 아닌, 데이터를 저장하고 전달할 목적으로만 만들어졌다. XML의 태그는 미리 정의되어 있지 않고, 사용자가 직접 정의할 수 있다. 〔그림 4-1〕에서 식빵이라는 이름을 가진 강아지의 품종(family)과 나이(age), 무게(weight)를 가진 XML 문서의 예시이다.

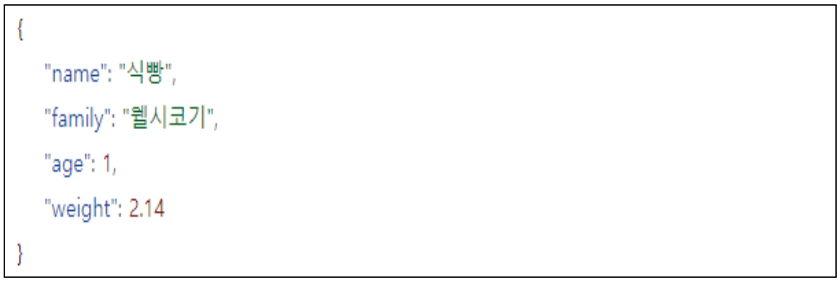
〔그림 4-1〕 XML 자료 예시

```
<dog>
   <name>식빵</name>
   <family>웰시코기<family>
   <age>1</age>
   <weight>2.14</weight>
</dog>
```

자료: http://tcpschool.com/json/json_intro_xml 인출일: 2019. 9. 1.

한편, JSON(javascript object notation)은 쉽게 데이터를 교환하고 저장하기 위해 만들어진 텍스트 기반의 데이터 교환 표준이다. 자바스크립트 기반으로 만들어졌으며 객체 표기법을 따른다. 프로그래밍 언어와 운영체제에 독립적이고 어떠한 프로그래밍 언어에서도 JSON 데이터를 읽고 사용할 수 있다. 〔그림 4-2〕는 JSON 형식에 대한 예제이다(TCPschool, 2018).

〔그림 4-2〕 JSON 자료 예시



자료: http://tcpschool.com/json/json_intro_xml 인출일: 2019. 9. 1.

반정형 데이터는 정형 데이터, 비정형 데이터와 명확한 구분을 하기 어렵다. 스키마가 잘 정의되어 있을수록 정형데이터화하기 쉽다고 볼 수 있다. 예를 들어 ① JSON 형태로 되어 있는 글 내용 본문은 반정형 데이터와 비정형 데이터가 합쳐진 구조라 할 수 있으며 ② 데이터베이스에 저장되어 있는 성별 값은 정형데이터로 볼 수 있다. ③ 데이터베이스에 저장된 글의 제목은 비정형 데이터라 볼 수 있다.

2. 임베딩 방법론 관련 연구

대표적인 비정형 자료인 텍스트에 대한 분석기술의 발전은 정형자료 분석에 대한 활용성 확장에 대한 도움을 주고 있다. 그러나 기본적으로 단어는 고차원벡터이므로 활용상에 상당한 제약이 수반된다. 그러므로 비정형 텍스트 분석에서는 단어벡터가 정의된 고차원 공간이 아닌 문장 내에서 의미를 이용하여 정보를 추출하고 이를 저차원상에 매핑하는 방법이 필요하다. 이러한 접근방법은 단어 임베딩(word embedding)과 같은 단어 특성 추출방법의 연구결과를 이용할 수 있을 것이다. 본 연구에서는 비정형빅데이터 확장을 위한 방법론을 위해서 임베딩 방법을 위주로 연계방안 등에서 대해서 알아보고자 한다. 다음과 같은 순서에 따라 논의를 진행한다.

① 임베딩 방법론 소개: 본 연구에서는 비정형빅데이터 확장을 위한 방법론을 위해서 임베딩 방법을 위주로 연계방안 등에서 대해서 알아보고자 한다.

② 임베딩 방법론에 기반한 연계방법론 소개: 연계방법론에 활용되는 정준상관분석을 살펴보고, 표현학습(representation learning)의 개념 및 고도화 방법론을 살펴본다.

③ 임베딩 방법론의 고도화: 심층합성망(deep convolution neural net), 순환신경망(recurrent neural net), 어텐션(attention) 구조 등 최근 딥러닝 모형 학습방법을 살펴보고, 보건/복지 분야의 비정형빅데이터 활용성 확장을 위해, 추천화시스템(recommendation system) 등을 포함한 딥러닝 방법 고도화에 대한 제언을 담고 있다.

제2절 임베딩방법론

다양한 구조를 가지는 비정형자료로부터 의미 있는 정보를 추출하거나 연계하기 위한 방안 등 다양한 분석 방법론들이 연구되고 있다. 이 절에서는 임베딩 기법에 기반한 기계학습방법론/딥러닝 학습방법론들을 자세히 소개하기 위해서 필요한 배경 개념 및 기본 연구들을 소개한다. 임베딩 방법론의 기초가 되는 대표적인 연구들로 오토인코더, 변분오토인코더 방법을 소개하고자 한다.

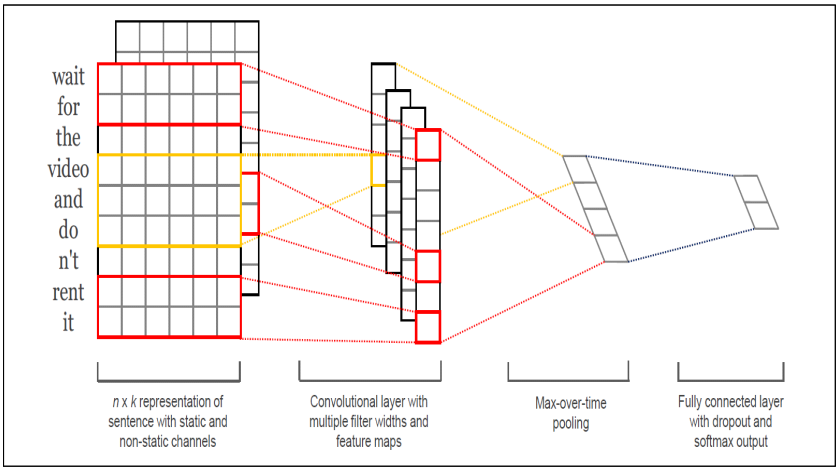
1. 임베딩(embedding)

자연어 처리 분야에서의 임베딩이란 기계가 이해할 수 있는 실수값으로 이루어진 벡터로 단어표현을 학습하는 과정과 그 결과를 나타낸다(Bengio et al., 2003). 대표적인 비정형 형태의 자료인 텍스트자료에서는 단어를 모형 적합에 사용하기 위해서 단어를 숫자 벡터로 변환하는 과정이 반드시 필요하다. 이를 워드 임베딩(word embedding)이라 부르며, 전통적인 방법에서는 전체 단어 집합의 개수를 V개라고 하면 하나의 단어를 해당하는 단어 위치에 1의 값을 가지고 있는 지시 벡터(indicator vector)로 변환하는 방법을 활용한다. 가장 기본적인 임베딩 방법인 TF-IDF(Term Frequency Inverse Document Frequency)는, 문서를 단어 단위로 분해하여 생성된 문서-단어 행렬(document-term matrix)에서 사용할 수 있는 방법으로 단어의 등장 여부를 0과 1을 통해 표현하는 지시 벡터로 표현한다. TF-IDF 방법은 단어의 개수에 비례하기 때문에 고차원을 가지게 된다.

최근 단어표현학습에 범용적으로 활용되고 있는 word2vec(Mikolov

et al., 2013), Glove(Pennington et al., 2014) 등의 기술들은 밀집한(dense)한 저차원의 실수벡터로 표현함으로써 비슷한 의미의 단어는 가까운 거리(유클리디언 거리 혹은 코사인 유사도)에 위치하도록 임베딩 공간을 만들 수 있다.

[그림 4-3] 1차원 합성망



자료: Kim (2014). "Convolutional neural network for sentence classification", Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, p. 1746-1751.

임베딩 방법은 특징을 분산하여 표현할 수 있는 장점을 가지고 있다. [그림 4-3]은 텍스트 자료에 대한 1차원 합성망모형을 활용하여 단어를 임베딩하는 기법을 나타내는 것으로 단어 색인을 저차원 벡터 표현으로 대응시킨다. 합성망모형은 문장의 단어들의 전후의 순서 정보를 보존함으로써 단어/표현의 등장순서를 학습에 반영하는 구조를 반영할 수 있다. 컴퓨터 비전 분야에서 널리 활용되는 합성망모형이 자연어 처리 문제에서도 효과적일 수 있음을 확인할 수 있다. 중간층으로 한 층을 사용하는 합성망모형에 대한 설명은 다음과 같다.

주어진 하나의 문장이 n개의 단어로 구성되어졌다고 하고, 각 입력단어는 k차원의 임베딩 실수벡터라고 하자. 즉, 한 문장에서 i번째 단어를 $x_i \in R^k$라 놓자. i번째 단어를 포함하여 연속된 j개의 단어로 구성된 문장을 연쇄(concatenation)하여 표현하면

$$x_{i:i+j} = x_i \oplus x_{i+1} \oplus \cdots \oplus x_{i+j}$$

와 같다. 필터의 커널(kernel)의 크기를 h(대역폭, bandwidth)라 놓자. 여러 필터에 적용하여 합성곱연산을 수행하고 입력정보로부터 특징들을 추출한다.

[그림 4-3]은 $n=9$개의 단어의 $k=6$차원의 사전 임베딩 행렬로부터 인접한 $h=3$개, 크기가 7($n-h+1=9-3+1=7$)인 벡터를 얻는 과정을 보여준다. 필터의 개수는 4이며, 이미지의 색상에 해당하는 채널의 수는 2이다. 문장의 양끝은 인접한 단어의 커널의 크기는 2이다. 한 필터의 가중치를 $w \in R^{hk}$로 두고, 절편 b가 주어지며 i번째 단어임베딩벡터에 대한 특징맵

$$c_i = f(w \cdot x_{i:i+h-1} + b)$$

를 얻을 수 있다. 여기서, f는 활성함수(activation function)이다.

h개의 인접 부분단어집합 전체 단어에 적용하여 $[x_{1:h}, x_{2:h+1}, \cdots, x_{n-h+1:n}]$와 같이 중복된 임베딩벡터를 가진 행렬로 표현하자. 전술한 특징맵 벡터를 뽑는 과정을 반복적용하면 특징맵

$$c = [c_1, c_2, \cdots, c_{n-h+1}] \in R^{n-h+1}$$

을 얻을 수 있다. $\max_{i=1,\cdots,n-h+1} c_i$와 같은 최대풀링(max pooling)

을 통해 c_i값 중 최대값을 추출한다. 이와 같은 과정을 4개의 필터에 적용하면 인접한 단어의 특징을 공유하는 특징맵을 만들 수 있고, 이 벡터를 통해 최상층의 소프트맥스(softmax)로 산출한다.

2. 오토인코더(autoencoder)

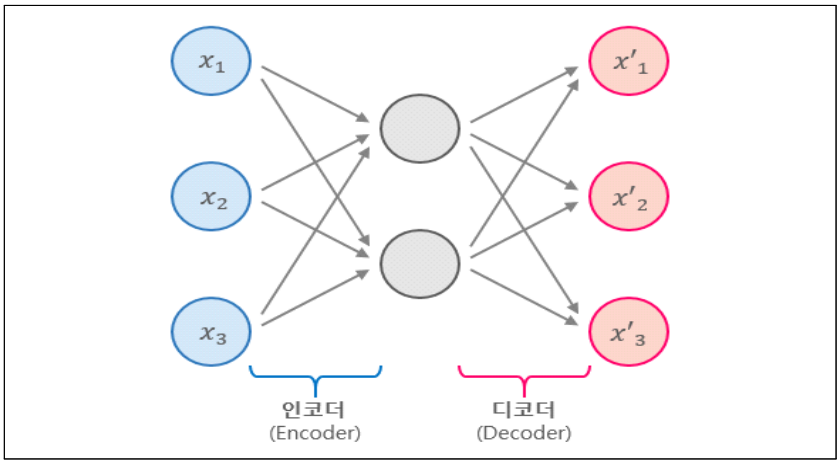
임베딩방법론 중 가장 기본적으로 알려진 비지도 학습(unsupervised learning)인 오코인코더(autoencoder)에 대해 알아보도록 한다. 오토인코더는 입력으로부터 출력을 하는 신경망으로 오토인코더, 과소완전 오토인코더, 희소 오토인코더, 노이즈 제거 오토인코더, 변분 오토인코더 등의 목적별로 여러 변형이 있다.

오토인코더는 인코더와 디코더로 구성되어 있는데 〔그림 4-4〕와 같이 단순히 입력을 출력으로 복사하는 신경망이다. 은닉층(hidden layer)의 뉴런 수를 입력층(input layer) 보다 작게 해서 차원을 축소(dimension reduction)하거나, 입력 데이터에 노이즈(noise)를 추가한 후 원본 입력을 복원(reconstruction)할 수 있도록 네트워크를 학습시키는 등 다양한 오토인코더가 있다. 이러한 다양한 제약들은 오토인코더가 단순히 입력 데이터의 특징을 바로 출력으로 단순 복사하지 못하도록 방지하며, 데이터를 효율적으로 표현(representation)하는 방법을 학습하도록 제어한다. 오토인코더는 〔그림 4-4〕와 같이 인코더(encoder)와 디코더(decoder), 두 부분으로 구성되어 있다.

① 인코더(encoder): 인지 네트워크(recognition network)으로 입력 벡터를 응축된 표현벡터로 변환한다.

② 디코더(decoder): 생성 네트워크(generative network)으로 표현을 출력한다.

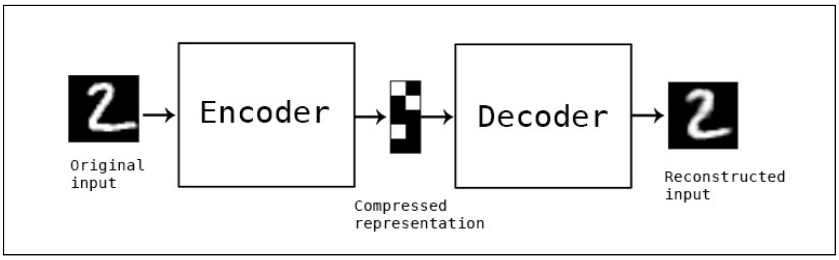
〔그림 4-4〕 오토인코더의 구조



자료: https://excelsior-cjh.tistory.com/187 인출일: 2019. 9. 1.

오토인코더는 입력과 출력층의 뉴런 수가 동일하다는 것만 제외하면 일반적인 다층 신경망인 MLP(multi-layer perceptron)과 동일한 구조이다. 오토인코더는 입력으로부터 새로운 특징들로 재구성하기 때문에, 최적화 대상이 되는 목적식은 입력과 새로 구성된 입력간의 복원의 정확도에 근거한다. 이를 보통 복원오차(reconstruction error)라 부르며 예를 들어, 연속형 입력자료의 경우 가장 일반적인 손실함수로써 제곱 손실함수를 들 수 있다. 손실함수는 입력과 출력의 차이인 잔차를 가지고 계산한다.

〔그림 4-5〕 Mnist 숫자 이미지에 대한 오토인코더와 입력과 복원된 출력 이미지



주: 그림 중간의 응축된 표현(compressed representation)은 임베딩 벡터를 나타냄.
자료: https://keraskorea.github.io/posts/2018-10-23-keras_autoencoder/ 인출일: 2019.9. 12.

오토인코더는 은닉층의 뉴런(노드)의 수가 입력층보다 작기 때문에 입력이 저차원으로 표현되는데, 이러한 오토인코더를 과소완전(undercomplete) 오토인코더라고 한다. 이 방법은 저차원을 가지는 은닉층에 의해 입력을 그대로 출력으로 복사할 수 없기 때문에, 출력이 입력과 같은 것을 출력하기 위해 학습해야 한다. 이러한 학습을 통해 입력벡터에서 가장 중요한 특성(feature)을 학습하도록 만든다. 즉, 은닉층을 h로 놓고, 인코더 함수를 $h=f(x)$로 $r=g(h)$를 복원된 값을 산출하는 디코더라고 놓자. 오토인코더로부터 h가 x보다 낮은 차원을 가지도록 제한한 과소완전 오토인코더는 훈련자료(training data)에서 가장 두드러진 특징을 포착하는 것으로 알려져 있다.

오토인코더의 목적함수를 $L(x,g(f(x)))$라 둘 때, 이는 $g(f(x))$는 인코딩된 은닉층이 주어져 있을 때 복원오차를 최소화하는 해당 손실함수로 제곱오차 $(x-g(f(x)))^2$를 예로 들 수 있다. 만약 오토인코더가 선형함수이고 L이 평균제곱오차일 때, 과소완전 오토인코더는 차원축소방법인 주성분분석(principal component analysis)와 동일하다고 할 수 있다. 오토인코더의 최적화문제는 다음과 같이 주성분분석에서 활용되는 행렬의 랭크-K근사문제로 해석된다.

M차원의 실수공간의 기저벡터(basis vector)행렬을 $W=[w_1 w_2 \cdots w_k]$라 놓자. 랭크-K근사문제는 주어진 $a_1, a_2, \cdots, a_N (a_i \in R^M)$ 자료벡터를 정규직교인 기저벡터 $w_1, w_2, \cdots, w_N$로 이루어진 K차원 벡터공간으로 투영하여 가장 비슷한 N개의 K차원 벡터 $a_1^{\|W}, a_2^{\|W}, \cdots, a_N^{\|W}$를 생성하는 정규직교 기저벡터(orthonormal basis vector) $w_1, w_2, \cdots, w_N$를 찾는 문제라 볼 수 있다. 정규직교 기저벡터에 대한 투영된(projected) $a_i^{\|W}$는 각 기저벡터에 대한 내적으로 만들 수 있다.

$$\begin{aligned} a_i^{\|W} &= (a_i^T w_1) w_1 + (a_i^T w_2) w_2 + \cdots + (a_i^T w_k) w_k \\ &= \sum_{k=1}^{K} (a_i^T w_k) w_k \end{aligned}$$

벡터 $a_1, a_2, \cdots, a_N$을 행벡터로 가지는 행렬 $A = \begin{bmatrix} a_1^T \\ a_2^T \\ \vdots \\ a_N^T \end{bmatrix}$를 가정하자. 모든 점들과의 거리의 제곱의 합은 행렬의 노음(norm)으로 계산 가능하며

$$\sum_{i=1}^{N} (\|a_i\|^2 - \|a_i^{\|W}\|^2) = \|A\|_2^2 - \sum_{i=1}^{N} \|a_i^{\|W}\|_2^2$$

이 된다. 이를 분산행렬 형태로 표현하면 다음과 같다.

$$\sum_{i=1}^{N} \|a_i^{\|W}\|^2 = \sum_{i=1}^{N} \sum_{k=1}^{K} \|(a_i^T w_k) w_k\|^2 = \sum_{i=1}^{N} \sum_{k=1}^{K} \|a_i^T w_k\|^2 = \sum_{k=1}^{K} w_k^T (A^T A) w_k$$

이를 분산행렬의 고유분해 사용하여 표현하면

$$\sum_{k=1}^{K} w_k^T (A^T A) w_k = \sum_{k=1}^{K} w_k^T (V \Sigma V^T) w_k \qquad (4\text{-}1)$$
$$= \sum_{k=1}^{K} w_k^T (\sum_{i=1}^{M} \sigma_i^2 v_i v_i^T) w_k = \sum_{k=1}^{K} \sum_{i=1}^{M} \sigma_i^2 \| v_i^T w_k \|^2$$

가 된다. 따라서 가장 큰 K개의 특이값에 대응하는 오른쪽 특이벡터가 기저벡터일 때 가장 값이 커지게 됨을 알 수 있다. 이는 식 (4-1)의 값을 가장 크게 하는 K개의 직교하는 단위벡터 w_k는 고유분해의 성질로부터 오른쪽 기저벡터 중 가장 큰 K개의 특이값에 대응하는 오른쪽 특이벡터가 최적해가 된다. 이중 랭크-K 근사화의 형태는

$$a_i^{\|\mathrm{W}} = (a_i^T w_1) w_1 + (a_i^T w_2) w_2 + \cdots + (a_i^T w_K) w_K = \begin{bmatrix} w_1 w_2 \cdots w_K \end{bmatrix} \begin{bmatrix} w_1^T \\ w_2^T \\ \vdots \\ w_K^T \end{bmatrix} a_i$$
$$= W W^T a_i$$

와 같다. 그러므로 투영벡터를 모아놓은 행렬 $\hat{A}$는

$$\hat{A} = \begin{bmatrix} (a_1^{\|\mathrm{W}})^T \\ (a_2^{\|\mathrm{W}})^T \\ \vdots \\ (a_N^{\|\mathrm{W}})^T \end{bmatrix} = \begin{bmatrix} a_1^T W W^T \\ a_2^T W W^T \\ \vdots \\ a_N^T W W^T \end{bmatrix} = \begin{bmatrix} a_1^T \\ a_2^T \\ \vdots \\ a_N^T \end{bmatrix} W W^T = A W W^T$$

으로 주어짐을 알 수 있다. 그러므로 원래 행렬 A에 랭크-K 행렬 WW^T를 곱해서 원래의 행렬 A와 가장 비슷한 행렬 $\hat{A}$을 만드는 문제와 동일하므로 다음과 같이 원 문제를 표현할 수 있다.

$$argmin_{w_1, \cdots, w_K} \| A - \underbrace{A W W^T}_{\hat{A}} \|_2^2.$$

여기서 추정량 $\hat{A}$는 입력자료 A로부터 AW로 인코딩되고, $(AW)W^T$로 다시 디코딩된다고 볼 수 있다. 이는 인코딩과 디코딩의 계수행렬 W은 동일하게 설정된 선형 오토인코더로 해석할 수 있다.

한편, 희소 오토인코더 복원오차뿐만 아니라 은닉층에 대한 희소성 벌점함수(penalty function)도 추가한 다음의 목적함수(objective function)를 고려할 수 있다.

$$L(x, g(f(x))) + \Omega(h),$$

여기서, $g(h)$는 디코더의 출력이며, $h = f(x)$는 인코더이다. 그러나 오토인코더의 함수공간이 크거나 차원이 너무 크면 훈련자료에서 학습이 잘 수행되지 않을 수 있다.

오토인코더는 입력 데이터와 출력 데이터가 동일하다는 점에서 지도 학습(supervised learning)과의 차이가 있다. 오토인코더는 지도학습에서 필수적으로 요구되는 출력변수가 없어도 비지도 학습방법(unsupervised learning)의 유형으로 학습을 할 수 있다. 최근 텍스트 분석에서 가장 널리 쓰이는 트랜스포머(transformer; Vaswani et al., 2017), GPT2(Radford et al., 2019)와 GPT3(Brown et al., 2020) 등은 구체적 학습의 목표(task)가 주어지지 않을 상황에서도 오토인코더와 유사한 목적의 자기주도 지도학습(self-supervised)을 통해 원래 지도학습의 정도를 향상시킬 수 있음을 실증하고 있다.

요약하면 오토인코더를 활용한 임베딩 방법은 크게 두 가지 특징을 가진다. ① 입력이 곧 출력이므로 복원손실이 필수적으로 발생하며 ② 분석자의 주관에 의존하기보다는 데이터를 통한 학습을 하며 대부분 신경망 구조하에서 진행된다.

3. 변분 오토인코더(variational autoencoder)

변분추론(variational inference)은 계산이 어려운 확률분포를, 다루기 쉬운 분포 $q(z)$로 근사하는 방법이다. $p(x)$는 x를 관찰할 확률에 해당되며 이 분야에서는 evidence라고 불린다. 이 항의 로그값의 하한(evidence lower bound, ELBO)을 다음과 같이 구할 수 있다. $p_\theta(x)$를 θ를 모수로 하는 확률모형이라고 하고 이 확률을 최대화한다고 하자. $q_\phi(z|x)$는 모수 ϕ를 가지는 확률모형이라고 놓자.

$$
\begin{aligned}
& \log p_\theta(x) \\
= \; & \log p_\theta(x) \int q_\phi(z|x) dz \\
= \; & \int q_\phi(z|x) \frac{\log p_\theta(x,z) p_\theta(x)}{p_\theta(x,z)} dz \\
= \; & \int q_\phi(z|x) \log \frac{p_\theta(x|z) p_\theta(z)}{p_\theta(z|x)} dz \\
= \; & \int q_\phi(z|x) \log p_\theta(x|z) dz - \int q_\phi(z|x) \log \frac{q_\phi(z|x)}{p_\theta(z|x)} dz \\
+ \; & \int q_\phi(z|x) \log \frac{q_\phi(z|x)}{p_\theta(z)} dz \\
= \; & E_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{KL}(q_\phi(z|x)|p_\theta(z)) + D_{KL}(q_\phi(z|x)|p_\theta(z|x)) \\
= \; & L(\theta,\phi;x) + D_{KL}(q_\phi(z|x)|p_\theta(z|x)) \\
\geq \; & L(\theta,\phi;x)
\end{aligned}
$$

위 부등식에서 로그 확률 $\log p_\theta(x)$의 우변, 변분 하한(variational lower bound)인 ELBO

$$L(\theta,\phi;x) = E_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{KL}(q_\phi(z|x)|p_\theta(z))$$

를 최대화하면 $\log p_\theta(x)$를 최대화할 수 있다. 여기서, $q_\phi(z|x)$는 자료 x에 대한 z에 대한 사후확률로 인코딩 함수에 해당된다. $p_\theta(x|z)$는 z가 주어지면 x에 대한 사후확률로 디코딩함수이다. 생성모형(generative model)에 기반한 임베딩 방법의 기저방법으로 널리 활용되고 있다.

변분 오토인코더(variation autoencoder; VAE)는 생성모형(generative model) 중 하나로, 확률분포 $p_\theta(x)$의 최대하한을 학습함으로써 데이터를 생성하는 것을 목표로 한다. 인코더 네트워크는 학습용 데이터(x)를 입력으로 받고 잠재 변수(z)의 확률분포에 대한 모수를 출력한다. 예를 들어 정규 분포의 경우 $\phi = (\mu, \sigma)$이다. 디코더는 잠재변수에 대한 확률 분포 $p(z)$에서 샘플링한 벡터를 입력받아, 원본 이미지를 복원한다.

변분추론은 이상적인 확률분포를 모르지만, 이를 추정하기 위해서 다루기 쉬운 분포(approximation class, 대표적으로 Gaussian distribution)를 가정하고 이 확률분포의 모수를 바꿔가며, 이상적인 확률분포에 근사하게 만들어 그 확률분포를 대신 사용하는 것이다. 다루기 쉬운 분포를 q_ϕ라고 한다면, 인코더 $q_\phi(z|x)$는 이상적인 확률 분포 $p(z|x)$에 근사시키는 역할을 수행한다. 보통 q_ϕ은 알려진 분포를 활용하며, 예를 들어 다변량 정규 분포 등을 활용한다.

z의 역할을 보면, z값은 적당한 분포(보통 정규분포)로부터 임의 생성되는 랜덤벡터이다. 이는 곧 매 실행마다 임의 생성됨을 의미하므로 결과가 달라질 수 있는 점을 뜻한다. 최근에는 z로부터 모형을 생성하여 결과를 산출하고 난 뒤 다시 의미 있는 z의 공간을 찾기 위한 연구들을 진행하고 있으며, 이를 통해 모형의 결과를 제어할 수 있게 된다.

정리하면, 변분 오토인코더는 최적화를 통해 ① 주어진 데이터를 잘 설명하는 잠재 변수의 분포를 찾고(encoder의 역할) ② 잠재변수로 부터

원본 이미지와 유사한 이미지를 복원하는 것을 목표로 한다(decoder의 역할). 인코더의 역할은 데이터가 주어졌을 때 디코더가 원래의 데이터로 잘 복원할 수 있는 잠재 벡터 z를 샘플링 할 수 있는 이상적인 확률분포 $p(z|x)$을 찾는 것이다. 그러나 이상적인 확률분포 $p(z|x)$는 알 수 없기 때문에 전술한 변분추론(variational inference)을 활용한다.

4. 어텐션(attention) 구조

최근 번역 모형에 많이 활용되고 있는 자기주도 어텐션(self-attention) 또는 트랜스포머(transformer; Vaswani et al., 2017), BERT 등은 어텐션 구조에서 발전된 형태이다. 어텐션(attention)은 자료에서 주의 집중할 부분에 관한 방법으로 주어진 질의(query)에 대해 키(key) 값들을 탐색하여 해당하는 key 값들의 값(value)에 의해 결정된다.

n개의 자료 $\{(x_i, y_i)\}_{i=1}^{n}$가 설명변수 $x_i \in R^p$와 종속변수 $y_i \in R$의 짝으로 구성되어 있다고 하자. 그러면 새로운 x에 대한 예측값 $\hat{y}(x)$를 가중평균

$$\hat{y}(x) = \sum_{i=1}^{n} \alpha(x_i, x) y_i \qquad (4\text{-}2)$$

를 통해서 구할 수 있다. 여기서 $\alpha(x_i, x)$는 0과 1사이의 값으로 i번째 x_i자료와의 관련성, 유사성에 기반한 가중치로 그 값이 클수록 1에 가까운 값을 가진다. 이 때, x를 query, x_i들을 key, y_i를 value라 매칭할 수 있다.

이러한 어텐션(attention) 개념은 커널밀도함수(kernel density estimation) 추정방법에서 널리 쓰이는 Nadaraya-Watson의 커널회귀

(kernel regression) 회귀추정량으로 설명할 수 있다. 다음의 추정량을 고려해보자.

$$\widehat{m}_h(x) = \frac{\sum_{i=1}^{n} k_h(x - x_i) y_i}{\sum_{j=1}^{n} k_h(x - x_j)} \qquad (4\text{-}3)$$

여기서, k_h는 대역폭(bandwidth)이 h인 커널함수(kernel function)이다. 식 (4-3)은 식 (4-2)으로부터 유래한다고 볼 수 있는데, 이를 확인하기 위해 식 (4-3)의 회귀추정량을 고려하자. 가중치 역할을 하는 값을 $\alpha(x_i, x) = \frac{k_h(x - x_i)}{\sum_{j=1}^{n} k_h(x - x_j)}$로 놓으면 0-1으로 표준화한다. 식 (4-3)의 식은 $\widehat{m}_h(x) = \sum_{i=1}^{n} \alpha(x_i, x) y_i$로 표현할 수 있으며 이는 어텐션에서 살펴본 식 (4-2)과 동일함을 확인할 수 있다.

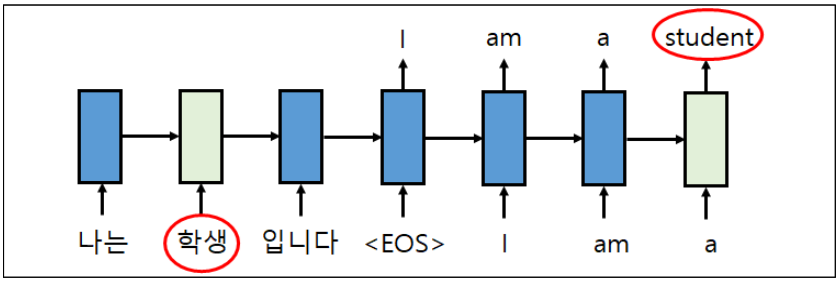
대표적인 비정형 자료인 텍스트 자료에 대해서 어텐션을 구체적 활용 사례는 다음과 같다. 어텐션 기법의 기본적인 아이디어는 디코더에서 출력 단어를 예측하는 매 시점마다 인코더에서 전체 입력 문장을 해당 시점에서 예측해야 할 단어와 연관이 있는 단어 부분을 좀 더 높은 가중치로 반영하는 것이다.

어텐션 구조는 전술한 query, key, value의 구조를 활용한다. 어텐션 함수는 주어진 query에 대해서 모든 key와의 유사도를 각각 구하고 이 유사도를 key와 맵핑 되어있는 각각의 값에 반영해 주는 원리이다. 이후 유사도가 반영된 어텐션 값을 가중합산하여 산출한다. Query는 t 시점의 디코더 셀에서의 은닉 상태이며, key는 모든 시점의 인코더 셀의 은닉

상태이다. Value는 모든 시점의 인코더 셀의 은닉 상태들의 값이다.

〔그림 4-6〕은 seq2seq 모형(4절 참고)에서의 어텐션의 활용에 대한 예제이다. 입력문장 “나는 학생입니다”에 대한 출력문장 “I am a student”에 대한 번역 태스크에 해당된다.

〔그림 4-6〕 seq2seq 모형에서 어텐션의 활용



자료: https://m.blog.naver.com/PostView.nhn?blogId=ckdgus1433&logNo=221608376139&proxyReferer=https:%2F%2Fwww.google.com%2F. 인출일: 2019. 9. 12.

예를 들어 “나는 학생입니다” 라는 문장을 영어로 번역시 “I am a” 의 세 단어 뒤에 어떤 단어가 올지 최종 출력을 결정할 때 “학생” 이란 두 번째 입력시점에 집중해서 “student”의 디코딩을 계산해야 한다. 이 때 어텐션 값은 한 디코딩 입력으로부터 모든 입력에 대해 계산이 되고 이 계산은 모든 디코딩 단어에서 반복된다. seq2seq과 어텐션을 결합한 모델에 사용될 수 있는 어텐션의 종류는 다양하다. 이 중 가장 간결한 구조를 가진 어텐션인 “dot attention”에 대한 구조는 다음과 같다. 어텐션 값은 디코딩 시에 인코더의 어느 입력시점 t에 집중할 것인지를 점수화한 값이다. 어텐션은 i시점의 디코더 벡터가 s_{i-1}라고 할 때 인코더의 j시점의 벡터 h_j와의 관계이다. h_j와 s_{i-1}를 인코더와 디코더 역할을 하는 은닉층이라고 할 때 스칼라값인 e_{ij}는 $e_{ij} = attention(s_{i-1}, h_j)$으로 정의될 수 있다.

다음으로 계산한 어텐션 값을 소프트맥스(softmax) 활성함수에 적용하여 어텐션에 관한 시간의 가중치 $\alpha_{ij}(0 \leq \alpha_{ij} \leq 1)$에 관한 분포를 구성한다.

$$\alpha_{ij} = softmax(e_{ij}) = \frac{e^{e_{ij}}}{\sum_{k=1}^{T_x} e^{e_{ik}}}.$$

이러한 어텐션 값으로부터 입력 은닉 상태들의 가중합을 계산하여 하나의 벡터로 만드는데 이를 맥락벡터(context vector)라 부른다. 인코더의 각 시점별 h 정보를 활용하여 디코더의 각 시점에 반영하는 어텐션 개념을 계산하기 위해 어텐션 벡터를 a라고 두면 a 벡터는 디코더의 각 시점에서 인코더의 시점과의 관련성 정보를 갖는 벡터이다. 벡터의 원소 중 가장 높은 값을 갖는 시점이 디코더 시 높은 가중치를 갖는다.

어텐션 출력벡터와 이전 디코더의 은닉 상태, 이전 디코더의 출력이라는 세 가지 값을 이용해 다음 시점의 디코더의 은닉 상태를 출력한다. 이는 가변적인 맥락벡터를 고려한 은닉상태를 계산할 수 있게 되므로 어텐션이 적용된 인코더-디코더와 적용되지 않은 인코더-디코더의 출력에 차이를 크게 한다.

제3절 임베딩에 기반한 연계방법론

이 절에서는 임베딩에 기반한 정형자료와 비정형자료를 연계하기 위한 방법론을 살펴본다. 일반적으로 입력과 출력간의 인과관계를 규명하는 것은 어렵기 때문에 상관성과 같은 연관성을 모형화한다. 이 절에서는 정형자료와 비정형자료를 연계하기 위한 방법론에 대해서 살펴보고 임베딩 방법론을 적용하는 연구결과를 소개하기로 한다.

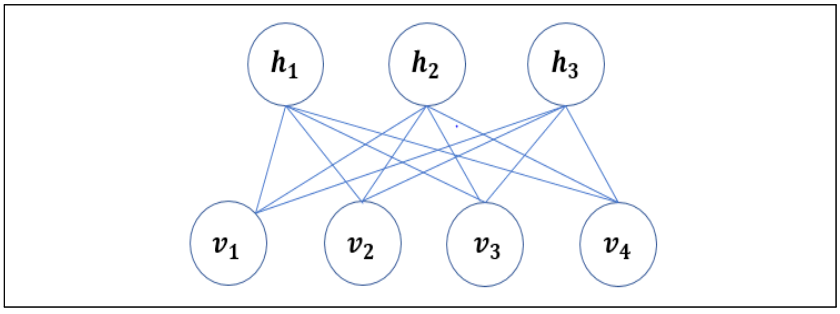
1. 표현학습(representation learning)

최근 딥러닝(deep learning)은 텍스트/이미지/오디오 분석 분야 등을 포함하여 여러 분야에서 광범위하게 활용되고 있다. 딥러닝 방법의 기저 방법으로 활용되는 표현학습(representation learning)은 심볼릭(symbolic) 오브젝트 형태를 특징을 가지는 자료들을 다차원의 실수공간으로 매핑하는 임베딩(embedding)을 찾는 학습방법을 통칭하여 일컫는다. 딥러닝의 심층신경망의 학습 과정에서 데이터의 다차원 공간에서의 다양한 표현을 찾는 것이 필요하며, 이를 통해 다양한 분석을 수행할 수 있다. 분석이나 예측에 좋은 자료는 유용한 특징을 가져야 되며, 이로부터 학습에 도움을 주는 유용한 표현을 찾는 것이 필요하다.

표현학습은 학습 시 유용한 표현을 만드는 과정이라 할 수 있는데 심층신경망 구조의 최상단 은닉층과 출력변수와의 학습을 제외한 나머지 은닉층들은 학습에 사용될 표현들을 생성하는 과정으로 이해할 수 있다. 표현학습은 학습 방법과 네트워크 설계로 사물과 관념, 개념에 대한 표현을 학습하는 과정을 다룬다.

표현학습의 대표적인 예로 제한된 볼츠만 머신(Restricted Boltzmann Machine, RBM)을 살펴보기로 한다. 먼저 RBM에서 활용하는 용어에 대해 간단히 설명한다. [그림 4-7]과 같이 차원의 수가 4인 관측 가능한 가시변수(visible variable) $v=(v_1,v_2,v_3,v_4)^T$와 차원의 수가 3인 은닉층 $h=(h_1,h_2,h_3)^T$이 연결된 그래프의 구조로 주어졌다고 가정하자. 7개의 모든 변수는 0과 1의 이진값(binary value)을 가진다고 가정하자.

[그림 4-7] RBM의 모형구조



자료: 저자 작성

관측한 가시변수로부터 관찰되지 않은 은닉변수를 추정하는데 적절한 모형이 필요하다. 관측 x를 종속변수로 하고, 은닉변수 h를 설명변수로 하는 회귀문제를 고려해보자. 그러면 목적함수로 제곱손실 $\|x-W^Th\|_2^2$을 고려할 수 있다. RBM에서는 이를 에너지(energy) 함수로 정의하고 이를 보다 일반화하여 전개하면 다음의 에너지함수를 얻을 수 있다.

$$\begin{aligned} E(x,h) &= -b^Tx-c^Th-h^TWx \\ &= -b^Tx-\sum_{k=1}^{3}(h_k(c_k-W_kx)). \end{aligned}$$

식에서 b와 c는 절편역할을 하는 항이라 할 수 있으며, W는 h와 x의 상호작용항이라 할 수 있다. 모형의 용어에서 제한(restricted)의 의미는 에너지 함수의 이차전개 시 얻게 되는 x^Tx와 $\|Wh\|_2^2$와 같은 자기상관항은 RBM에서는 고려하지 않음에서 유래한다.

이진값을 가지는 q차원의 h에서 모든 가능한 조합에 따른 개별 에너지를 적분한 주변(marginal) 에너지를 전에너지(total free energy)라 정의하며 다음과 같다.

$$\begin{aligned}
F(x) &= -\log\sum_h e^{-E(h,x)} \\
&= -\log e^{b^Tx}\sum_{h=\{0,1\}^q} e^{c^Th+h^TWx} \\
&= -b^Tx-\log\sum_h e^{c^Th+h^TWx} \\
&= -b^Tx-\log\sum_{h_1} e^{c_1h_1+h_1W_1x}\sum_{h_2} e^{c_2h_2+h_2W_2x}\cdots\sum_{h_q} e^{c_qh_q+h_qW_qx} \\
&= -b^Tx-\sum_{k=1}^{q}\log\sum_{h_k} e^{c_kh_k+h_kW_kx} \\
&= -b^Tx-\sum_{k=1}^{q}\log(1+e^{c_k+W_kx}).
\end{aligned}$$

식으로부터 관측 x가 주어졌을 때의 q차원의 은닉변수의 조건부 확률은 $p(h|x)=\prod_{k=1}^{q}p(h_k|x)$으로 됨을 유도할 수 있으며, 또한, 특정 k번째 항이 1의 상태를 가질 확률은 시그모이드(sigmoid) $p(h_k=1|x)=\frac{1}{1+e^{-(c_k+W_kx)}}$로 주어짐을 유도할 수 있다. 동일한 방법으로 은닉이 주어졌을 때의 x의 조건부확률은 $p(x|h)=\prod_{j=1}^{p}p(x_j|h)$로 성립함을 보일 수 있으며, 특정 j번째 항이 1을 가질 확률은

$p(x_j = 1|h) = \frac{1}{1 + e^{-(b_j + W_j^T h)}}$로 유도할 수 있다. 이 경우, x가 주어졌을 때의 q차원 $h = (h_1, h_2, \cdots, h_q)$는 임베딩벡터로 x에 의한 표현(representation)으로 여길 수 있다. 단, h는 관측 가능하지 않으므로 x가 주어졌을 때의 조건부 확률을 기반으로 모의생성하게 된다.

RBM의 예에서와 같이 관측한 자료로부터 구한 새로운 특징의 표현을 학습된 표현(learned representation)이라고 하며, 지도학습 문제에서 예측성능을 향상시키는 데 주로 사용되고 있다. 앞에서 살펴본 오토인코더는 이러한 표현학습(representation learning)의 한 유형으로 특징 추출을 수행하는 방법이라 할 수 있다. 표현학습은 대개 원본 특징의 표현을 학습하고, 이후의 계층은 학습된 은닉화된 표현을 기반으로 구축된다. 계층이 심화될수록 단순한 표현에서 점점 더 복잡한 표현을 학습하면서 점점 더 추상적인 개념 층을 구축하는 것으로 알려져 있다.

비정형 데이터의 정형 데이터와의 연계는 텍스트, 이미지, 오디오 등의 각기 다른 다양한 신호로부터 이들의 상관관계를 고려하여 결합, 공통적으로 표현하는 개념을 학습함으로써 이루어진다. 비정형자료는 대개 음성, 텍스트, 이미지 등인데 이와 같이 한 관측치에 대해 정의역이 다른 다양한 자료들을 모달리티(modality)라 일컫는다. 여기서 한 측정변수는 다차원(multi-dimension)을 가질 수 있다. 다중 모달리티 데이터는 각 모달리티가 상이한 통계적 특성과 각기 다른 표현 체계를 갖고 있다. 또한, 다양한 환경에서 측정될 수 있으므로 대개 척도가 상이하며 서로 다른 잡음을 수반한다. 따라서 다중 모달리티 학습은 이러한 데이터를 효율적으로 처리하여 공통적으로 표현하는 개념에 상응하는 결합된 표현 체계를 학습하는 것이 중요하다.

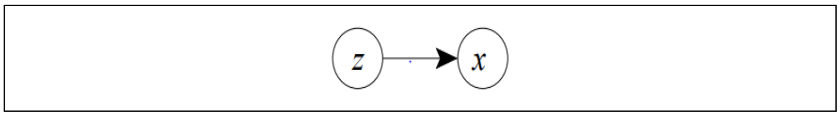
표현학습의 장점으로는 ① 특정 데이터 도메인(domain)에 대한 선험

적 지식이나 통찰이 다소 부족하더라도 관측된 자료와 자료 분석에 적합한 다양한 손실함수를 활용하여 임베딩을 구할 수 있다. 예를 들어, 의학에서의 전문용어도 처방전이나 의학사전 등에서 전문용어의 전후 사용된 단어들의 문맥정보로부터 다차원의 임베딩 벡터로 표현될 수 있다. ② 어떠한 유형의 데이터도 서로 다른 데이터와 연계를 할 수 있는데 예를 들어, 서로 다른 모달리티를 가지는 비정형형태의 이미지, 영상, 텍스트 등 숫자로 표현할 수 있기 때문에 특징맵 간의 융합이 가능하다. ③ 감정 등과 같은 표현하기 힘든 느낌이나 캐릭터와 같은 심볼릭(symbolic) 형태의 자료나 기호 등에 대해서도 표현할 수 있다.

2. 정준상관분석(canonical correlation analysis, CCA)

다양한 모달리티들은 서로 다른 측도를 가지고 있으므로 이들 간의 연관성을 모형화하여 분석하기가 어려운데 정준상관분석은 이러한 서로 다른 형태의 모달리티를 가진 자료의 연관성 분석에 활용될 수 있다. 정준상관분석은 적합한 특정 형태의 방향성을 가정하지 않고도 여러 양식 간의 공통의 변동 원인을 식별하고 모형화할 수 있으므로 다중 모달리티 데이터의 융합에 주요한 도구라 할 수 있다. 이러한 분석목적을 필요로 하는 여러 분야에서 비선형 투영을 수행하는 커널 CCA(kernel CCA), 제약 CCA(constrained CCA), 딥 CCA(deep CCA) 및 다중 세트 CCA를 포함한 CCA 변형방법 등이 널리 적용되고 있다(Zhuang et al., 2020).

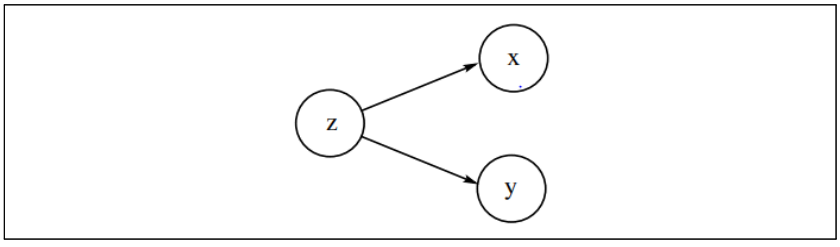
〔그림 4-8〕 요인모형



자료: 저자 작성

정준상관분석모형에 앞서 〔그림 4-8〕과 같이 표현된 간단한 요인(factor) 모형을 살펴보자. 요인모형에서는 잠재요인 z가 주어졌을 때, x가 관측된다고 가정한다.

〔그림 4-9〕 요인모형



자료: 저자 작성

이를 두 변수에 대한 관계식으로 확장한 경우의 표현은 〔그림 4-9〕와 같다. 보자. 〔그림 4-9〕는 두 변수의 공통요인의 역할을 하는 z가 주어졌을 때, 두 변수 x와 y의 관계를 보여준다. 연속형 자료에 대한 정준상관분석의 확률적 해석으로는 잠재확률변수 $z \sim N(0, I)$가 주어져 있을 때, x와 y의 조건부확률분포를

$$x|z \sim N(W_x z, \Phi_x),\ W_x \in R^{p_x \times p_z}$$
$$y|z \sim N(W_y z, \Phi_y),\ W_y \in R^{p_x \times p_z}$$

와 같이 구조화하면 x와 y의 관련성을 구조화하여 표현할 수 있다.

요인모형에 기반하여 확률적인 모형으로 해석하는 것과 별개로 정준상관분석은 주어진 변수들의 선형결합변수들의 상관성을 찾는 방법으로 설명될 수 있다. 즉, 두 변수 집단의 상관성을 각 변수집단에 속한 변수들의 선형결합(linear combination)들의 상관계수를 이용하여 설명하고자 한다. 각 변수집단에 속하는 변수들의 선형결합은 선형결합들 사이의 상관관계가 최대가 되도록 가중치(weight)를 결정하여 구성할 수 있다. 새로 만들어진 선형결합변수를 정준변수(canonical variable)라고 하며, 이들 정준변수들 사이의 상관계수를 정준상관계수(canonical correlation coefficient)라 부른다(Hotelling, 1936). 정준상관분석은 회귀분석과 달리 같은 종속변수와 설명변수간의 설명력을 설명하는 인과관계를 모형화하지 않는다.

정준상관계수는 두 투영된 스코어(projected score)간의 차이를 제곱손실로 여길 경우 오차가 되며 이를 최소화하는 투영된 방향을 찾는 문제가 된다. 투영된 방향의 크기에 제약값을 부여함으로써 유일하지 않는 해를 추정하는 실질적인 방안이 된다. 먼저 p_1차원의 y_1과 p_2차원의 y_2 두 잠재변수에 대한 상관성을 최대화하는 것을 찾는다. 두 잠재변수의 원 변수들이 관측되는 정의역 즉, 서로 다른 모달리티에 의존하지 않는다. $Y_k, k=1,2$를 $N\times p_k, k=1,2$의 크기의 행렬이라고 하자. N개의 관측치와 각 변수의 크기가 p_1, p_2이라고 하자. 정준상관계수 p_1차원의 u_1과 p_2차원의 u_2를 각각 y_1와 y_2의 투영된 스코어라 하자. 그러면 원문제는 Y_1u_1과 Y_2u_2사이의 상관계수를

$$\max_{u_1,u_2}\rho = corr(u_1^T Y_1, u_2^T Y_2) = \frac{u_1^T \Sigma_{12} u_2}{\sqrt{u_1^T \Sigma_{11} u_1}\sqrt{u_2^T \Sigma_{22} u_2}} \qquad (4\text{-}4)$$

최대화 하는 방향벡터 u_1과 u_2를 찾는 문제가 된다. 여기서, Σ_{11}와 Σ_{22}는 각 그룹의 급내변동행렬(within variance matrix)으로 공분산행렬이며, Σ_{12}는 급간변동행렬(between covariance matrix)이다.

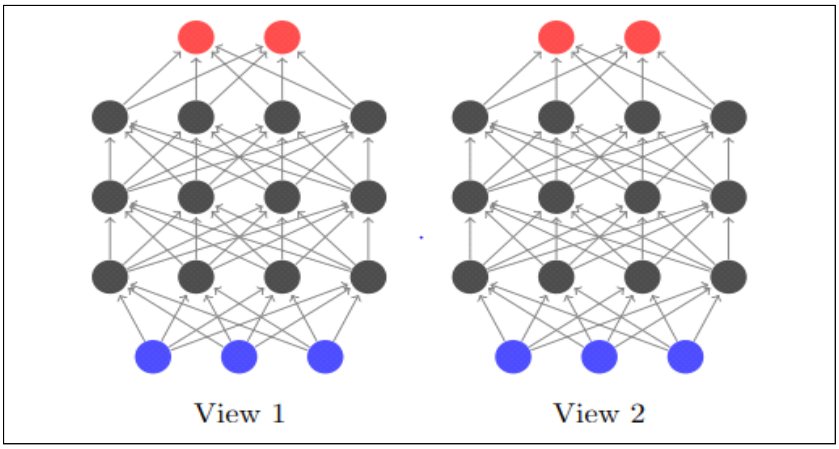
식 (4-4)에 대한 해는 변수집단 1과 변수집단 2의 공분산행렬에 관한 다음의 스펙트럴 분해(spectral decomposition)에 의한 고유벡터(eigen-vector)와 고유값(eigen-value)에 의해 주어진다.

$$\Sigma_{12}u_2 = \rho\Sigma_{11}u_1 \text{와 } \Sigma_{21}u_1 = \rho\Sigma_{22}u_2 \qquad (4\text{-}5)$$

여기서, ρ는 공통 고유값이다. 한편, 고유벡터 u_1과 u_2에 l_1-노음 제약조건을 통해서 고유벡터의 방향의 해석을 보다 용이할 수 있는데 대표적인 모형으로 Witten et al.(2009)의 연구를 예롤 들 수 있다.

여러 변형적인 방법 중에서 심층신경망 구조를 가진 정준상관분석모형(Andrew et al., 2013)에 대해서 살펴보도록 한다. 서로 다른 두 모달리티의 예로 한 관측치에 대한 두개의 이미지 view1와 view2가 주어졌다고 가정하자. 〔그림 4-10〕은 정준상관분석에서 3차원의 view1와 view2의 관측벡터들로부터 3층의 은닉층을 가진 심층 신경망 구조를 통과해서 최상단의 2차원의 임베딩 벡터로 표현됨을 나타낸다.

〔그림 4-10〕 심층 정준상관분석



자료: Andrew et al. (2013). "Deep canonical correlation analysis". Proceedings of the 30th International Conference on Machine Learning, 28(3), p. 1247-1255 재인용.

심층 정준상관분석은 다층의 비선형 변환을 통과하여 서로 다른 두 입력 view의 표현을 산출한다(〔그림 4-10〕 참조). 단순화를 위해 첫 번째 view에 대한 네트워크의 각 중간 계층에는 p_1 차원이라 하고 최종 (출력) 계층에는 크기가 o인 차원이라고 가정하자. 입력 $x_1 \in R^{n_1}$을 첫 번째 view의 객체라 하자. 추정모수 $W_{11} \in R^{p_1 \times n_1}$은 가중치 행렬, 절편을 $b_{11} \in R^{c_1}$이라고 놓자. 비선형함수 $s : R \rightarrow R$에 대해 출력값을 $h_1 = s(W_{11}x_1 + b_1) \in R^{p_1}$라고 하면 h_1은 다음 층 $h_2 = s(W_2 h_1 + b_2) \in R^{p_1}$의 입력으로 활용된다. 따라서 신경망이 d층일 때 최종 표현이 되는 값은 $f_1(x_1) = s(W_d h_{d-1} + b_d) \in R^o$이 된다. 두 번째 view에 대한 입력 $x_2 \in R^{n_2}$에 대해서도 표현 $f_2(x_2)$는 동일한 방식으로 계산할 수 있다.

〔그림 4-10〕에서 두 view에 대한 입력층은 각각 $n_1 = n_2 = 3$이다. 은닉층의 차원 $p_1 = p_2 = 4$, 최종 출력층의 차원은 $o = 2$이다. 중간 은닉층

의 개수는 $d=4$이다. 목적식은 양쪽의 입력이 주어졌을 때, 두 출력변수 간의 상관계수 $corr(f_1(X_1),f_2(X_2))$가 가능한 크게 하는 모수를 추정한다. 만약 θ_1는 첫 번째 view1에 대한 모수 W_{1l}과 b_{1l}, $l=1,\cdots,d$라 놓고, θ_2에 대해서도 view2에 대해 동일한 방식으로 매개화하면

$$(\hat{\theta}_1,\hat{\theta}_2)=argmax_{(\theta_1,\theta_2)}corr(f_1(X_1;\theta_1),f_2(X_2;\theta_2))$$

로 나타낼 수 있다. $\hat{\theta}_1$와 $\hat{\theta}_2$는 각 심층신경망에서의 추정된 가중치 행렬과 절편이다.

정준상관분석은 다중 모달리티 데이터 분석을 위해서 단순한 선형 상관관계를 다룬다. 표현의 다양성을 보다 확장하기 위한 방안으로 함수공간의 복잡도를 늘릴 필요가 있다. 비선형 상관관계를 추정하기 위해 신경망을 기저방법으로 한 다양한 정준상관분석의 변형방법들이 개발되었다. 다음 절에서는 표현학습에 기초하여 텍스트와 수치자료에 대한 정준상관분석의 여러 응용 방법들에 살펴본다.

제4절 딥러닝에 기반한 연계방법론 고도화 방안

이 절에서는 앞 절에서 제시한 임베딩 추출 방법들을 고도화하기 위한 방안에 대해 살펴보기로 한다. 특히, 비정형자료를 시간과 공간정보가 축척된 정형자료와 연계하기 위한 방법을 딥러닝 모형에 기초하여 살펴보기로 한다. 나아가 이 방법들의 대표적인 확장모형에 대해서도 알아본다. 끝으로 수치정보를 이미지화하여 합성신경망을 적용하는 방법과 보건복지 서비스 사용자와 서비스와의 관계를 통한 추천서비스 방법에 대해 알아본다.

1. 순환신경망

대표적인 비정형 자료인 텍스트 데이터에서 한 문장에서 추출한 단어는 맥락(context)에 의해서 의미론적인 중요성을 가진다고 할 수 있다. 개별 단어의 중요도를 기준으로 단순 합산하여 점수화 방식과 달리 전체 문장의 의미를 파악한 후에 전체 문장에서의 개별 단어가 가지는 가중치를 평가하는 방안을 고려해볼 수 있다. 또한 문장은 단어들에 따른 순서 구조가 가지고 있기 때문에 순서가 있는 시퀀스 자료(sequence data)(시간에 따라 서로 독립되지 않고 연관성, 의존성을 갖는 데이터)로 처리하는 방식을 고려할 필요성이 있다. 시간의 순서를 반영하는 다양한 방법들 중 이 절에서는 심층신경망 기법에서 대표적으로 많이 사용하는 방법인 순환신경망(recurrent neural net; RNN) 에 대해 간단히 살펴본다.

가. 순환신경망의 구조

순환신경망은 신경망에 시간의 개념을 도입하여 처리할 때 가장 기본적으로 활용되는 기법이다. 순환신경망은 주어진 t번째 입력정보에 대한 상응하는 출력을 생성하기 위해 t번째 입력(input)과 그 이전의 $t-1$번째 은닉상태(hidden state)의 정보를 활용하는 모델이다. 은닉상태의 노드(node)들은 방향을 가진 엣지(edge) 혹은 연결선으로 연결되어 있는 방향그래프의 구조를 가진 신경망의 한 종류이다. 순차적인 정보가 입력이 되고 이전 시점의 정보들이 자연스럽게 반영되므로 문장에서의 단어들 간의 순차적인 종속성을 반영할 수 있다. 그러므로 장기기억 의존성(long-range memory dependency)을 포착하는 모형을 만들 수 있다. 이를 수식으로 작성하면 다음과 같다.

x_t : 시간 t에서의 입력값(input)
h : 시간 t에서의 중간층(hidden layer)
$h = f(Ux_t + Wh_{t-1} + b_h)$
여기서 비선형 함수 f는 tanh, $h_0 = 0$ 초기화
$\hat{y}_t = softmax(Vh_t + b_y)$으로 시간 t에서의 출력값

각 층마다 모수의 값들이 일반적인 신경망과는 달리 시간에 의존하지 않는 모수(U, V, W)를 공유한다. 학습해야 하는 모수의 수가 감소한다. 해당 시점 이전의 정보를 모두 저장하는 중간은닉층 h_t는 네트워크 메모리라 부르기도 한다. 순환신경망은 시간에 따라 펼쳐진 구조에서 시간역전파 방법 BPTT(back propagation through time)을 사용하여 훈련을 한다. 순환신경망은 장기 기억의존성(long-term dependency)에 취약한 것으로 알려져 있다. 현재 시점 t에서 이 시점 이전의 먼 과거시점들

의 자료들의 특징을 잘 반영하지 못한다. 기본 순환신경망이 가진 장기 기억의존성 소실 문제를 개선하기 위한 여러 방안과 모형들이 제시되었다. 이 중 LSTM 모형(Hochreiter & Schmidhuber, 1997)이나 GRU(Cho et al., 2014) 모형을 대표적인 방법으로 들 수 있다.

나. Sequence to Sequence 모형

예를 들어, 언어를 다른 언어로 번역하는 학습문제에서 가장 널리 쓰이는 대표적인 모형은 sequence to sequence(seq2seq)이다. 순환신경망의 일종인 seq2seq 모형의 구조는 입력 문장을 받는 인코더와 출력 문장을 출력하는 디코더로 구성된다. 인코더는 입력 문장의 모든 단어들을 순차적으로 입력받은 뒤에 단어 정보들을 순환신경망을 적용하여 하나의 벡터로 만든다. 이를 맥락(context) 벡터라 한다. 입력 문장의 정보가 하나의 맥락 벡터로 모두 압축되면 이를 출력단계인 디코더로 전송하여 출력단계의 새로운 첫 입력정보로 활용하게 되고 순환신경망을 통해 각 시점에서 이전 시점까지 단어들이 주어졌을 때 하나씩 순차적으로 출력한다.

기본 seq2seq 모형은 크게 인코더와 디코더의 구조로 이루어져 있다. 인코더에서는 입력 시퀀스(sequence)를 하나의 고정된 크기의 맥락 벡터로 압축하고 이를 활용하여 디코더에서는 출력 시퀀스(sequence)를 만들어낸다. 그러나 seq2seq 모델에는 몇 가지 단점을 내재하고 있다. 순환신경망 계열이 가진 기울기 소실(vanishing gradient) 문제 외에 가장 큰 구조적인 단점은 하나의 고정된 크기의 벡터에 모든 정보를 압축하는 구조로 인해서 여러 시점에 관한 정보가 손실이 되는 점이다. 입력 시퀀스의 길이가 길어지게 되면 출력 시퀀스의 결과를 신뢰할 수 없게 된다. 이러한 단점을 개선하기 위한 방안으로 제안된 기법은 여러 시점에 대한

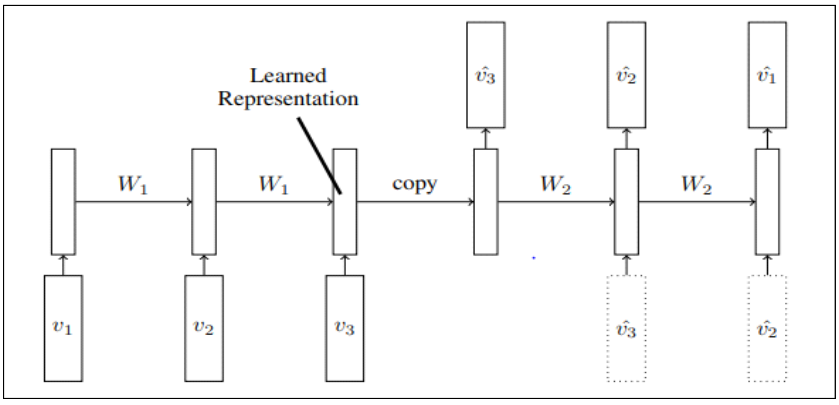
가중치를 반영하여 맥락벡터의 표현력을 높이는 방법이 제안되었으며 2절에서 소개한 어텐션(attention) 기법이 대표적인 방법이라 할 수 있다.

다. 시계열 자료에 대한 오토인코더

seq2seq 모형의 한 예로써 오토인코더 모형을 살펴보기로 한다. 입력값이 시계열의 종단구조로 얻어지는 경우, 시간에 따른 종속성을 반영할 수 있는 오토인코더를 고려할 수 있다. [그림 4-11]은 입력이 시계열 자료일 때의 오토인코더의 모형을 보여준다. 그림에서 학습된 표현이 임베딩벡터인 맥락벡터를 나타낸다. 전반부는 인코더 단계로 이때의 가중치 행렬은 W_1이고 후반부는 디코더 단계로 시점에 상관없이 공유되는 가중치 행렬은 W_2이다. 후반부의 첫 번째 입력은 인코더를 통한 맥락벡터가 되며 $\hat{v}_3$는 첫 번째 시점에서 추정된 값이며 그 다음의 출력시점의 입력으로 활용되게 된다.

순환신경망 기반의 오토인코더에서는 인코더단계는 입력 시퀀스마다 인코더를 수행하여 전체 시퀀스의 정보를 축약한 맥락(context) 임베딩 벡터로 변환한다. 만약 n을 출력 시퀀스의 시간의 길이라고 하면 디코더는 출력 시퀀스의 길이인 n번만큼 디코딩을 반복한다.

〔그림 4-11〕 LSTM 오토인코더 모형



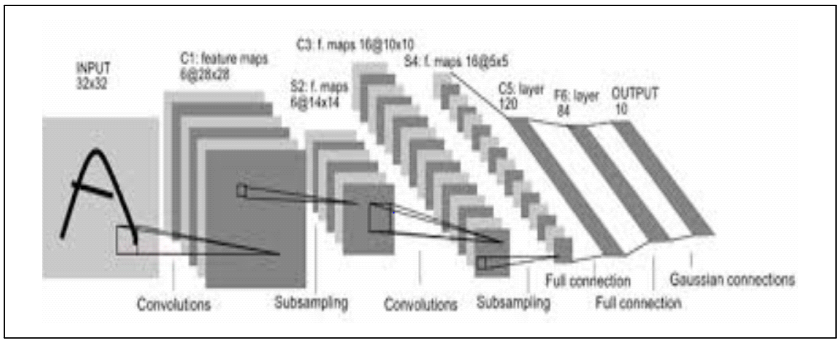
자료: Srivastava et al. (2015). "Unsupervised Learning of Video Representations using LSTMs", Proceedings of Machine Learning Research, 37, p. 843-852 재인용.

2. 합성망모형

가. 합성망모형

딥러닝이 대중화되고 그 중요성이나 활용성이 부각되게 한 주요한 요소로써, 풍부한 특징을 추출할 수 있는 특징공학(feature engineering)을 빠르게 학습할 수 있는 기법의 개발을 들 수 있다. 이미지 자료의 분류분석문제에서 대표적으로 활용되는 합성망모형(convolutional neural network)은 다단계 다층의 중첩된 컨볼루션(convolution)의 연산을 처음 실현한 모형이라 할 수 있다. 컴퓨터 비전분야에서 유래되었지만, 뛰어난 성능과 폭넓은 활용성으로 인해 합성망모형은 이미지분석 및 여러 분야의 분석에서 가장 널리 쓰이는 보편적인 방법으로 자리매김하였다. 〔그림 4-12〕는 LeNet-5 합성망모형으로 이미지로부터 심층 구조의 네트워크를 활용하여 다단계의 정보추출 과정을 학습하는 과정을 나타낸다.

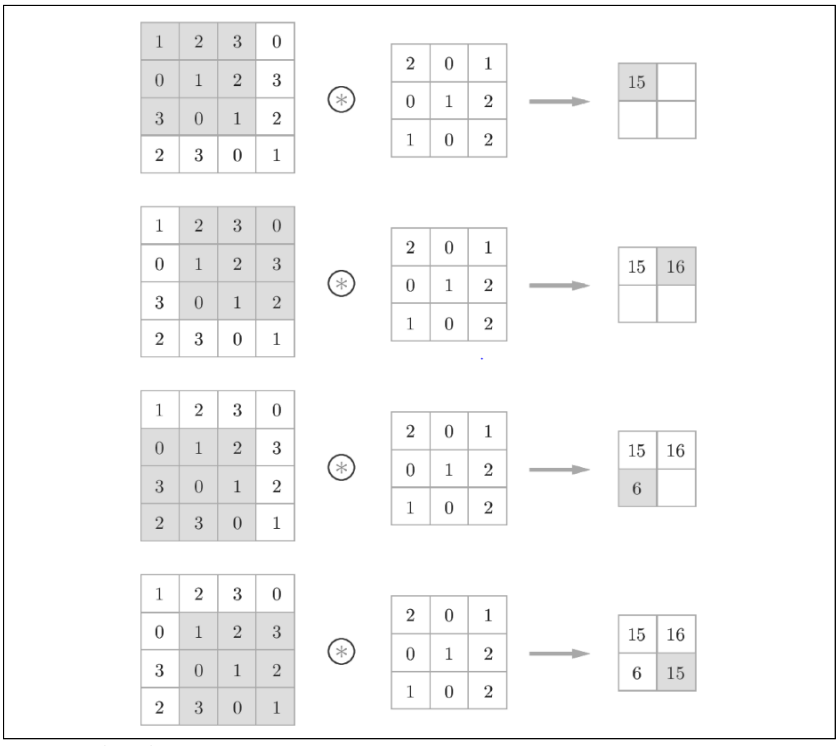
〔그림 4-12〕 LeNet-5 모형구조



자료: LeCun et al. (1998). "Gradient-based learning applied to document recognition", Proceedings of the IEEE, 86(11), p. 2278-2324, 재인용.

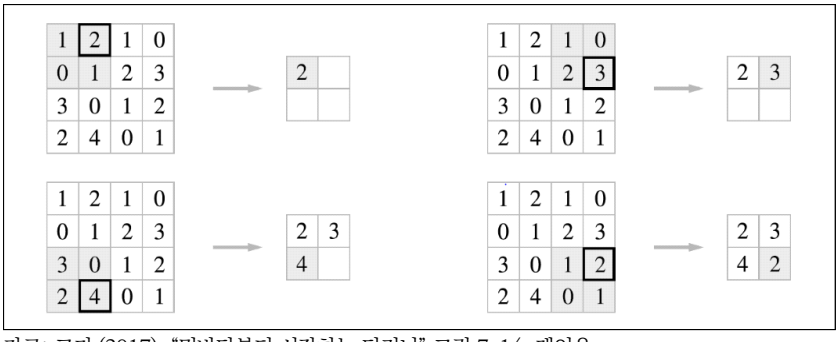
역전파를 통해 주어진 목적함수를 최소화하는 방법으로 적절한 특징(feature)을 추출하는 학습을 진행한다. 이는 종래의 컴퓨터 비전의 특징 공학에 의존한 처리가 합성망 모형을 통해 관련성 있는 특징을 다수의 입력과 네트워크 구조를 통해 학습하여 찾는 과정으로 대치되었다고 할 수 있다.

〔그림 4-13〕 합성곱층(convolution layer)



자료: 고키 (2017). "밑바닥부터 시작하는 딥러닝" 그림 7-14, 재인용.

〔그림 4-14〕 풀링층(pooling layer)



자료: 고키 (2017). "밑바닥부터 시작하는 딥러닝" 그림 7-14, 재인용.

〔그림 4-13〕과 〔그림 4-14〕는 각각 합성곱층과 풀링층에서 활용되는 필터(filter)를 나타낸다. 커널(kernel)이라고도 부른다. 합성망모형에서의 필터를 통한 특징추출은 이미지 인식에서 특징은 이미지내의 이동 불변성(translation invariance)에 주로 근거를 둔다. 여기서, 이동 불변성은 객체(object)를 인식하는 것은 그 객체가 이미지 내 특정 위치에 상관없이 동일한 영향을 가질 때를 말한다(LeCun et al., 1998)(〔그림 4-12〕 참조). 그러나 동일 부분 이미지이더라도 이미지내의 특징의 위치의 차이에서 오는 다른 의미성도 중요할 수 있다. 이는 사물을 인식하는 것 뿐만이 아닌 사물의 위치를 인지하는 것이 필요하기 때문이다. 그러므로 2절의 텍스트자료에서 살펴본, 문장의 맥락(context)과 동일한 개념인 이미지의 맥락을 추출할 필요성이 있다. 이와 같이 이미지를 의미를 가지는 부분이미지로 분할하는 의미론적 분할(semantic segmentation) 문제에서 주로 관심을 두는 문제가 된다.

합성망모형에서는 하위 계층부터 상위 계층을 통과하면서 점차 높은 수준의 특징을 추출한다. 하위 계층에서는 주로 복수의 합성곱과 풀링층을 반복적으로 적용하여 특징맵을 구성한다. 합성곱 층에서는 직전단계층의 값을 입력받아 공유된 가중치 연산처리를 수행하며 최대값 풀링층에서는 필터 내의 값 중 최대값을 취함으로써 특징이 되는 값은 보존하고 특징맵의 크기를 줄여 연산을 빠르게 해주는 역할을 한다. 최상위 완전연결 층에서는 추출된 최상단층의 특징을 입력으로 하여 학습문제를 푼다.

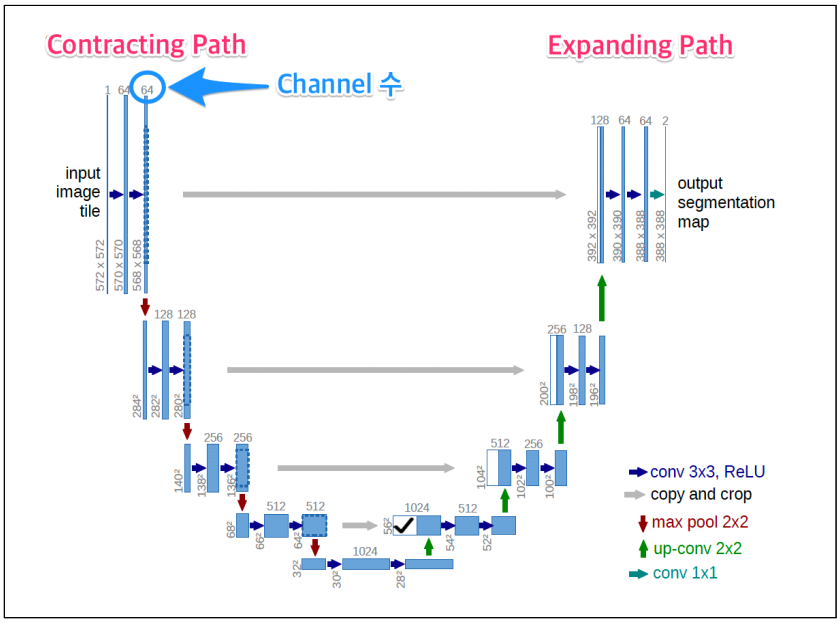
나. 합성망모형의 확장

이미지를 인식하는 기존 구조에서 다양한 변형적 모형이 파생되었으며, 대표적으로 U-net, W-net 등의 합성망 모형을 예로 들 수 있다.

체인형태의 네트워크 연결구조가 가지는 단순한 정보전달력을 개선하기 위한 방안으로 개발되었다. 이러한 모형은 직전의 여러 은닉층이 현재 은닉층에 미치는 영향력을 반영할 수 있는 구조를 가지고 있다. 〔그림 4-15〕는 U-net을 뜻하고, 〔그림 4-16〕은 W-net의 모형구조를 나타낸다.

U-net은 바이오메디컬 분야에서 이미지 분할(image segmentation)을 목적으로 제안된 end-to-end 방식의 완전연결망(fully convolutional network) 기반 모델이다. 전반적인 네트워크 구성의 형태가 U자 형태를 띠고 있다. U-Net은 이미지의 전반적인 맥락 정보를 얻기 위한 네트워크와 정확한 지역화(localization)를 위한 네트워크가 대칭 형태로 설계되어 있다. U-net의 전반부는 contracting path라 부르며 후반부는 expanding path 단계라 부른다. 이러한 구조는 2절에서 살펴본 오토인코더의 두 단계와 동일하다고 할 수 있으며, 순환신경망에서는 인코더와 디코더의 구조와 유사하다고 할 수 있다. 그림의 전반부와 후반부를 가로지르는 직선은 입력으로 전반부의 층과 출력으로 후반부의 동일 층의 관계를 모형화할 수 있음을 나타낸다. 전반부는 입력 이미지의 차원의 크기가 572×572부터 시작하여 28×28로 축소되어 정보가 응축이 되며, 후반부는 28×28을 입력으로 하여 388×388의 크기로 확대(up-sampling)되어 출력된다.

〔그림 4-15〕 U-net의 구조

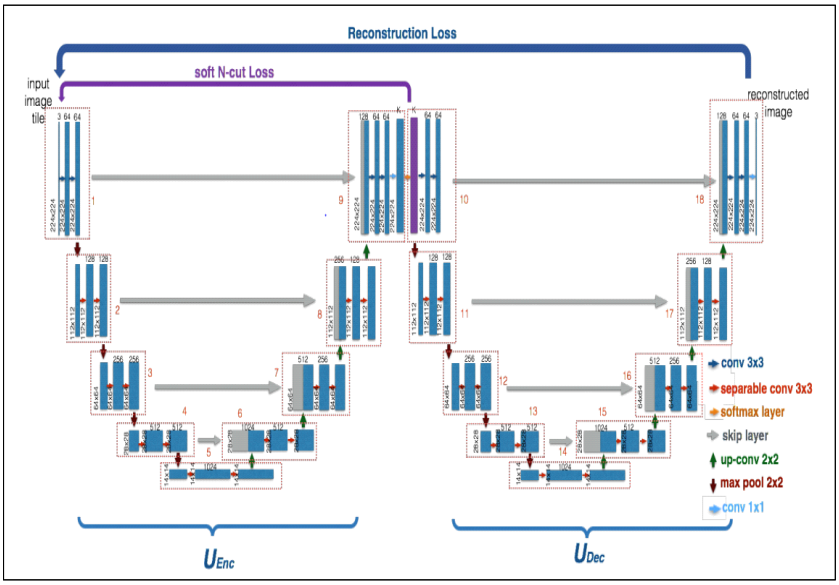


자료: Ronneberger et al. (2015). U-Net: convolutional networks for biomedical image segmentation, ArXiv, abs/1505.04597, 재인용.

Contracting path는 입력이미지의 수축을 통한 문맥(context) 정보를 추출하기 위한 인코더 과정에 해당된다. 보통 VGG(Simonyan & Zisserman, 2014) 계열의 합성망이 활용된다. 〔그림 4-15〕의 U-Net 구조에서 특징맵 추출과 생성이 깊은 구조로 반복되므로 입력 저층 부분의 층에서 고층으로 연결되는 skip connection을 추가하였다. 여기서, skip connection은 훈련자료에서의 간결하거나 얕은(shallow) 모형보다 딥모형의 훈련오차가 개선이 잘 이루어지지 않은 경우(degradation)를 개선하기 위하여 고안되었다(He et al., 2016). 바이오메디컬 분야에서는 이미지의 영역별로 높은 해상도의 결과를 도출할 필요성이 있기 때문에 저차원의 특징맵을 up-sampling 과정을 통해서 고차원으로 확장

한다. U자형의 구조를 통해 특징맵 추출과 up-sampling을 수행한다.

[그림 4-16] W-net의 구조



자료: Xia and Kulis (2017). W-Net: a deep model for fully unsupervised image segmentation, ArXiv, abs/1711.08506, 재인용.

이 외에도 확장된 합성곱(dilated convolution 또는 atrous convolution)는 합성곱 커널 필터의 내부에 0값을 넣는 방법으로 가로 세로 모두에 적용하면 가중치(weight)의 개수를 늘리지 않고 윈도우 크기(window size)를 늘릴 수 있는 방법으로 동일 객체이나 척도가 상이한 이미지를 인식하는데 매우 효과적이다.

다. 해석 가능한 이미지 영역 추출

이미지 영역의 중요도 파악문제에서 유용하게 활용되는 CAM(class activation map)은 합성망의 해석을 명시적으로 해주는 방법이라 할 수 있다. CAM의 구조를 살펴보자. CAM의 최상단 층은, 여러 채널의 결과를 합산한 값(global average pooling; GAP)을 소프트맥스를 통해 확률값을 산출한다. k번째 채널의 값 중 (x,y)에 위치한 값을 $f_k(x,y)$라고 표현하고, 합산값 F_k는 $\sum_{x,y} f_k(x,y)$가 된다. 클래스 c에 대해 소프트맥스 값은

$$S_c = \sum_k w_k^c F_k$$

이다. 여기서 w_k^c는 클래스 c에 대한 k번째 채널 F_k의 중요성을 나타낸다고 할 수 있다. 즉 w_k^c가 클수록 c에서 F_k가 미치는 영향은 커진다. $F_k = \sum_{x,y} f_k(x,y)$로부터

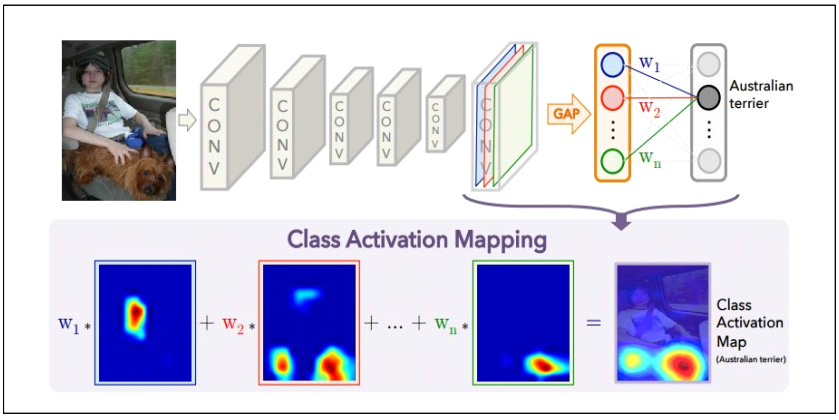
$$\begin{aligned} S_c &= \sum_k w_k^c F_k \\ &= \sum_{x,y}\sum_k w_k^c f_k(x,y) \\ &= \sum_k w_c^k \sum_{x,y} f_k(x,y) \end{aligned}$$

가 된다. M_c를 클래스 c에 대한 CAM이라 하고

$$M_c(x,y) = \sum_k w_k^c f_k(x,y)$$

라고 정의한다면 결국 $S_c = \sum_{x,y} M_c(x,y)$를 얻을 수 있다. 즉 $M_c(x,y)$는 (x,y)에 위치한 값이 c라는 클래스 분류에 미치는 중요도를 나타내는 것으로 결국 어텐션과 비슷한 역할을 한다고 할 수 있다.

〔그림 4-17〕 CAM 방법의 모형구조



자료: Zhou et al. (2016). Learning deep features for discriminative localization, 2016 IEEE Conference on Computer Vision and Pattern Recognition(CVPR), p. 2921-2929, 재인용.

〔그림 4-17〕에서 GAP을 취하기 전에 위치한 특징맵들이 f_k가 된다. 이를 채널 단위로 합해주면 F_k(그림의 각 하나의 영역)가 되고 이를 완전연결계층을 통한 소프트맥스를 산출 시 가중치를 w_k^c로 한 가중합산을 한다. 어떤 클래스 c로 분류될 확률을 구할 때 곱해지는 각각의 가중치들을 특징맵에 곱해준 다음 가중 합산한다.

라. 이미지 분석목표의 여러 유형

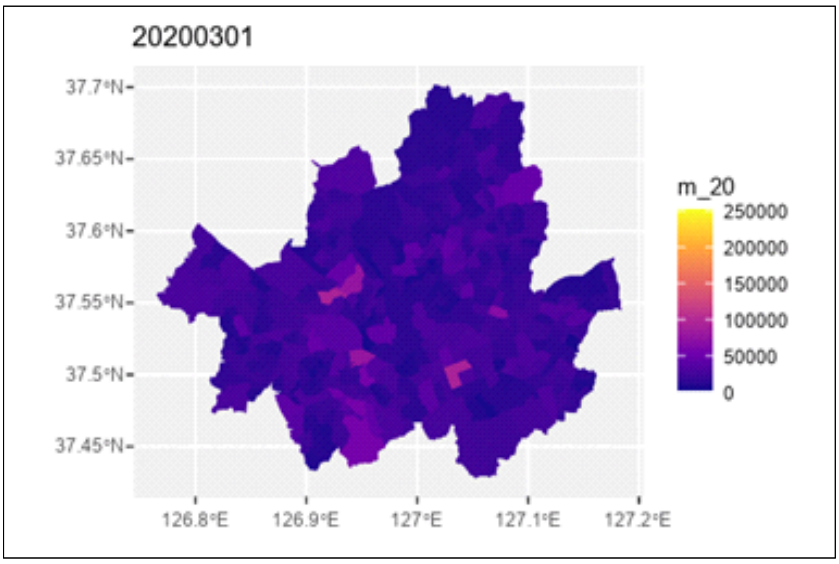
이미지 자료의 분석목표의 대표적인 4가지 유형을 살펴보기로 한다.

첫 번째 유형으로는 이미지의 클래스를 분류하는 문제가 있다. 이미지 전체 클래스를 분류하고자 하는 것이 목적으로 보통 상위 5개의 클래스로 분류하고 이를 태깅한다. 다음으로 이미지내의 여러 개의 다중 오브젝트 분류를 다루는 다중분류(multi-label classification) 문제를 들 수 있다. 이미지 분류문제를 보다 확장한 것으로 이미지의 분류클래스가 다중으로 주어진 경우이다. 세 번째는 지역화(localization) 개념을 결합한 연구 분야로 이미지내의 객체의 클래스 분류와 동시에 위치를 찾는 문제이다. 보통 바운딩박스(bounding box)라는 개념을 활용하여 이미지내의 객체를 탐지하고 위치를 찾는데 이를 오브젝트 탐지(object detection)라 한다. 마지막 문제는 이미지분할(image segmentation)로서 오브젝트 탐지(object detection)와 오브젝트의 경계와 형체까지 탐지하는 문제를 말한다. 이미지가 가지는 특징추출을 통해 기본적인 분류부터 이미지의 의미를 파악하는 문제까지 점차 복잡해짐을 확인할 수 있다.

3. 합성망모형을 통한 비정형빅데이터의 활용성 확장

수치/범주형 값으로 작성한 그래프 등의 데이터를 이미지로 간주하면 전술한 합성망모형을 활용할 수 있게 되며, 비정형 자료인 텍스트 데이터와 결합 방법 등 다양한 방안을 고려할 수 있게 된다. 여기에서는 비정형 빅데이터 활용성 확장을 위한 한 방안으로 서울시 공공 빅데이터에서 제공하는 다양한 수치형 데이터를 지도에 시각화하고([그림 4-18] 참조), 이를 합성망모형으로 분석하는 방법에 대해 다루기로 한다.

〔그림 4-18〕 2020년 3월 1일 서울시 관내의 생활인구(20대)



자료: 저자 작성

합성망 모델을 이용하여 이미지-정형데이터 결합을 통해 서울시생활인구 자료의 예를 들어 설명하기로 한다. 예시로 활용한 데이터는 서울시 생활인구자료로 성별, 연령대별로(10세 미만, 10~79세[5세 단위], 80세 이상) 매 시간마다 서울시 행정구역 공간상에 통신자료를 근거로 하여 식별된 인구수 자료이다. 단, 개인정보 비식별화를 위하여 '3명' 이하인 경우 "*" 처리되었다.

본 연구에서는 2020년 3월의 442개 행정동별 생활인구자료를 활용하여 20대와 70대의 두 연령대에 따른 행정동의 권역별 생활인구 유동량의 차이를 시각적으로 확인하여 본다. 〈표 4-1〉은 3월 1일 서울시 관내의 20대의 생활인구 자료에 대한 구체적인 정보를 나타낸다.

〈표 4-1〉 자료의 설명

구분	내용
1	3월의 시간대별 통신자료를 근거로 생활인구
2	9시-18시 주간 데이터 활용(1개월*30일*10시간*2집단=600장 이미지)
3	합성망을 적용하여 20대와 70대 집단 분류분석 수행
4	이미지의 가로, 세로를 5개씩 나누어 총 25개의 권역으로 나누어 분류에 영향력이 있는 지표를 권역별 가중치 추정(attention weight)을 통해 살펴봄

수치정보를 이미지로 변환 및 학습에 활용하기 위해 다음의 절차에 따라 진행하였다.

① 수치정보가 포함된 이미지에서 범례 등 공간상의 불필요한 정보를 나타내는 이미지 영역은 삭제한다.

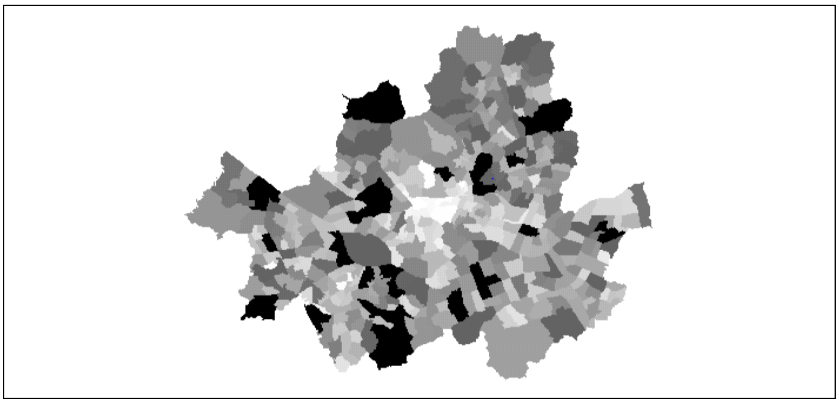
② 수치를 행정구역의 표시된 공간상에 흑백으로 픽셀값을 표현하고 지도에 표시한다.

- 흑백의 픽셀값은 0-255의 값을 가지므로 3월달의 집단별 최대인구수(픽셀값 255)를 기준으로 최대값이 1, 최소값이 0이 되는 0-1 표준화를 시행함.

③ 한 이미지를 25개 권역으로 분할하여 25장의 이미지를 구성한다.

④ 25개 분할된 부분 이미지에 어텐션 벡터를 가진 합성망을 적용한다.

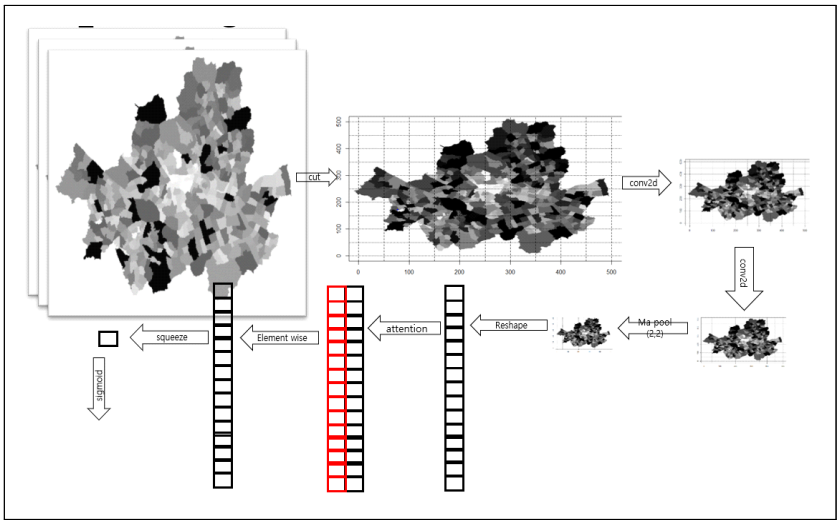
〔그림 4-19〕 2020년 3월 1일 서울시 관내의 생활인구(20대)



자료: 저자 작성

〔그림 4-19〕는 표준화작업 등 전처리 과정을 거친 수치정보를 표시한 이미지이다. 그림에서 흰색에 가까울수록 3월 1일 9시 해당 시간대에 생활인구가 적은 지역을 나타내고, 검은색에 가까울수록 생활인구가 많은 지역을 나타낸다. 〔그림 4-20〕은 연구에서 활용한 합성망 모형을 나타낸다. 〔그림 4-21〕은 어텐션 벡터를 추가한 합성망모델을 추정결과이다. 25개 권역별 가중치는 각 그림에서 영역의 색이 짙을수록 두 집단을 분류하는 정도가 상대적으로 높은 권역임을 나타낸다. 색깔이 진한 영역이 20대와 70대의 인구집단의 생활인구 차이가 큰 지역을 의미한다.

〔그림 4-20〕 합성망모형



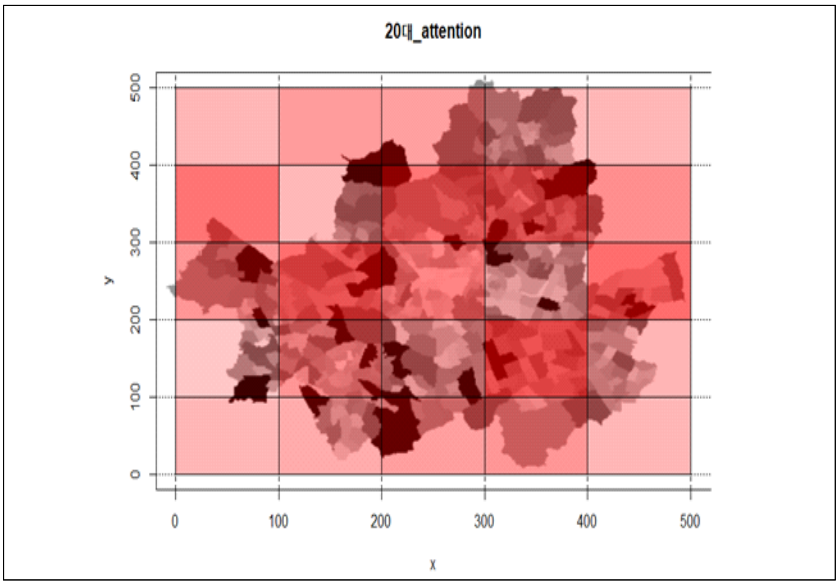
자료: 저자 작성

이러한 결과를 바탕으로 다음의 두 가지 방안으로 보다 세밀한 분석을 수행할 수 있을 것으로 기대한다.

① 공간상 단위의 세분화 행정구(약 440개)를 선거집계구단위(약 14,000개)로 세분화하여 더 정밀한 모형을 생성할 수 있다.

② 다양한 보건복지 정보들을 다채널(channel) 이미지로 텐서(tensor) 입력이미지로 입력을 통하여 공간과 수치가 복합적인 자료에 대해서 여러 사회적 수치에 대한 이미지를 여러 채널(channel)로 여기고 이를 적용할 수 있다. 이는 여러 종류의 시공간상의 정보를 이미지 데이터로 표현하고 출력변수와 연계할 수 있음을 시사한다. 또한, 어텐션 벡터의 가중치를 통해 연령별로 행정동, 자치구에 따른 입력정보의 차이가 발생하는지를 확인할 수 있게 된다. 이를 통해 보건복지정보의 상대적으로 비교우위나 비교열위 지역을 효과적으로 파악할 수 있다.

〔그림 4-21〕 25개 구역의 어텐션 벡터



자료: 저자 작성

4. 추천시스템 알고리즘을 통한 비정형빅데이터 활용성 확장

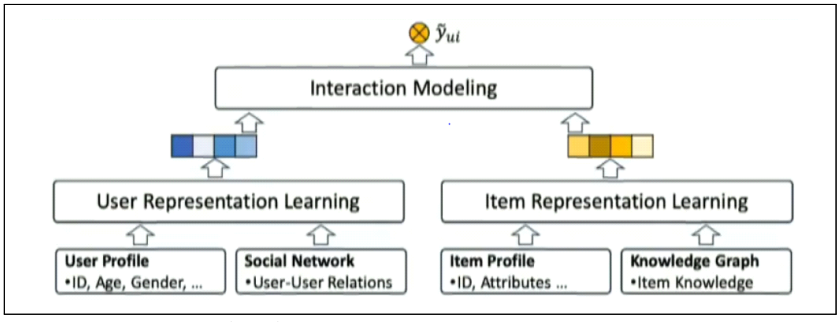
시간이 흘러감에 따라 사용자의 생애의 모습이 달라지거나 환경변화가 일어나므로 보건복지서비스에 대한 수요가 변화된다. 또한, 생애주기별로 서비스를 제공하는 서비스 주체자도 변화하게 되거나 추천되는 공공아이템도 달라진다. 향후 이러한 수요에 시의적절하게 제공할 수 있는, 적응형 공공 서비스의 중요성이 점차 부각될 것이라고 예상되는바 추천시스템 알고리즘 연구결과를 리뷰하고 이에 대한 적절한 제언을 담고자 한다. 아울러 시간 정보도 포함된 보건 자료 비정형 자료 분석에 비정형빅데이터 활용성 확장 방안을 모색해본다.

이 절에서는 정준상관분석, 임베딩 등 앞에서 설명한 주요한 방법들에

대한 다양한 예를 들어 이해를 돕고자 한다. 우선 보건사회와 관련된 공공서비스 등을 아이템, 이에 대한 수혜자는 사용자라 매칭하자. 그리고 이 들 간의 관계는 아이템과 사용자 사이의 다양한 행위의 결과 혹은 형태라 놓자. 이를 기초로 사용자-아이템(user-item) 등의 추천시스템을 서울시 공공데이터들의 연계분석에 적용할 수 있다.

보다 정확하고 다양하며 설명 가능한 추천을 제공하려면 사용자 항목 상호 작용 모델링을 넘어서 부가 정보(side information)를 고려해야한다. 지식그래프와 사용자 항목 그래프의 융합 구조에서 두 항목을 하나 또는 여러 개의 연결된 속성으로 연결하는 고차 관계가 성공적인 추천을 위한 필수 요소라고 할 수 있다. 노드의 인접 항목(사용자, 아이템 또는 엔터티가 될 수 있음)에서 임베딩을 재귀적으로 전파하여 임베딩을 갱신하고 어텐션 구조를 사용하여 관심 있는 아이템의 인접 아이템의 중요도를 계량화하고 판별할 수 있다. [그림 4-22]는 사용자, 사회관계망, 아이템 정보, 지식정보 등 가용한 입력정보로부터 임베딩과 이 임베딩벡터의 관계에 대한 모형을 나타낸다.

[그림 4-22] 사용자 임베딩과 아이템 임베딩의 작용 모델

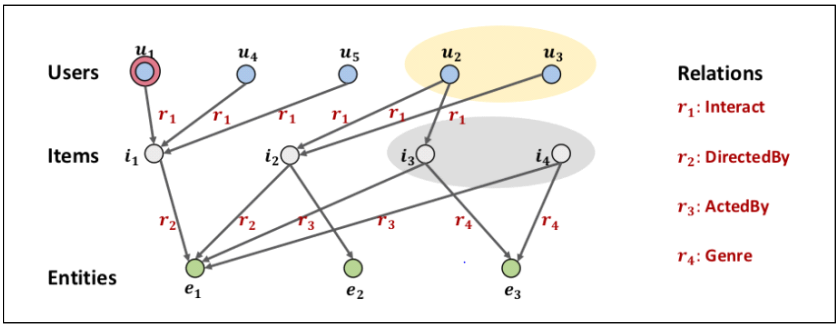


자료: Wang et al., (2019). "KGAT: Knowledge Graph attention network for recommendation", Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 재인용.

Wang et al.(2019)이 제안한 어텐션 기능을 가진 협동 지식그래프에 대한 방법론을 소개한다. [그림 4-23]은 영화에 대한 지식그래프의 한 예이다. 하나의 영화는 영화감독, 캐스트, 장르 등에 의해 설명되어질 수 있다. 그림에서 u는 사용자(user), i는 아이템(item)을 나타낸다. 사용자가 활용하는 서비스 정보 e는 엔터티(entity)를 나타내는 것으로 아이템과 연관성이 있는 부가 정보를 뜻한다. r은 관계(relation)를 뜻하는 것으로 엔터티와 아이템간의 관계, 아이템과 사용자간의 관계를 나타낸다.

협동필터링(collaborative filtering) 방법은 프로파일상 유사한 패턴을 보이는 사용자의 사용정보에 중점을 두는데 사용자와 아이템간의 관계만을 주로 반영하여 다룬다고 볼 수 있다. [그림 4-23]에서는 추천대상인 사용자 u_1에 아이템 i_1에 같이 반응한(interact) 사용자 u_4와 u_5의 특징에 주로 참고하여, u_4와 u_5이 반응한 아이템 중 u_1이 아직 반응하지 않은 아이템이 있다면 이 아이템을 추천을 하는 방식이다.

[그림 4-23] 지식그래프(knowledge graph)의 예시(Wang 등, 2019)

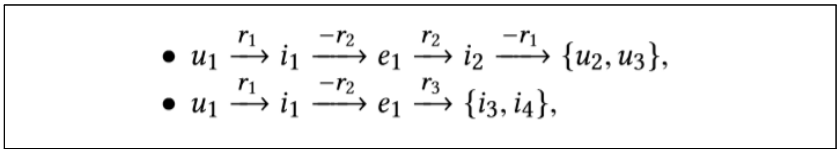


주: u_1은 target user이며, u_2부터 u_5는 다른 user들이다. i_1, i_2, i_3, i_4는 4종류의 아이템, e_1, e_2, e_3은 엔터티임.

자료: Wang et al., (2019). "KGAT: Knowledge Graph attention network for recommendation", Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 재인용.

반면 협동지식그래프(collaborative knowledge graph)는 사용자-아이템(user-item) 그래프 외에 지식그래프(knowledge graph)를 동시적인 정보를 담고 있다. 여기서 사용자-아이템 그래프는 u-i의 그래프를 의미하며 지식그래프는 e와 r의 그래프를 의미한다. 따라서 지식그래프를 활용한 추천시스템의 핵심 요소는 고차의 다항이 관련된 관계(high-order relations)를 탐색하는 것에 있다고 볼 수 있다.

〔그림 4-24〕 지식그래프에서의 도달 경로(path)의 예시



자료: Wang et al., (2019). "KGAT: Knowledge Graph attention network for recommendation", Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.

협업 지식 그래프에서 아이템에 연결된 부가 정보를 활용하여 보다 나은 추천을 제공하기 위해 고차 관계를 명시적으로 모델링하는 것이 중요하다. 〔그림 4-24〕는 경로상의 고차단계의 관계를 탐색하기 위한 과정을 예시적으로 보여준다. 첫 번째 경로는 u_1에서 4단계의 관계를 거쳐 u_2나 u_3로 도달과정을 보여주며, 두 번째 경로는 u_1에서 i_3와 i_4까지의 도달 경로를 나타낸다. 지식그래프를 활용한 방법에서는 도달 경로마다 가중치가 달라지기 때문에 적절한 가중치 연산을 통해서 이를 반영할 수 있다.

가. 임베딩 층

그래프에서 head와 tail간의 관계(relation)가 주어진 그래프가 있다고 가정하자. 서로 다른 모달리티(modality)간의 관련성은 번역원리(translation principle)에 의해 모형화할 수 있다. 즉, 한 모달리티는 다른 모달리티로의 변환이 자연스럽게 '번역'이 최대화되어야 함을 뜻한다.

이러한 목표 하에 엔터티와 관계의 임베딩을 학습시키는 과정을 수리적 유도를 통해 살펴보자. 먼저 $e_{h,}e_t \in R^d$, $e_r \in R^k$를 각각 h,t,r의 임베딩 벡터이며, e_h^r와 e_t^r는 각각 e_h와 e_t의 관계 공간으로의 투영한 벡터를 나타낸다고 하자. h는 경로 상에서 관계의 시작점을 뜻하며, t는 경로에서 종착점을 나타낸다. 그러므로

$$e_h^r + e_r \approx e_t^r$$

와 같은 관계를 성립하는 것을 기대할 수 있다. 따라서 다음의 세 임베딩 벡터에 대한 손실함수를 고려하고

$$g(h,r,t) = \| W_r e_h + e_r - W_r e_t \|_2^2$$

이를 최소화하는 W_r를 학습하는 문제를 구성할 수 있다. 여기서, $W_r \in R^{k \times d}$는 관계 r에 대한 d차원의 엔터티로부터 k차원 관계 공간으로 변환하는 행렬이다. 또한, $g(h,r,t) = \| W_r e_h + e_r - W_r e_t \|_2^2$값이 작을수록 세 항들 간의 관계에 대한 설명력이 높다고 할 수 있다. 세 항목간의 정의역은 다르나 상관성 분석에 활용되는 벡터는 임베딩을 통해서 사칙연산을 수행할 수 있게 된 점을 주목할 수 있다.

임베딩이 주어지면 훈련자료에서 관찰된 세 항목들 간의 관계 (h,r,t)

와 관찰되지 않은 관계 (h,r,t')간의 쌍별 랭킹 손실(pairwise ranking loss)을 통해 손실을 계량화할 수 있다.

집합 $\Gamma = \{(h,r,t,t')|(h,r,t) \in G, (h,r,t') \notin G\}$에 대해 지식그래프에서의 손실은

$$L_{KG} = \sum_{(h,r,t,t') \in \Gamma} -\log \sigma(g(h,r,t') - g(h,r,t)) \qquad (4\text{-}6)$$

으로, 여기서 $\sigma(\cdot)$는 시그모이드(sigmoid) 함수이다. 이외에도 다양한 랭킹 손실함수를 고려할 수 있다.

나. 어텐션

하나의 엔터티 e는 여러 개의 관계에 포함될 수 있다. 그래프의 엣지 연결강도인 가중치에 어텐션을 적용하는 방안으로 그래프의 중심에서 그래프 상에서 합성망 방법론 적용할 수 있다. 사용자-아이템 이분그래프(user-item bipartite graph)는 추천시스템 활용 시 구매이력 정보를 활용한다. 예를 들어, 구매여부와 클릭 수 등을 들 수 있다. 사용자와 아이템간의 상호작용 자료를 이분그래프의 구조 G_1로 이해할 수 있다.

U와 I는 각각 사용자들의 집합과 아이템의 집합이라고 놓자. 그러면 이분그래프의 구조 G_1의 원소는 $\{(u,y_{u_i},i)|u \in U, i \in I)\}$로 나타내어진다. y는 링크를 나타내고 이진값을 가진다. 가령 $y_{u_i} = 1$이면 사용자 u와 아이템 i간에 상호작용이 있음을 나타낸다.

집합 $\Omega = \{(u,i,i')|(u,i) \in R^+, (h,r,t') \in R^-\}$이며, R^+와 R^-는 각각 사용자와 아이템간의 관찰된 상호작용과 관찰되지 않은 상호작용 집합을 뜻한다. $y(u,i)$는 사용자 임베딩, i는 아이템 임베딩 벡터간의 내적

이라 하면 다음의 협동 필터링에 관한 랭킹 손실을 L_{KG}와 같이 정의할 수 있다.

$$L_{CF} = \sum_{(u,i,i') \in \Omega} -\log \sigma(y(u,i) - y(u,i')) \qquad (4\text{-}7)$$

이제 식 (4-6)과 식 (4-7)을 동시에 고려하여 $L_{KGAT} = L_{KG} + L_{CF}$을 최소화하는 임베딩벡터를 추정하는데 필요한 모수를 학습하면 된다.

〈표 4-2〉는 협동지식그래프 추천시스템방법에서의 입력과 최종 출력을 나타낸다.

〈표 4-2〉 협동지식그래프 추천시스템방법의 입력과 출력

구분	내용
입력	사용자-아이템 이분그래프 G_1과 지식그래프 G_2
출력	사용자가 u가 아이템 i를 수용할 확률 $\Pr(y_{ui} = 1)$를 산출

이러한 협동지식그래프를 통한 추천시스템방법을 활용하면 보건복지 서비스 수혜자 u에게 서비스 아이템 i를 추천할 확률을 산출할 수 있을 것으로 기대한다.

이상을 정리하면

① 추천시스템을 통한 임베딩방법론은 특정 사용자가 개별 아이템을 구매하는 이를 기반으로 추천하는 시스템에서 사용자와 아이템 모두 심볼릭(symbolic)의 자료구조로 여기로 이를 임베딩하여 수치적 공간에서 다양한 기계학습 알고리즘을 적용할 수 있다.

② 사용자와 공공서비스에 새로운 아이템이 유입되어 새롭게 데이터가 구성되어 유입되더라도 추천모형에 관한 실시간학습을 진행할 수 있다.

5. 심층 정준상관분석과 심층 오토인코더 등을 통한 비정형빅데이터 활용성 확장

Andrew et al.(2013)에서 활용한 입력 데이터가 view1(영상)과 view2(자막)과 같이 서로 다른 관점(view)에서 데이터들이 수집된다고 하자. 심층 오토인코더와 정준상관분석을 이용해서 서로 다른 모달리티의 데이터들을 비선형 변환을 거친 후 상관관계를 높게 하는 잠재변수 z와 선형변환(정준상관분석의 선형변환)을 구할 수 있다. 만약 view2이 없더라도 학습된 오토인코더와 정준상관분석의 선형변환을 통해 view2의 정보를 view1에 반영할 수 있다.

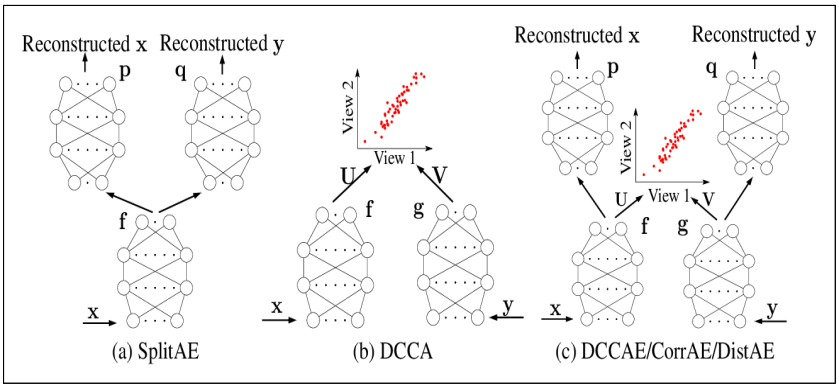
n을 자료의 크기라 할 때, 두 view에 대한 쌍 관측 자료 $(x_1, y_1), \cdots, (x_n, y_n)$에 대해 $x_i \in R^{D_x}$와 $y_i \in R^{D_y}$라 하자. 각 view에 대한 자료 행렬 $X = [x_1, \cdots, x_n]$과 $Y = [y_1, \cdots, y_n]$라 놓자. 또한, 심층신경망을 통해 투영된 산출행렬을 $f(X) = [f(x_1), \cdots, f(x_n)]$라 두자.

Ngiam et al.(2011)은 평가 시 하나의 view로부터 두 view를 모두 복원하여 공유된 표현을 추출하는 방법을 제시하였다. G_1와 G_2를 심층신경망 구조를 가진 그래프라 하자(그림에서는 $p = G_1$와 $q = G_2$). 〔그림 4-25〕의 (a)에 개략적으로 제시된다. 이 방법에서 특징맵을 추출하는 네트워크 f는 공유된다. 목적식은 두 view에 대한 복원오차의 합을 최소화하는 다음의 식으로

$$min_{W_f, W_{G_1}, W_{G_2}} \frac{1}{n} \sum_{i=1}^{n} ||x_i - G_1(f(x_i))||_2^2 + ||y_i - G_2(f(x_i))||_2^2$$

정의된다.

〔그림 4-25〕 심층정준상관분석 모형 구조들



자료: Wang et al. (2015). "On deep multi-view representation learning", Proceedings of the 32nd International Conference on International Conference on Machine Learning, 37, p. 1083-1092, 재인용.

정준상관분석은 직교성 제약조건을 만족하면서 최대화하는 최적화문제로 구성될 수 있다. 〔그림 4-25〕의 (b)의 심층 정준상관분석은 모수에 관한 제한조건 집합을

$$C(U,V) = \{(U,V) : U^T g(\frac{1}{n} f(X) f(X)^T + \lambda_x I) U = I,$$
$$V^T g(\frac{1}{n} g(Y) g(Y)^T + \lambda_y) V = I,$$
$$u_i^T f(X) g(Y)^T v_j = 0 \ \ for \ \ i \neq j\}$$

라 두자. 그러면 모수행렬 $U = [u_1, u_2, \cdots, u_L]$, $V = [v_1, v_2, \cdots, v_L]$일 때, 제약조건 $U, V \in C(U,V)$을 모수로 한 심층 정준상관분석은

$$min_{U,V \in C(U,V), W_f, W_g} - \frac{1}{n} tr(U^T f(X) g(Y)^T V)$$

문제를 푼다. 〔그림 4-25〕의 (c)의 심층 정준상관분석을 수행하는 오토인코더에 대한 적합도는

$$min_{\ U, V \in C(U,V), W_f, W_g, W_{G_1}, W_{G_2}} - \frac{1}{n} tr\left(U^T f(X) g(Y)^T V\right) + \lambda \sum_{i=1}^{n} \left(\|x_i - G_1(f(x_i))\|_2^2 + \|y_i - G_2(g(y_i))\|^2 \right)$$

이다. 여기서 $U^T f(X)$와 $V^T g(Y)$는 투영 스코어이다. [그림 4-25]의 (a)와 (b), (c) 모형간의 차이는 (a)는 임베딩으로 한 입력으로 오토인코더를 수행하고 복원오차는 자기복원오차와 서로 다른 자료의 복원오차항으로 구성되어 있다. 반면 (b)와 (c)는 서로 다른 두 입력 데이터가 들어가는 것에 차이가 있다. (b)와 (c)가 두 변수의 상관관계를 고려하는 것에는 차이가 있지만 (c)는 오토인코더로 복원오차도 동시 고려하는 것에서 차별성을 가진 모형이다.

비정형 자료의 연계적용방법으로 확장하면 여러 모달리티가 있는 경우에도 적용할 수 있다. 예를 들어, 다변량 연속벡터 $x \in R^{p_1}$과 다변량 이진벡터 $y \in \{0,1\}^{p_2}$가 주어진 경우를 고려하여 보자. 이진벡터는 p_2개의 복지서비스에서 수혜여부라고 생각할 수 있다.

$\|u_1^T f(X)\|_2^2 = \|v_1^T g(X)\|_2^2 = 1$로 표준화한 경우, 다음의 문제

$$min_{U,V} - tr\left(U^T f(X) g(Y)^T V\right)$$

에 대한 해는 두 투영벡터의 제곱손실에 관한 다음의 문제

$$min_{U,V} \frac{1}{2} \| U^T f(X) - g(Y)^T V \|_F^2$$

의 해와 동일하다. U, V의 방향에 대한 척도 표준화가 필요하므로 $tr(U^T f(X) f(X)^T U)$와 $tr(V^T g(Y) g(Y)^T V)$를 상수로 놓으면 n개의 자료가 주어진 경우 목적함수는

$$min_{U,V\in C(U,V),W_f,W_g,W_{cG_1},W_{G_2}} -\frac{1}{n}tr\left(U^T f(X)g(Y)^T V\right)$$
$$+\lambda\sum_{i=1}^{n}\left(\alpha l_1(x_i,G_1(f(x_i)))+(1-\alpha)l_2(y_i,G_2(g(y_i)))\right.$$

이 된다. 단, $\alpha\in[0,1]$는 두 항을 조율하는 모수이다. 두 손실 $l_1(x_i,G_1(f(x_i)))$와 $l_2(y_i,l_2(g(y_i)))$는 복원오차이며, $l_1(\cdot,\cdot)$는 제곱손실함수 또는 $l_2(\cdot,\cdot)$는 음이항손실함수(negative binomial loss function)에 해당될 수 있다. 또한, 자료가 얻어지는 f를 수치형자료로부터 얻어지는 임베딩이라 하고, g를 텍스트를 입력으로 하여 순환신경망을 통한 임베딩 벡터라 놓으면, 서비스 수혜자의 정보와 서비스 이용패턴에 대한 자료에 대해서 심층 정준상관분석을 적용할 수 있다.

제5장

결론

제 5 장 결론

2020년 한 해는 국내외 모든 이슈가 COVID-19로 집중되었다. 2장, 3장에서 보건, 복지, 사회보장 키워드로 수집된 문서(2020.1.1.-2020.11.31.)들을 분석해본 결과, 전체 기간 동안 코로나, 확진, 지역, 환자, 바이러스, 지원, 감염, 신종, 검사, 정부, 서울, 방역, 보건, 발생, 마스크, 중국, 사회, 격리, 병원, 확산 키워드 순으로 코로나 관련 이슈가 상위 키워드를 차지하였다. 연구진에서 선정한 주요 키워드인 '코로나'는 2월에 가장 높은 빈도를 보였으며, '마스크' 키워드는 2, 3월에 최대치로 나타났다. '백신' 관련해서는 '접종' 키워드와 더불어 10월에 가장 높은 빈도를 나타냈고, '예방' 키워드는 2월에 가장 높은 빈도를 나타내다가 10월에 다시 증가하였다. '지원금', '지급', '재난' 모두 공통적으로 4월에 높은 편이며, '지원금'은 9월에 한 번 더 증가한 것으로 나타났다. 상반기에 집중된 키워드는 '발병, 체온, 유증, 소독제, 본부장, 감시, 전염병, 증상' 등이 있고, 하반기에 집중된 키워드는 '내년, 대선, 결합, 백신' 등의 키워드이다. 3장 클러스터링 분석 결과 각 월별 클러스터 명칭은 코로나 관련 국내외 정세, 코로나 확진 관련 조치, 코로나 지원 대책, 코로나 환자 확진에서 크게 벗어나지 않았다. 이는 코로나 19 관련 이슈가 각 월별로 차이가 없음을 의미한다.

4장에서는 비정형 형태의 자료의 활용성을 기계학습과 딥러닝의 여러 모형 등을 살펴봄으로써 활용성을 확장하기 위한 방안을 모색해보았다. 최근 딥러닝 방법은 임베딩에 기반하여 다양한 측도를 융합하여 활용하

는 방법으로 발전되고 있다. 이러한 방법들 중 가장 성공적인 사례는 적대적 신경망(generative adversarial network; GAN)을 들 수 있다. GAN은 목적함수에서 상반되는 역할은 하는 두 개의 적대적 함수를 활용하여 원 함수에 대한 추정의 효율성을 높였다. GAN 방법의 핵심은 자료에 정의되는 거리의 정의와 적분상수의 연산에 있으며, 적분상수 연산의 효율성을 높이기 위한 방안이 계속 개발되고 있다. 그러나 고차함수를 명시적으로 적분하거나 수치적분(numerical integration)하지 않은 이상 제한적인 해결방법이라 할 수 있다. 이중 샘플링을 통한 근사 방법은 해결에 대한 가장 현실적인 방안이라 할 수 있는데, 이에 근거한 방법들은 생성된 자료를 부수적으로 활용할 수 있는 장점을 가지고 있다. 적대적 신경망기법을 적용하면 다양한 데이터를 효율적으로 생성할 수 있으며, 또한, 생성된 모의자료 중 일부는 관측된 데이터와 거의 구별할 수 없을 정도의 높은 수준을 가진다. 이러한 모의자료는 다양한 관점을 사전검토하거나 대응시나리오 점검 등 보건 복지 서비스 프로그램 기획 등 공공서비스 전반에 걸쳐 사전, 사후분석에 활용될 수 있을 것으로 기대한다.

보건복지분야 주요 키워드로 살펴본 위 분석결과는 당연한 결과일 수 있으며 코로나 19 상황으로 인해 2020년 각 월별 주요 이슈 변화가 크지 않았다. 그럼에도 불구하고, 소셜 빅데이터 분석은 보건복지 정책 영역에서 국가적·사회적으로 관심이 있는 이슈에 대해 현 상황을 파악하는 데 중요한 경쟁력으로 작용할 수 있으며, 앞으로의 정책 관련 이슈를 도출하고 연구 전략을 세우는 데 근거자료로 활용될 수 있다. 비정형 빅데이터는 비정형데이터 및 정형데이터와 연계될 수 있는 가능성이 있다. 앞서 살펴본 임베딩 방법론에 기반한 연계 분석 기술을 바탕으로 주요 보건복지 정책에 관한 사회적 관심도, 영향력 등을 분석하고 그 변화 과정을 살펴본다면 시의성 높은 보건복지 정책 연구의 기반을 마련할 수 있을 것이다.

참고문헌

박대현, 송동현 (2014년, 2월). 비정형 데이터 활성화의 정치, 경제, 문화적 함의. Internet & Security focus, 4-20.

인코돔 (2016). *스키마 아키텍처(人CoDOM).* Retrieved from http://www.incodom.kr/스키마

케라스 블로그 (2018). 오토인코더. Retrieved from https://blog.keras.io/building-autoencoders-in-keras.html.

Andrew, G., Arora, R., Bilmes, J., and Livescu, K. (2013). Deep canonical correlation analysis. *Proceedings of the 30th International Conference on Machine Learning, 28*(3), 1247-1255.

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research 3*, 1137-1155.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., …, Amodei, D. (2020). Language models are few-shot learners. arXiv: 2005.14165 [cs.CL].

Cho, K., Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv: 1406.1078 [cs.CL].

DATANET (2011). BI 연계된 정형-비정형 데이터 통합 검색으로 패러다임 전환. *Networks Times, 2011*(8), 137-148.

Data on-air (2020a). 수집 데이터의 형태에 따른 분류.

Data on-air (2020b). 정형, 비정형, 반정형 데이터란?

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning

for image recognition, IEEE Conference on Computer Vision and Pattern Recognition, 770-778.

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory, *Neural Computation, 9*, 1735-1780.

Hotelling, H. (1936). Relations between two sets of variates, *Biometrika, 28*, 321-377.

Kim, Y. (2014). Convolutional Neural Network for Sentence Classification. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, 1746-1751.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278-2324.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., and Dean, J. (2013). Distributed Representations of Words and Phrases and their compositionality, *Advances in Neural Information Processing Systems 26*, 3111-3119.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and Its Applications. 9* (1): 141-142.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. (2011). Multimodal deep learning. In ICML, 689-696.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing,* EMNLP 2014, 1532-1543.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: convolutional networks for biomedical image segmentation, ArXiv, abs/1505.04597.

SD아카데미 (2018). 인공지능(A.I)기술과 자연어 처리.

Simonyan, K. and Zisserman, A. (2014). Very Deep convolutional networks for large-scale image recognition. CoRR2014.

Srivastava, N., Mansimov, E., and Salakhudinov, R. (2015). Unsupervised learning of video representations using LSTMs. *Proceedings of Machine Learning Research, 37,* 843-852.

TCPschool (2018). JSON과 XML의 구조. http://tcpschool.com.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need, Neural Information Processing Systems (NIPS 2017)

Wang, W., Arora, R., Livescu, K., and Bilmes, J. (2015). On deep multi-view representation learning. *Proceedings of the 32nd International Conference on International Conference on Machine Learning, 37,* 1083-1092.

Wang, X., Xiangnan, H., Cao, Y., Liu, M., and Chua, T-S. (2019). KGAT: Knowledge Graph attention network for recommendation. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.*

Witten, D., Tibshirani, R., and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis, *Biostatistics, 10*(3), 515-534.

Xia, X. and Kulis, B. (2017). W-Net: a deep model for fully unsupervised image segmentation, ArXiv, abs/1711.08506.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2016). Learning deep features for discriminative localization, 2016 IEEE

Conference on Computer Vision and Pattern Recognition(CVPR), 2921-2929.

Zhuang, X., Yang, Z., and Cordes, D. (2020). A technical review of canonical correlation analysis for neuroscience applications. 41(13), 3807-3833.

[부록 1] 경험위험 관점에서의 표본주성분분석의 이해

이 부록에서는 자료의 크기가 n인 p차원의 관측치 $x_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^T i = 1, \cdots, n$이 주어졌을 때, 차원축소된 공간에서 자료를 근사하는 관점에서 표본주성분분석을 유도하기로 한다. 원변수와 차원축소된 주성분변수간의 정보의 손실을 제곱손실함수를 이용하여 추정하는 방법을 설명한다. 이 방법은 선형 주성분분석을 비선형 주성분분석으로 확장이 가능하다는 장점이 있다. 다음으로 스펙트럴 분해와 특이치 분해의 관계를 살펴보고 이 절을 맺도록 한다.

X를 한 행이 p차원 관측치 x_i으로 구성된 자료행렬로 놓자. $e_j = (e_{j1}, \cdots, e_{jp})^T, j = 1, \cdots, p$을 p차원 유클리드 공간의 표준기저벡터(standard basis vector)라고 하자. 여기서 $e_{jk} = I(j = k)$으로 정의된다. 그러면 임의의 관측치 x_i은 $x_i = \sum_{j=1}^{p} x_{ij} e_j = \sum_{j=1}^{p} e_j^T x_i e_j$으로 표현된다. 평균 μ을 고려할 때, 일반적으로 임의의 p차원 직교 기저벡터 $v_j, j = 1, \cdots, p$(즉, $|v_j| = 1$이고 $j \neq k$에 대해서 $v_j^T v_k = 0$)에 대하여 $x_i = \mu + \sum_{j=1}^{p} v_j^T (x_i - \mu) v_j$으로 표현된다. 이 경우, 차원축소는 p보다 작은 q개의 직교 기저벡터를 사용하여 관측치 x_i을 근사적으로 표현하는 것으로 볼 수 있다. 즉, $x_i = \mu + \sum_{j=1}^{p} v_j^T (x_i - \mu) v_j$을 $v_1, \cdots, v_q$공간에서의 근사해인 $z_i = \mu + \sum_{j=1}^{q} (v_j^T (x_i - \mu)) v_j$로 가장 잘 근사하는 직교 기저벡터

$v_{1,}\cdots,v_q$을 구하면 된다. 자료의 원분산이 가능한한 유지되기 위해서는 2차 적률(moment)인 다음의 제곱손실함수

$|x_i-z_i|^2=|\sum_{j=1}^{p}v_j^T(x_i-\mu)v_j-\sum_{j=1}^{q}v_j^T(x_i-\mu)v_j|^2$를 최소화하는 해를 찾으면 된다. n개의 자료점 모두를 고려하면, 표본주성분분석은 모든 $q\le p$에 대해서 다음의 제약조건

$z_i=\mu+\sum_{j=1}^{q}(v_j^T(x_i-\mu))v_j,\ v_j^Tv_k=\begin{cases}1\ j\neq k,\\0\ j=k\end{cases}$하에서

$\min_{\mu,v_1,\cdots,v_q}\frac{1}{n}\sum_{i=1}^{n}|x_i-z_i|^2$이 되는 기저벡터 $v_{1,}\cdots,v_p$을 찾는 문제로 볼 수 있다. 최적화의 제약조건은 새로 생성된 방향벡터 v_j들이 서로 직교하며 단위벡터임을 의미한다. $V_q=[v_{1,}\cdots,v_{q]}$`--으로 정의하면 ({dim_pca_model2})는 $z_i=\mu+V_qV_q^T(x_i-\mu)$으로 표현된다. 따라서 모든 자료 $x_i,i=1,\cdots,n$에 대한 차원 축소로 인한 평균제곱손실은 $R=\frac{1}{n}\sum_{i=1}^{n}|(I_q-V_qV_q^T)(x_i-\mu)|^2$이며, 복원오차(reconstruction error)라 일컫는다. V_q이 주어지면 μ의 추정치는 표본평균벡터 $\bar{x}=\sum_{i=1}^{n}x_i/n$임을 알 수 있다. μ를 추정치 $\bar{x}$으로 대체하면 문제는

$$min_{V_q\in R^{p\times q}}\frac{1}{n}\sum_{j=1}^{n}|(x_i-\bar{x})-V_qV_q^T(x_i-\bar{x})|^2 \qquad \text{(A-1)}$$

이 된다. 논의의 편의상 식 (A-1)에서 $x_i-\bar{x}$ 대신 x_i으로 표기하기로 한다. 앞으로 표기상의 오해가 없으면 x_i은 중심화된 $\tilde{x_i}$으로 간주한다. 또한 $V=[v_1,\cdots,v_p]$인 경우를 생각하자. 그러면 $\Pi=V(V^TV)^{-1}V^T=VV^T$은

정사영 행렬(projection matrix)으로써 각 관측값 x_i 를 차원축소된 근사 관측값 $z_i = \Pi x_i$으로 매핑하는 역할을 수행한다. 행렬 VV^T과 $I_p - VV^T$은 대칭이며 멱등(idempotent)이므로 복원오차 (A-2)는

$$\begin{aligned} R &= \frac{1}{n}\sum_{i=1}^{n} x_i^T (I_p - VV_i^{T)x} \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i^T x_i - \frac{1}{n}\sum_{i=1}^{n} x_i^T VV^T x_i \\ &= \frac{1}{n} tr\left(\sum_{i=1}^{n} x_i x_i^T\right) - tr\left(V^T \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T V\right) \end{aligned} \qquad \text{(A-2)}$$

으로 표현된다. 즉, $R = tr(X^T X) - tr(V^T X^T XV)$이다. 따라서 식(식 pca)은 $tr(V^T X^T XV)$을 최대화 하는 문제와 동치이다. 제 1주성분 방향벡터 v_1의 분산은

$$max_{v_1 \in R^p} \sum_{i=1}^{n} (x_i^T v_1)^2 = \max_{v_1 \in R^p} v_1^T (X^T X) v_1 = \lambda_1^2$$

임을 알 수가 있다. 여기서 λ_1^2은 $X^T X$의 최대 고유값 또는 X의 최대 특이값(singular value) λ_1의 제곱에 해당된다. 특히 p이 큰 경우에는 모공분산행렬 Σ 또는 표본공분산행렬 S의 크기가 $p \times p$으로 커지므로 저장공간을 적게 쓰면서 해를 찾는 방법이 요구된다. 이 목적에 부합하는 방법이 특이치 분해(singular value decomposition)라 할 수 있다. $V = [v_1, \cdots, v_p]$을 구하는 방법으로는 정리의 스펙트럴 분해 방법을 직접 적용하기보다는 다음의 자료행렬 X에 대한 특이치 분해에 의해 구한다.

$$X = U\Lambda V^T, \qquad \text{(A-3)}$$

여기서 $r = \min(n,p)$은 자료행렬 X의 계수(rank)이고 $n \times r$행렬 $U = [u_1, \cdots, u_r]$은 직교행렬 $U^T U = I_r$을 만족한다. 열벡터 u_j들은 좌 특이치 벡터(left singular vector)라 부른다. 또한 $p \times r$인 행렬 V도 직교성 $V^T V = I_r$을 만족하며 v_j들을 우 특이치 벡터(right singular vector)라 일컫는다. Λ은 특이값 λ_j들로 이루어진 대각행렬로, 특이값들은 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r \geq 0$을 만족한다. 특이치 분해 (A-3)의 관계식으로부터 $XV = U\Lambda$이 성립하고, 우변의 $n \times p$ 행렬 $U\Lambda$을 X의 주성분점수(principal component score)라고 부른다. 만약 q개의 주성분만을 고려한다면 $XV_q = U_q \Lambda_q$으로 표현된다. 여기서 V_q, U_q, Λ_q은 각각 대응되는 행렬의 앞 q열을 추출한 부분행렬을 의미한다.

참고로 특이치 분해와 스펙트럴 분해의 관계를 확인할 수 있다. 행렬계수가 r일 때, $(p-r)$개의 고유값이 0이므로 표본공분산행렬 $S = X^T X/(n-1)$이 둘 때, 스펙트럴 분해에 의해 다음의 식이 성립한다. $(n-1)S = X^T X = \lambda_1^2 v_1 v_1^T + \cdots + \lambda_r^2 v_r v_r^T$. $p < n$일 때는 $r = p$으로 0인 고유값은 관찰되지 않는다. 만약 $p > n$인 경우에는 $r = n$으로 $p - n$만큼의 0인 고유값이 존재하며, 이들에 대한 고유벡터는 항상 $Xv_{k=0}, k = n+1, \cdots, p$을 만족한다. V을 k번째 열벡터로 v_k을 가지는 행렬로, Λ은 대각원소로 λ_k인 $r \times r$ 크기의 대각행렬로 정의하자. U을 k번째 열벡터를 $u_k = \lambda_k^{-1} X v_k, k = 1, \cdots, r$ 로 놓으면

$$
\begin{aligned}
U\Lambda V^T &= U\begin{bmatrix} \lambda_1 v_1^T \\ \lambda_2 v_2^T \\ \vdots \\ \lambda_r v_r^T \end{bmatrix} \\
&= \sum_{k=1}^{r} (\lambda_k^{-1} X v_{k)} \lambda_k v_k^T \\
&= \sum_{k=1}^{r} X v_k v_k^T \\
&= \sum_{k=1}^{p} X v_k v_k^T = X \sum_{k=1}^{p} v_k v_k^T = X
\end{aligned}
$$

으로 귀결된다. 마지막 두 등식은 $v_{k,}k = r+1, r+2, \cdots, p$은 0의 고유값에 해당하는 고유벡터이므로 $Xv_{k=0}$로부터, v_k들의 직교성(orthogonality)으로부터 각각 성립한다.