COVID-19 감염경로 분석을 위한 역학조사서 전처리 방법론에 관한 연구

김시원, 김지온*

한림대학교

siwonkim@hallym.ac.kr, *jion972@hallym.ac.kr

A Study on Preprocessing Methodology for Epidemiological Investigation Data to Analyze COVID-19 Transmission Routes

Siwon Kim, Jion Kim*

Hallym Univ.

요 약

본 논문은 반정형 형태로 기록된 역학조사서를 전처리하여 확진자-접촉자 연결망을 구축하고, 이를 네트워크 분석에 적용한 과정을 제시한다. 역학조사서는 일정한 필드 구조를 갖추고 있으나, 실제 자료는 작성자와 시기마다 기록 방식이 달라 데이터의 불일치 및 결측이 빈번하게 발생한다. 이를 해결하기 위해 개인 식별자 부여, 역할 전환 처리, 필드 표준화 등의 전처리 과정을 수행하였다. 전처리 결과를 기반으로 관계망을 시각화하고, 중심성 지표를 산출하여 주요 전파자와 매개자를 규명할 수 있었다. 이러한 반정형 데이터 전처리 방법론은 전염병 기반의 역학 연구와 더불어 관계정보를 다루는 다양한 연구 분야에도 적용될 수 있다.

Ⅰ. 서 론

사회 연결망 분석(Social Network Analysis) 원리는 개인 간 상호작용을 구조적으로 이해하고, 사건이나 현상의 확산 경로를 규명하는 데 중요한 방법론이다 [1]. 특히, 감염병 역학 분야에서는 확진자와 접촉자 간의 관계를 네트워크로 전환함으로써 감염경로, 전파자의 특성, 슈퍼 전파자 존재 여부, 전파 경로의 특성을 밝히는 데 기여해왔다.

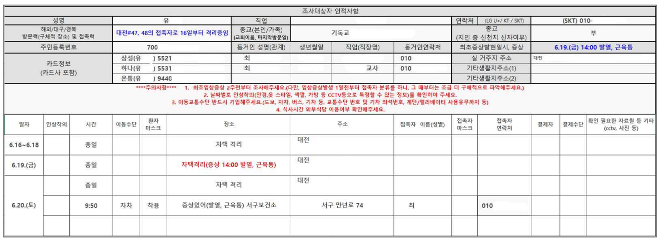
그러나 이러한 감염경로 추적을 위한 분석의 성패는 원본 자료의 품질과 정형화 수준에 크게 의존한다. 대부분의 역학조사서는 일정한 필드 구조를 갖춘 반정형(semi-structured) 데이터로 수집되지만, 실제 현장에서는 작성자, 시기, 기관마다 기록 방식이 상이하여 데이터의 일관성이 부족한 경우가 많다 [1]. 역학조사서는 성명, 성별, 연령 등의 정형 필드와 증상 발현일, 감염 경위와 같은 자유 기입식 서술형 텍스트가 혼재된 형태로, 동일한 항목이 문서마다 다르게 표현되거나 일부 필드가 공란으로 남겨지는 사례가 빈번하다. 반정형 데이터의 이러한 비표준적인 특성은 특히 의료 및 역학 분야에서 데이터 전처리 과정에 소모되는 시간을 늘리고 복잡하게 만들며 연결망 분석의 정확성을 저해하는 주요 요인이 된다 [2]. 따라서 원본 데이터를 정제하고 관계정보의 추출이 가능한 형태로 재구성하기 위한 전처리 과정이 필수적이다. 그러나 기존 연구들이 대부분 정형화된 자료에 기반을 두고 있어, 실제 현장의 반정형 데이터를 분석 가능한 형태로 전환하는 과정에 대한 논의는 상대적으로 미흡한 실정이다 [3].

이에 본 연구는 COVID-19 역학조사서를 사례로 반정형 형태의 역학조사서를 관계분석이 가능한 형태로 전처리하여 연결망 분석 가능한 데이터셋으로 정제하는 과정을 제시하고자 한다. 이러한 전처리 방법은 전처리 역학 분야를 넘어, 사회 조사, 범죄 수사 기록 등 비정형 또는 반정형 데이터가 빈번히 발생하는 다양한 분야에 확장 적용될 수 있는 잠재력을 탐색하고자 한다.

Ⅱ. 연구 방법

본 연구는 역학조사서 원본 데이터를 연결망 분석에 활용하기 위해 원본 자료→ 1차 전처리 → 2차 전처리의 세 단계 과정을 거쳤다.

1) 원본 데이터 (심층 역학조사서)


조사대상자 인적사항

성명	유	직업				연락처 (LG U+/ KT / SKT)	(SKT) 010-
해외/대구/경북 방문력(구체적 장소) 및 접촉력	대전#47, 48의 접촉자로 16일부터 격리중임	종교(본인/가족) (교회이름, 마지막방문일)	기독교			종교 (지인 중 신천지 신자여부)	부
주민등록번호	700	동거인 성명(관계)	생년월일	직업(직장명)	동거인연락처	최초증상발현일시, 증상	6.19.(금) 14:00 발열, 근육통
카드정보 (카드사 포함)	삼성(유) 5521	최			010	실 거주지 주소	대전
	하나(유) 5531	최		교사	010	기타생활지주소(1)	
	온통(유) 9440					기타생활지주소(2)	

****주의사항**** 1. 최초임상증상 2주전부터 조사해주세요.(다만, 임상증상발생 1일전부터 접촉자 분류를 하니, 그 때부터는 조금 더 구체적으로 파악해주세요.)
2. 날짜별로 인상착의(안경,옷 스타일, 색깔, 가방 등 CCTV등으로 특정할 수 있는 정보)를 확인하여 주세요.
3. 이동교통수단 반드시 기입해주세요.(도보, 자차, 버스, 기차 등, 교통수단 번호 및 기차 좌석번호, 계단/엘리베이터 사용유무까지 등)
4. 식사시간 외부식당 이용여부 확인해주세요.

일자	인상착의	시간	이동수단	환자 마스크	장소	주소	접촉자 이름(성별)	접촉자 마스크	접촉자 연락처	결제자	결제수단	확인 필요한 자료원 등 기타 (cctv, 사진 등)
6.16~6.18		종일			자택 격리	대전						
6.19.(금)		종일			자택격리(증상 14:00 발열, 근육통)	대전						
6.20.(토)		종일			자택 격리	대전						
		9:50	자차	착용	증상있어(발열, 근육통) 서구보건소	서구 만년로 74	최		010			

〈그림 1〉 비식별화 완료된 심층 역학조사서 원본 데이터

원본 데이터는 정형 필드와 함께 자유 서술형 텍스트가 혼재된 반정형 형태를 가지고 있으며, 이러한 특성으로 인해 규칙 기반 자동 파싱만으로는 일관된 값을 안정적으로 추출하기에 어려움이 존재한다. 주요 원인으로는 ① 형식의 비일관성 ② 날짜·장소·시간 등의 값 표현의 자유도와 다수의 이형(異形) 존재 ③ 동일 인물의 중복 기재나 확진자·접촉자의 역할 반복 등 사람의 의미적 판독 필요 ④ 정확도를 위한 분석 규칙 적용 등이 있다. 따라서 1차 단계에서는 사람이 직접 항목을 판독하고 핵심 필드를 정리하는 수작업 전처리가 불가피했으며, 이후 2차 단계에서 코드화된 규칙을 적용해 관계정보를 구조화 및 검증하는 이중 처리 전략을 채택했다.

2) 1차 전처리

구분	컬럼명	설명
확진자 정보	성명	확진자 이름
	전화번호	확진자 연락처
	생년	출생년도
	성별	남/여
	주소	거주지 주소
	증상 발현일	최초 증상 발현 시점
	확진일	확진 판정일
동선 및 접촉자 정보	일자	해당 동선 발생일
	시간	방문/노출 시간
	장소	방문 장소명
	업종	장소의 업종 분류
	주소	방문 장소 주소
	확진자 마스크 착용 여부	O/X
	접촉자 이름	-
	접촉자 전화번호	-
	접촉자 성별	-
	접촉자 생년	-
	접촉자 주소	-
	접촉자 마스크 착용 여부	-

〈표 1〉 1차 전처리 후 데이터 컬럼 구성

원본 역학조사서에서 분석과 무관한 법령, 작성 지침, 중복 서술을 제거한 뒤, 분석에 필요한 항목만을 추출하여 총 19개 컬럼을 가진 표 형태로 정리하였다. 성명, 성별, 생년, 주소, 증상 발현일, 확진일 등 기본 필드를 선별하고, 문서마다 달리 기재된 날짜와 장소 표현은 단순화된 규칙에 따라 통일된 형식으로 정리하였다. 예를 들어, '6월 24일 16시경'과 같은 서술식 기록은 '2020-06-24 16:00'과 같이 표준화하였다. 이를 통해 불규칙한 반정형 자료를 연결망 분석이 가능한 구조로 가공하였다.

3) 2차 전처리

컬럼명	설명	예시
person_id	인물 고유 식별자	P000001
person_name	인물 이름(익명 처리)	이OO
role	확진자/접촉자 구분	confirmed
sex	성별	F
birth_year	출생년도	1997
onset_dt	증상 발현일	2020-02-17
diagnosis_dt	확진일	2020-02-21
district_code	거주 지역 코드	25
conversion_flag	접촉자→확진자 전환 여부	0

〈표 2〉 최종 산출 데이터 구조 persons 시트 예시

1차 전처리로 정리된 데이터를 기반으로, 코드화된 규칙을 적용하여 관계정보를 추출하기 위한 데이터 형태를 구조화하였다. 동일 인물이 문서 내에서 확진자와 접촉자로 반복 등장하는 경우에는 하나의 엔티티로 통합하였고, 접촉 관계는 확진자 중심으로 일관되게 연결되도록 변환하였다. 또한, 데이터 누락과 중복 문제를 최소화하기 위해 표현이 다른 값은 동일한 표준 형식으로 변환하며 이를 통해 연결망 구축 과정에서 발생할 수 있는 불필요한 노드 분산이나 잘못된 연결을 방지하였다.

최종적으로 산출된 데이터 시트는 persons, locations, edges의 세 개의 시트로 나뉘며 확진자와 접촉자를 구분하는 person_id, role 컬럼을 중심으로 접촉이 발생한 날짜(date), 장소(locations), 접촉자 식별자(contact_id) 및 변환 과정에서 부여된 변수값(conversion_flag, mask_flag) 등으로 구성된다. 이를 통해 각 행은 '확진자-접촉자 쌍'을 의미하며, 접촉 시점과 장소 정보가 결합되면서 네트워크의 엣지(edge) 정보로 변환될 수 있도록 구조화했다.

Ⅲ. 분석 및 결과

대전지역의 심층 역학조사서 200건을 데이터셋으로 사용하여 확진자-접촉자 네트워크 구축 및 분석을 진행하였다. 전처리를 거쳐 정제된 데이터셋은 총 200명의 확진자와 1,851명의 접촉자로 구성되었으며, 이들 간의 관계는 2,231건의 엣지(edge)로 표현된다.

1. 네트워크 기본 통계

확진자-접촉자 네트워크는 노드 수, 엣지 수, 평균 차수, 네트워크 밀도, 약한 연결 집합 수, 최대 연결 집합 크기, 노드 간 최대·최소 거리 등을 통해 구조적 특성을 파악하였다.

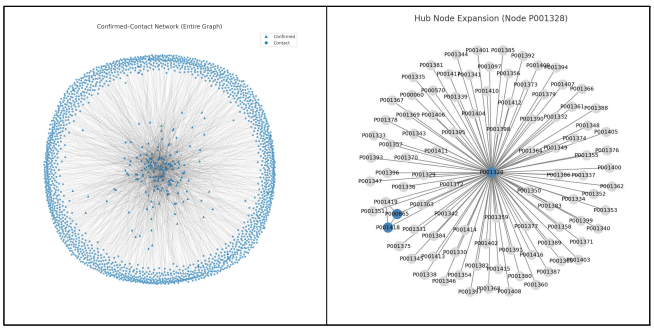
지표	값
노드 수	2,051
엣지 수	2,231
평균 차수(average degree)	2.12
네트워크 밀도	0.001
약한 연결 집합 수	44
최대 연결 집합 크기	1,698
노드 간 최대 거리	16
노드 간 최소 거리	1
노드 간 평균 거리	7.3

〈표 3〉 확진자-접촉자 네트워크의 기본 통계

평균 차수가 약 2라는 것은 대부분의 개인이 1~2명 수준의 제한적인 접촉 관계만 맺고 있다는 것을 의미한다. 그러나 일부 노드는 다수와 연결되어 있어, 전파가 특정 집단이나 개인에 의해 집중적으로 이루어졌음을 보여준다. 네트워크 밀도가 0.001로 낮게 나타난 것은 모든 사람이 서로 연결된 촘촘한 구조가 아니라, 허브 노드를 중심으로 방사형으로 뻗어 나가면서 주변 이웃 노드 간의 연결은 거의 없는 느슨한 네트워크라는 것을 의미한다. 이러한 구조적 특성은 수많은 개별 노드들이 산발적으로 분리되어 있으면서도, 동시에 슈퍼 전파자와 같은 소수 핵심 노드가 전체 확산을 좌우할 수 있는 위치에 있다는 것을 의미한다. 실제로 44개의 연결 집합은 지역 내 다수의 소규모 전파 집단이 존재함을 반영하며, 그중 최대 연결 집합이 전체의 약 82.8%를 포함하는 대규모 구조로 나타난 것은 특정 사건이나 집단이 지역 내에서 지배적인 영향력을 행사했음을 보여준다.

더불어 네트워크의 최소 거리가 1, 최대거리가 16으로 나타나는 것은 한쪽 끝의 개인에서 다른 쪽 끝의 개인까지 최장 16단계를 거쳐 연결될 수 있다는 것을 의미한다. 이는 전파 대부분이 짧은 거리에서 발생하지만, 동시에 일부 경로는 장기간 연쇄적으로 이어져 감염병이 확산될 수 있다는 구조적 가능성을 내포한다.

2. 확진자-접촉자 네트워크 구조



〈그림 1〉 확진자-접촉자 네트워크 전체 구조(왼)와 주요 허브 노드 확대(오)

앞서 확인된 바와 같이, 본 연구를 통해 도출한 대전지역 COVID-19 감염경로 네트워크는 낮은 밀도와 중심부에 다수의 소규모 연결 집합을 특징으로 한다. 이러한 구조적 특성은 시각화를 통해 명확하게 드러난다. 네트워크 주변부에 중심부와 연결된 수많은 개별 노드들이 분포되어 있으면서, 중심부에 소규모 허브 노드들의 하위집단이 감염네트워크의 코어를 구성하고 있다. 소규모 집단들은 주로 가족이나 지인 모임과 같은 밀접 생활 단위, 직장·학교·종교시설 등 특정 공간을 공유하는 단위, 혹은 특정 시점에 발생한 사건 단위로 형성되었다.

그림 1의 왼쪽 그래프는 전체 확진자-접촉자 네트워크를 나타낸다. 네트워크는 중심부에 연결 중심성이 높은 몇 명의 확진자를 중심으로 뭉쳐 있으며, 주변부에는 중심부와 연결된 다수의 개별 노드들이 분산되어 존재하는 형태를 보인다. 이는 대다수 개인이 제한적인 접촉 관계만 맺고 있었으나, 특정 확진자를 통해 네트워크가 전반적 연결이 일어났다는 것을 보여준다.

그림 1의 오른쪽 그래프는 대표적인 허브 노드인 P001328의 연결구조를 확대한 것이다. 해당 확진자는 95명의 접촉자와 직접 연결되어 있어 네트워크 내에서 스타형 구조를 형성하며, 가장 높은 연결 중심성을 보였다. 이는 감염병 전파에서 소수의 슈퍼 전파자 혹은 특정 사건이 전체 연결망 구조를 결정하는 데 큰 영향을 끼친다는 것을 의미한다.

Ⅳ. 결론 및 시사점

본 논문에서는 대전지역의 심층 역학조사서 200건을 기반으로 확진자-접촉자 네트워크를 구축하고, 이를 통해 감염병 전파 구조를 분석하였다.

첫째로, 반정형 형태의 역학조사서에 포함된 관계정보를 시각화 하여 감염경로 연결망을 분석하기 위해 2차에 걸친 전처리 방법론을 제시하였다.

둘째로 관계형 데이터로 전처리한 역학조사 데이터셋을 네트워크 형태로 시각화하여 분석한 결과, 대전지역 COVID-19 확진자 연결망은 주변부에 서로 연결되어 있지 않은 다수의 개별 노드들로 분산되어 있으면서도, 특정 사건이나 슈퍼 전파자에 의해 중심부에 소규모 집단이 형성되는 구조를 보였다. 이는 감염병 전파가 소수 핵심 노드에 의해 급격히 확산될 수 있음을 시사한다.

본 연구는 역학조사서와 같은 비정형·반정형 데이터를 네트워크 데이터로 전환할 수 있음을 입증하였으며, 이를 통해 전파 구조의 시각화, 주요 확산 경로 식별 등 감염경로 기반의 다양한 분석 가능성을 제시하였다. 나아가 제안된 전처리 방법론은 감염병 역학조사뿐만 아니라 반정형 형태의 보고서를 활용하는 다양한 분야에 적용될 수 있을 것으로 기대한다. 다만 향후 후속연구를 통해 전처리와 연결망 데이터 구축 과정을 LLM 등 AI 기술을 활용하여 최대한 자동화할 수 있어야만 이와 같은 분석 방법론을 효과적으로 실무에 적용할 수 있을 것이다.

ACKNOWLEDGMENT

이 논문은 2025년도 정부(경찰청)의 재원으로 과학치안진흥센터의 지원을 받아 수행된 연구임 (과제번호: RS-2025-02218280, 사이버범죄 수사단서 통합분석 및 추론 시스템 개발 사업).

참 고 문 헌

[1] Wasserman, Stanley, and Katherine Faust. 1994. Social Network Analysis: Methods and Applications. Cambridge University Press.

[2] 이승희, 이지원, 김연주, 김상희 and 조수현. "역학조사관의 감염병 대응 경험: 질병관리청 소속 역학조사관을 대상으로 한 포커스 그룹 인터뷰" 보건행정학회지 34, no.4 (2024) : 440-449.doi: 10.4332/KJHPA.2024.34.4.440

[3] Amiruddin, N. A., Wahid, N. A. A., & Yasin, M. M. 2024. "Pre-processing Approach for Semi-Structured Medical Data." In Proceedings of the International Conference on Innovation & Entrepreneurship in Computing, Engineering & Science Education (InvENT 2024). Atlantis Press. https://www.atlantis-press.com/proceedings/invent-24/126005585.

[4] Jo, W., Chang, D., You, M. et al. A social network analysis of the spread of COVID-19 in South Korea and policy implications. Sci Rep 11, 8581 (2021). https://doi.org/10.1038/s41598-021-87837-0