

# 이종 네트워크 연계 기반 HGT 모델을 활용한 범죄 사건 내 피의자 후보군 판별 기술 연구

한현규, 백명선  
세종대학교

hhg02258565@gmail.com, msbaek@sejong.ac.kr

## A Study on Criminal Suspects Identification Technology using a Heterogeneous Network-based HGT Model

Hyungyu Han, Myung-Sun Baek  
Sejong University

### 요 약

초동 수사 단계에서 신속하고 효과적으로 범죄에 대응하기 위해 수사관에게 잠재적 용의자 후보군 정보를 제공하는 것은 매우 중요하다. 기존 연구는 텍스트 분류나 개별 네트워크 분석에 초점을 맞추었으나 인물과 사건 간의 복합적인 관계를 통합적으로 모델링하는 데 한계가 있었다. 본 논문에서는 인물-사건 관계를 단일 이종 그래프(Heterogeneous Graph)로 모델링하고 Heterogeneous Graph Transformer(HGT)를 이용해 사건별 참여자 집합 내에서 피의자 여부를 분류하는 인공지능 기술을 설계한다. 제안 기술은 텍스트로 기술된 인물 및 사건 정보를 임베딩하고 이들 간의 유사도 및 소속 관계를 그래프 엣지로 구성한다. 본 기술은 초동 수사 시 제한된 자원을 효율적으로 배분하고 용의자 탐문 범위를 좁히는 데 주요한 참고 자료로 활용될 수 있다.

### I. 서 론

최근 스마트 치안 기술의 일환으로 인공지능 및 빅데이터를 활용하여 범죄에 대응하는 연구가 활발히 진행되고 있다 [1]. 효과적인 초동 수사를 위해서는 신규 사건 접수 시 해당 사건의 특성을 분석하고 관련성이 높은 인물을 신속하게 파악하는 것이 중요하다. 기존에는 텍스트 분류를 통해 범죄 유형을 예측하거나 준지도학습 기반의 네트워크를 통해 용의자를 추론하는 연구가 제안되었다[2]. 그러나 이러한 접근법들은 인물 간의 관계와 사건 간의 관계를 분리된 네트워크로 다루거나 단순 텍스트 유사도에 의존하여 다각적인 관계 정보를 통합하는 데 어려움이 있었다.

본 논문에서는 사건과 인물정보를 포함하는 누적된 가상의 사건 데이터 내 인물과 사건 정보를 하나의 이종 그래프에 통합 모델링하는 방법론을 제안한다. 본 논문에서는 사법시스템에서 가장 일반적으로 활용되는 가상의 형사사법정보시스템 (KICS: Korea Information System of Criminal Justice Services) 데이터를 활용하였다.

제안 방법은 다음과 같은 핵심 기술로 구성된다. 첫째, 최신 언어 모델을 사용해 텍스트 정보를 벡터로 임베딩한다. 둘째, 인물-인물(p-p) 및 사건-사건(c-c) 간에는 상호 k-최근접 이웃(mutual k-NN) 알고리즘으로 강건한 유사도 엣지를 생성하고 인물-사건(p-c) 간에는 원본 데이터의 소속 관계를 엣지로 연결한다. 셋째, 이종 그래프 구조 학습에 특화된 Heterogeneous Graph Transformer(HGT) [3] 모델을 적용하여 특정 사건과 관련된 인물이 '피의자' 관계로 연결될 확률을 예측한다.

상기 설계된 기술을 활용하면 신규 사건 발생 시 참여 인물들의 역할을 예측하여 수사 우선순위를 정하는 데 참고할 수 있다.

### II. 범죄유형 추론 기술 설계

입력 데이터는 KICS 양식의 사건 개요와 인물 관련 레코드로 구성되며 단일 엑셀 파일 내 KICS 입력, 피의자, 피해자, 참고인 시트를 사용한다. 인물 식별은 주민번호로 사건 식별은 최초접수번호로 정의한다. 각 시트는 스키마를 유지한 JSON 배열로 변환되어 후속 처리 단계의 일관성을 보장한다. 인물 1 인과 사건 1 건당 하나의 임베딩을 생성하며 임베딩 모델은 Qwen Embedding 4B 를 사용한다[4]. 임베딩 모델을 통해 생성된 2560 차원 벡터는 저장 후 모델 학습 시 노드 타입별 선형 투영층을 거쳐 128 차원으로 투영되며 임베딩 모델 파라미터는 동결하여 GNN 과 분류기만 학습한다. 이러한 전처리와 임베딩 정책은 대규모 배치 처리에서 메모리 사용을 예측 가능하게 만들고 운영 단계의 재현성을 높인다. 본 시스템의 파이프라인은 전처리 → 임베딩 → 그래프 구성 → HGT 인코딩 → 엣지 분류기의 순서로 진행되며 그림 1 은 이 과정을 요약한다.

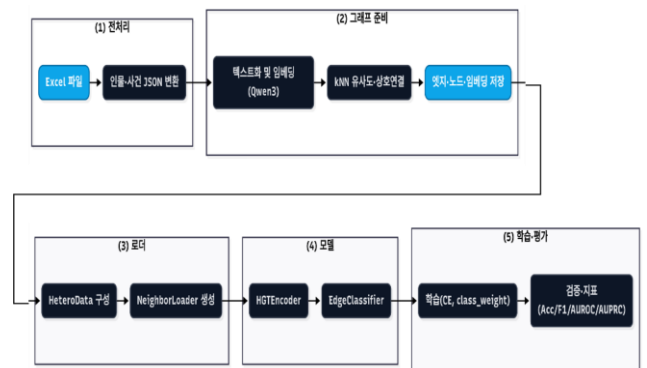


그림 1. 제안된 기술의 전체 절차를 포함하는 블록다이어그램

### III. 그래프 기반 피의자 후보군 예측 방법

그래프  $G=(V,E)$  는 Person 과 Case 두 노드 타입으로 구성되며 인물-인물(p-p)과 사건-사건(c-c) 유사도 엣지, 인물-사건(p-c) 소속 엣지 세 관계를 포함한다. p-p 와 c-c 는 각 타입의 임베딩 코사인 유사도에 기반한 k-최근접 이웃을 계산한 뒤 서로의 이웃 목록에 동시에 포함될 때만 엣지를 부여하는 상호 k-NN 전략을 사용한다. 이때  $k_{pp} = k_{cc} = 15$ , 임계값  $\tau = 0.4$  를 적용하고 엣지 가중치는  $\frac{1+\cos}{2}$  로 정규화하여 스케일 안정성을 확보한다.

p-c 엣지는 인물의 사건 소속 정보를 그대로 연결하며 역할 속성은 [피의자, 피해자, 참고인] 원-핫으로 인코딩한다. 예측 대상 사건의 입력에서는 피의자·참고인 속성을 0 으로 마스킹하고 피해자 정보만 유지하여 모델이 타깃 사건에서 알려진 정답을 직접 보지 못하게 설계한다. 이러한 마스킹은 훈련·검증·추론의 전 단계에 동일하게 적용되며 피해자 존재는 사건 맥락 신호로 작용한다. HGT 는 타입과 관계에 따라 서로 다른 투영과 어텐션을 적용하는 이중 그래프용 트랜스포머로 본 연구에서는 두 개의 HGT 레이어(hidden=128, heads=4)를 사용하여 노드 임베딩을 갱신한다. 입력 단계에서 타입별 선형 투영(2560→128) 후 Swish 활성화와 RMSNorm 으로 정규화하여 안정적으로 메시지 패싱을 시작한다. 갱신된 인물 임베딩  $h_p$  와 사건 임베딩  $h_c$  는 Swish-GLU 두 블록과 Dropout 을 통과한 뒤 2-클래스 로짓을 출력한다. 이 구성은 임베딩의 분포 차이와 역할 정보의 누수 문제를 동시에 고려하여 설계되었으며, 사건·인물 규모가 수만 단위를 넘어서는 환경에서도 NeighborLoader 기반의 타입 인식 샘플링으로 미니배치 학습이 가능하다.

최종 확률은 소프트맥스로 계산하며 사건별로 임계 0.5 초과 인물을 후보로 제시한다. 그림 2 는 인물-사건 이중 그래프의 개념도를 나타낸다.

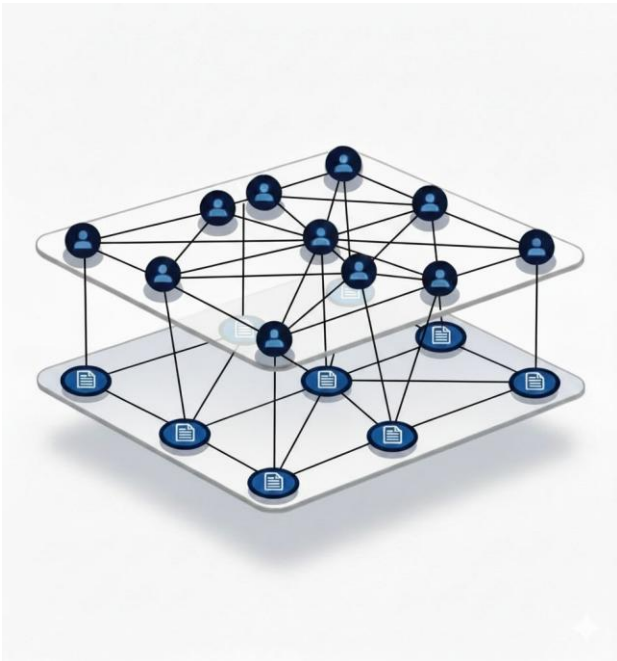


그림 2. 인물-사건 이중 그래프 개념도

### IV. 개발 기술의 성능 검증 방법

제안된 모델의 학습 목적은 특정 사건  $c$  에 이미 참여가 확인된 인물  $p$  의 역할이 '피의자'인지 여부를 분류하는 것이다. 따라서 출력 확률에 대해 가중 크로스엔트로피 손실을 최소화하도록 학습된다. 클래스 가중치는 라벨 분포의 역비율로 설정하여 불균형을 완화하며 손실 집계는 타깃 사건의 p-c 엣지 중 피의자·참고인만 포함하고 피해자 엣지는 자동 제외한다. 데이터 분할의 경우, 사건 노드를 8:1:1 분할하며 전체 그래프를 사용하되 손실 평가는 해당 분할 사건 집합에만 한정한다. 최적화는 AdamW 와 Dropout 을 사용한다. 추론 단계에서는 신규 사건 임베딩을 생성하여 그래프에 연결한 후 사건 참여자 집합 내 피의자 여부 분류 및 확률과 근거 이웃(유사 사건 링크, 관련 인물)을 함께 제시해 수사관이 맥락을 확인할 수 있게 한다. 주요 지표는 F1-score, Accuracy, Precision, Recall 을 중심으로 AUROC, AUPRC 를 함께 기록한다. 본 논문은 피의자 후보군 판별 기술 구현 방법론에 대한 것으로 추후 수치를 포함한 기술 검증 연구가 추가적으로 진행될 예정이다.

### V. 결론

본 논문은 인물-사건 이중 그래프와 HGT 를 결합해 사건별 피의자 후보 확률을 산출하는 방법을 제시하였다. 임베딩 유사도 기반 상호 k-NN 으로 유사 구조를 부여하고 역할 마스킹으로 정보 누수를 차단하며 HGT-기반 메시지 패싱과 GLU 분류기로 엣지 수준 확률을 출력하는 파이프라인을 정립했다. 향후에는 실제 데이터 분포에 대한 적응과 임계 재보정 절차를 포함한 운영형 학습 파이프라인을 확장하고 성능을 검증할 계획이다.

### ACKNOWLEDGMENT

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 정보통신방송혁신인재양성 (메타버스융합대학원) 사업 연구 결과로 수행되었음(IITP-2024-RS-2023-00254529).

### 참고 문헌

- [1] 장광호, “스마트치안”, 2020. 06.
- [2] 백명선 외, “효과적인 초동 수사를 위한 지도/준지도 학습 기반 범죄 사건 정보 예측 및 용의자 후보군 특징,” 한국인공지능학술대회, 2020.
- [3] Z. Hu, Y. Dong, K. Wang, and Y. Sun, “Heterogeneous Graph Transformer,” in *Proceedings of The Web Conference 2020 (WWW '20)*, 2020.
- [4] Y. Zhang et al., “Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models,” arXiv, 2025.