수사 데이터 중복 식별을 위한 알고리즘 기반 처리 프레임워크 연구

문병훈

성균관대학교 과학수사학과

sindori1@naver.com

An Algorithmic Framework for Identifying Duplicates in Investigative Data

Moon Byoung Hun

Sungkyunkwan Univ

요 약

수사기관에서 수집한 통신·계좌 데이터를 병합하는 과정에서 발생하는 중복 문제를 분석하고, 이를 해결하기 위한 알고리즘 기반 처리 프레임워크를 제안한다. 통신 데이터에는 착발신 번호와 시작일시를 활용한 복합키 알고리즘을 적용하고, 계좌 데이터에는 무방향 계좌쌍 기반 시간·금액 매칭 절차를 적용하여 중복을 정밀하게 식별한다. 이러한 프레임워크는 표준화, 대표 레코드 선정, 시각화, 로그 기록 전 과정을 체계화함으로써 분석의 정확성과 데이터 신뢰성을 동시에 확보한다.

Ⅰ. 서론

수사 현장에서 통신·계좌 데이터 수집·분석 과정은 범죄사실 규명과 증거 확보를 위해 필수적이다[1]. 은행과 통신사로부터 개별적으로 수신한 원본 데이터를 표준화하여 통합하는 절차는 수사 데이터 활용의 출발점이 된다[2]. 병합 과정에서 동일 이벤트, 동일 거래가 중복되는 현상이 빈번히 발생하며 분석 결과의 타당성을 저해할 수 있다. 통신 데이터는 두 개 이상의 전화번호에 대해 통신사실확인자료를 요청하면, 요청한 번호 간에 발생한 송수신 기록이 각각의 요청 건에 포함되어 중복된다. 계좌 데이터에서도 두 개 이상 계좌번호를 대상으로 금융계좌추적용 압수수색영장을 집행할 경우, 동일한 자금 이체가 송금 계좌와 수취 계좌의 거래 내역에 각각 기록되어 병합 시 이중으로 남는다. 이러한 중복은 시각화 과정에서 거래나 통화 횟수를 실제보다 과대하게 표현하여 네트워크 구조와 빈도 기반 지표를 왜곡한다. 특히 관계망 분석이나 시계열 패턴 분석과 같이 정량 지표를 활용하는 분석에서는 중복으로 인한 수치 왜곡이 수사 판단의 오류로 이어질 수 있다. 이 연구는 통신과 계좌 데이터 표준화 및 병합 과정에서 발생하는 중복을 체계적으로 식별하는 방안을 제안한다. 제안 방법은 시간, 금액, 착·발신 번호, 계좌번호 등 핵심 필드를 활용한 복합키 설계와 규칙 기반 필터링 절차를 중심으로 구성되며, 이를 통해 분석의 정확성과 시각화 결과의 신뢰성을 제고하는 데 목적이 있다.

Ⅱ. 국제적 수사 데이터 품질 문제와 사례

데이터 중복과 품질 저하 문제는 국내에 국한되지 않는다. 해외에서도 수사 데이터의 변환·통합 과정에서 유사한 오류가 발생하여 심각한 법적 영향을 초래한 사례가 보고된 바 있다. 덴마크 통신 스캔들은 2010년부터 2019년까지 덴마크 경찰 텔레센터(Telecentre)가 형사사건 수사에 활용되는 과거 통화기록(Historical Call Data Records, HCDR)을 처리하는 과정에서 다수 오류가 발생한 사건이다[3]. HCDR은 기지국 위치, 수·발신 번호, 통화 시간 등 통신 메타데이터를 포함하며, 수사에서 알리바이 검증, 범행 시간대 추정, 공범 관계 식별 등 핵심 증거로 사용된다. 이 사건에서는 데이터 변환 과정에서 행(row) 누락, 형식 해석 오류, 데이터 불일치 등 중대한 문제가 발견되었다. 덴마크 정부는 사건 심각성을 인식하고 10,000건 이상 형사사건을 재검토하였다. 오류의 근본 원인은 여러 통신사로부터 제공받은 다양한 원시 데이터 형식을 단일 형식으로 변환하는 과정에 있었다. 2011년부터 2019년 사이, 외부 조사에서 확인된 원시 데이터 형식만 100종 이상이었다. 변환 시스템은 이러한 변화를 완벽히 처리하지 못했고, 일부 행이 누락 되거나 데이터 의미가 변형되는 오류가 발생하였다. 텔레센터와 경찰청 간 데이터 품질 검증 절차도 미흡하였다. 변환된 데이터와 원시 데이터 행 수를 비교하는 단순 검증만 수행되었으며 내용 정확성이나 완전성은 확인되지 않았다. 오류의 지속에는 편향(bias)과 과신(overconfidence)이 중요한 역할을 하였다. 텔레센터 직원들은 변환 시스템이 정확하게 작동한다고 믿었고, 경찰청과 검찰도 이를 신뢰하였다. 보고서와 변환 데이터가 형식상 일치하는 것이 확인되면 데이터 품질이 보장되었다고 판단하였으며, 이는 각 단계에서 신뢰를 강화하는 편향 눈덩이 효과(bias snowball)로 작용하였다. 이러한 신뢰 구조는 오류를 조기에 발견하지 못하게 만들었고, 일부 사건에서는 잘못된 기소와 재판으로 이어질 위험을 높였다.

Ⅲ. 데이터 중복 발생 구조와 영향 분석

1. 통신데이터

통신 데이터는 수사기관이 통신사실확인자료를 요청하면 각 통신사에서 엑셀 파일 형태로 제공한다. 제공되는 항목은 발신번호, 착신(역발신)번호, 통화시작일시, 통화종료일시, 통신사명(업체명), 발신위치(기지국), 서비스 구분 등이다. 대상 전화번호의 통신사에서는 해당 번호의 발신과 역발신 내역, 다른 통신사에서는 역발신 내역만 제공한다. 요청 대상 전화번호가 두 개(A, B) 이상인 경우 A의 통신사에서 제공받은 A→B 발신 내역과, B의 통신사 외에서 제공받은 B←A 역발신 내역이 중복된다. A 또는 B의 통신 데이터를 개별 분석하면 문제가 되지 않지만, 두 전화번호 간 관계를 분석하기 위해 표준화 후 통합하면 중복이 불가피하다. 이런 중복값은 통신 3사(SKT, KT, LGU+)와 같은 경우 서비스 구분을 제외하고 대부분 항목이 일치하나, 별정통신사는 서비스 구분과 함께 통화종료일시

값이 다른 경우도 발생한다.

<표1> 통신 데이터 중복값 예시

구분	발신번호	착신번호	시작일시	종료일시	업체명	발신위치	서비스 구분	비고
통신3사	01053710000	01056690000	2017-10-24 17:00:05	2017-10-24 17:01:46	KT 이동	경기도 성남시 분당구 판교동 291-0	VOLTE-망외음성(VOLTE간)	
통신3사	01053710000	01056690000	2017-10-24 17:00:05	2017-10-24 17:01:46	KT 이동	경기도 성남시 분당구 판교동 291-0	VOLTE음성	
통신3사	01053710000	01056690000	2017-10-26 16:04:31	2017-10-26 16:06:59	KT 이동	서울특별시 구로구 구로동 1125-6	VOLTE-망외음성(VOLTE간)	
통신3사	01053710000	01056690000	2017-10-26 16:04:31	2017-10-26 16:06:59	KT 이동	서울특별시 구로구 구로동 1125-6	VOLTE음성	
별정통신사	01083190000	01075530000	2022-08-17 13:45:13	2022-08-17 13:45:21	LGU+	(04323) 서울시 용산구 동자동 12-0	VOLTE 음성	
별정통신사	01083190000	01075530000	2022-08-17 13:45:13	2022-08-17 13:45:22	스마텔	(04323) 서울시 용산구 동자동 12-0	volte 음성	
별정통신사	01058490000	01072370000	2022-08-17 16:13:18	2022-08-17 16:13:53	LGU+	(61957) 광주시 서구 지평동 1171-5	VOLTE 음성	
별정통신사	01058490000	01072370000	2022-08-17 16:13:18	2022-08-17 16:13:54	큰사람	(61957) 광주시 서구 지평동 1171-5	volte 음성	

이러한 중복은 데이터 생성 단계에서부터 구조적으로 내포되어 있다. 통신사와 별정통신사 간 망 연동 과정에서 각 사업자가 자체 과금·관리 목적의 로그를 생성하고, 수사기관에는 사업자별 로그를 개별 제공하기 때문이다. 여기에 전화번호 단위로 통신사실확인자료를 요청하는 절차가 더해지면서, 요청한 번호 간에 발생한 송수신 기록이 각 요청 건에 중복된다. 수사기관에서 사용하는 표준화 프로그램(i1, A1)도 데이터 중복의 원인이 된다. 이 프로그램은 각 통신사에서 제공받은 데이터를 표준화한 뒤 하나의 파일로 병합하는 방식이므로, 동일 이벤트가 중복될 가능성이 존재한다. 데이터의 중복은 네트워크 분석에서 동일 노드 간 연결(링크)을 증가시키고, 요청 대상 간 통화 빈도를 실제보다 과대(2배) 계산하게 만든다. 시계열 분석에서는 동일 시점에 중복이벤트가 집중되는 것처럼 표현되어 특정 시기 활동량이 왜곡된다. 그 결과 범죄 연루 가능성 평가나 공범 관계 추정과 같은 분석에서 부정확한 판단이 이루어질 가능성이 존재한다.

2. 계좌거래데이터

계좌 거래 데이터는 수사기관이 금융계좌추적용 압수수색영장을 집행하여 금융기관으로부터 제공받는다. 제공되는 항목에는 입출구분, 계좌주, 금융기관명, 계좌번호, 거래일자, 거래시간, 거래구분, 취급점, 입금·출금 적요, 거래금액, 거래 후 잔액, 상대은행, 상대예금주, 상대계좌번호, 단말번호, IP 주소, MAC 주소 등이 포함된다. 요청 대상 계좌번호가 두 개(A, B) 이상인 경우, A의 은행에서 제공받은 A→B 출금 내역과 B의 은행에서 제공받은 B←A 입금 내역이 동일 거래임에도 불구하고 각각 독립적으로 기록된다. 개별 계좌를 분석할 때는 문제가 없지만, 두 계좌 간 자금 흐름을 파악하기 위해 표준화 후 병합하면 동일 거래가 이중으로 남게 된다. 통신 데이터의 경우 종료일시와 서비스 구분을 제외하면 대부분 항목이 일치하지만, 계좌 거래 데이터는 계좌번호와 거래일자를 제외하고 대부분 항목이 일치하지 않는다. 동일 거래라도 수수료 부과 여부에 따라 거래금액이 일치하지 않을 수 있다. 추가로 송금 후 입금이 이루어져야 함에도 각 금융기관의 로그 생성 시점 차이로 인해 출금 시각이 입금 시각보다 늦게 기록되는 경우도 발생할 수 있다. 예를 들어, <표2>에서 2011년 5월 6일의 동일 거래는 출금 시각이 11:44:06이고 입금 시각이 11:44:00으로, 시간상 역전된 형태로 나타나기도 한다.

<표2> 계좌 데이터 중복값 예시

구분	계좌주	금융기관	계좌번호	거래일자	거래시간	거래구분	취급점	입금적요	입금액	출금적요	출금액	상대계좌번호
출금	박OO	우체국	7004300000000	2011-04-25	08:17:00	전자금융타행이체	0310431		-	박OO	3,000,000	043400000000
입금	박OO	대구은행	043400000000	2011-04-25	08:17:17	폰뱅킹	청평우체국	체신박OO	3,000,000		-	7004300000000
출금	박OO	대구은행	043400000000	2011-05-06	11:44:06	인터넷	고객김동센터		-	체신박OO	2,000,600	7004300000000
입금	박OO	우체국	7004300000000	2011-05-06	11:44:00	전자금융타발타행이체	0310431	박OO	2,000,000		-	043400000000
출금	박OO	우체국	7004300000000	2011-08-22	11:56:00	전자금융타행이체	0724399		-	박OO	1,000,000	043400000000
입금	박OO	대구은행	043400000000	2011-08-22	11:56:46	폰뱅킹	청평우체국	체신박OO	1,000,000		-	7004300000000
출금	박OO	우체국	7004300000000	2011-09-26	09:55:00	전자금융타행이체	0724399		-	박OO	470,000	043410390343
입금	박OO	대구은행	043400000000	2011-09-26	09:55:09	폰뱅킹	청평우체국	체신박OO	470,000		-	7004300000000

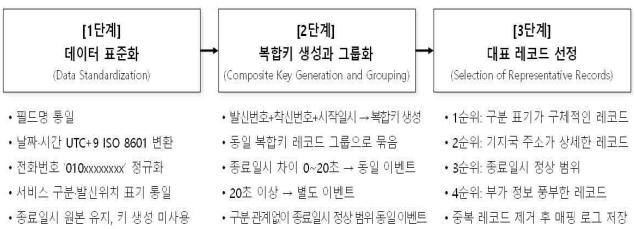
이러한 중복은 거래 데이터의 구조적 특성에서 비롯된다. 금융기관은 자금 이체 발생 시 송금자와 수취자 각각의 계좌 원장에 독립적으로 기록을 남기며, 수사기관은 각 계좌의 거래 내역을 별도로 수집한다. 특히 여러 금융기관에서 데이터를 제공받는 경우, 기관별 파일 포맷과 필드 구성, 날짜·시간 표기 방식이 달라 단순 비교만으로 동일 거래 여부를 판별하기 어렵다. 이러한 중복은 금전 흐름 네트워크에서 링크 수를 증가시키고, 총 거래금액과 거래 횟수를 실제보다 과대 산출한다. 그 결과 자금 추적의 정확성이 저하되고 수사 자원이 불필요하게 소모되며 분석 전반의 신뢰성이 떨어진다.

Ⅳ. 데이터 중복 식별을 위한 프레임워크

1. 통신데이터 중복 식별 알고리즘 설계

통신 데이터 중복 문제를 해결하기 위해 발신번호·착신번호·시작일시를 핵심 필드로 한 복합키를 중심으로 중복을 식별하되, 종료일시는 복합키에 포함하지 않고 검증 단계에서만 활용하는 알고리즘을 설계하였다. 1단계는 데이터 표준화이다. 서로 다른 사업자에서 제공받은 데이터의 필드명을 통일하고, 날짜·시간은 UTC+9 기준 ISO 8601(YYY-MM-DD hh:mm:ss) 형식으로 변환한다. 발신번호와 착신번호는 모든 구분기호를 제거하고 숫자로만 구성된 '010xxxxxxxx' 형태로 정규화한다. 서비스 구분과 발신위치는 대소문자, 공백, 약어를 일관되게 변환하고, 종료일시는 원본값을 유지하되 식별 키 생성에는 사용하지 않았다. 2단계는 복합키 생성과 그룹화이다. 발신번호, 착신번호, 시작일시 세 필드를 결합하여 복합키를 생성하고, 동일 복합키를 가진 레코드를 하나의 그룹으로 묶었다. 그룹 내에서는 종료일시를 비교하여 0~20초 범위의 차이는 동일 이벤트로 간주하고, 20초 이상의 차이는 별도 이벤트로 재분류한다. 이러한 기준은 통화 종료 기록이 사업자별로 최대 수십 초 차이가 날 수 있지만, 20초 이상 차이는 데이터 매핑 오류 가능성이 높다는 실무 분석 결과에 따른 것이다. 서비스 구분 값이 서로 다르더라도 종료일시가 정상 범위에 있으면 동일 이벤트로 처리하였다. 3단계는 대표 레코드 선정이다. 동일 이벤트 그룹 내에서 대표 레코드를 선정하기 위해 우선순위 규칙을 적용하였다. 첫째, 서비스 구분 표기가 구체적인 레코드를 우선 보존하였다. 둘째, 기지국 주소가 상세하게 기록된 레코드를 보존하였다. 셋째, 종료일시가 정상 범위에 있는 레코드를 우선하였다. 넷째, 부가 필드(비고란)의 정보가 풍부한 레코드를 보존 대상으로 선정하였다. 나머지 레코드는 중복으로 분류하고 중복 레코드와 대표 레코드 간 매핑 정보를 별도의 로그 파일로 저장하여 후속 검증과 복원에 활용할 수 있도록 하였다. 이 알고리즘은 네트워크 분석에서 동일 노드 간 불필요한 링크를 줄이고, 시계열 분석에서 특정 시점의 통화량이 과대 표시되는 현상을 완화한다. 특정 인물의 통신량이 실제보다 부풀려지는 오류를 방지하여 범죄 연루 가능성 평가와 공범 관계 추정의 정확성을 높인다.

<그림1> 통신 데이터 중복 식별 알고리즘

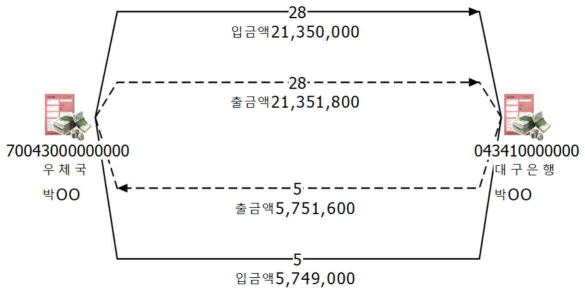


2. 계좌데이터 중복식별 알고리즘 설계

계좌 거래 데이터의 중복은 다계좌 병합 분석 과정에서 거래 건수와 금액 합계를 왜곡시켜 분석 정확성을 저하시킨다. 이를 해소하기 위해 무방향 계좌쌍을 기반으로 한 중복 식별 알고리즘을 설계하였다. 1단계는 데이터 표준화이다. 여러 금융기관에서 제공받은 데이터의 필드명을 통일하고, 거래일시는 UTC+9 기준 ISO 8601(YYYY-MM-DD hh:mm:ss) 형식

으로 변환한다. 계좌번호는 모든 구분기호를 제거하고 숫자만 남도록 정규화하며, 입출구분은 '입금'과 '출금'으로 통일한다. 거래구분·적요·취급점 표기는 대소문자, 공백, 약어를 일관되게 변환하여 비교 가능성을 높인다. 2단계는 무방향 계좌쌍 생성과 그룹화이다. 출금계좌번호와 입금계좌번호를 비교하여 작은 값을 앞에, 큰 값을 뒤에 두는 방식으로 정렬함으로써 방향성을 제거한다. 동일 계좌쌍을 하나의 그룹으로 묶고, 거래일시와 거래금액은 그룹 키에서 제외하되, 이후 비교 단계에서 보조 조건으로 활용한다. 3단계는 그룹 내 순차 비교와 후보군 생성이다. 동일 계좌쌍 그룹 내 거래를 시간순으로 정렬한 후 인접 거래를 비교한다. 거래일시 차이가 ±60초 이내이고 거래금액 차이가 허용오차(예: ±3,000원, 수수료) 범위 안에 있으면 후보군으로 지정한다. 이 과정에서 상대계좌와 상대은행의 교차 일치 여부를 우선 확인하며, 거래금액, 거래구분·채널 표준화 일치 여부, 적요·취급점의 유사도를 함께 평가한다. 4단계는 수수료 포함 여부 판별이다. 거래구분에서 이체 행위를 식별한 뒤, '전자금융타행이체', '폰뱅킹' 등 타행송금에 해당하는 경우 수수료 부과 가능성이 높다고 판단한다. 이어서 후보군의 출금액과 입금액 차이를 계산하여, 400원, 500원, 1,000원 등 사전에 정의된(은행 거래구분별 수수료 데이터 확보) 소액 패턴과 일치하면 수수료로 판정한다. 5단계는 매칭 결정과 대표 레코드 선정이다. 후보군 내에서는 출금 집합과 입금 집합 간 최소 비용 매칭을 수행하여 1:1 대응을 확정한다. 다대일·일대다 매칭의 경우 금액 부분합 검증을 통해 일치 여부를 확인한다. 대표 레코드는 출금에서 입금 방향으로 정규화하며, 순금액은 입금액, 수수료는 출금액과 입금액의 차이로 산출한다. 거래구분·채널은 구체적인 표기를 우선 채택하고, 적요·취급점은 정보량이 많은 항목을 보존한다. 6단계는 시각화와 중복 링크 표기이다. 네트워크 분석 시 실거래는 실선 링크로, 중복 거래는 점선 링크로 표시하여 분석자가 직관적으로 구분할 수 있도록 한다. 점선 링크에는 해당 거래의 출금액과 입금액을 함께 표기하고, 실선과 점선을 한 쌍으로 배치해 금액 차이와 중복 여부를 한눈에 확인할 수 있게 한다. 이 방식은 중복 거래를 단순히 제거하는 대신 시각적으로 보존함으로써 재현성과 투명성을 확보하며, 법정 제출 자료나 보고서에서 데이터 처리 과정을 설명하는 근거로 활용할 수 있다.

<그림2> 계좌 데이터 시각화 및 중복 링크 표기



7단계는 로그 기록 및 예외 처리이다. 대표 레코드와 원본 레코드 간 매핑 정보를 별도의 로그 파일에 저장하여 검증과 복원에 활용할 수 있도록 한다. 거래일시·거래금액이 허용 범위를 초과하거나 교차 일치가 확인되지 않는 건은 예외 큐로 이관하여, 보조 메타데이터 확보 또는 수동 검증 후 재처리한다. 이러한 로그는 추후 재분석이나 패턴 학습(머신러닝 기반 사기 탐지)에도 활용 가능하다. 이 알고리즘은 동일 거래의 중복 반영을 방지하여 다계좌 병합 분석 시 네트워크 분석 왜곡을 최소화하고, 거래량 및 중심성 지표의 과대 계상을 완화한다. 수수료로 인한 송금액 부풀림 오류를 방지하고, 점선 표기를 통해 중복 거래 정보를 투명하게 전달함으로써 범죄 연루 가능성 평가와 공범 구조 추정의 정확성을 높인다.

<그림3> 계좌 데이터 중복 식별 알고리즘

V. 결론

이 연구는 수사 과정에서 수집되는 통신 데이터와 계좌 거래 데이터의 병합·분석 시 발생하는 구조적 중복 문제를 실증적으로 분석하고, 이를 식별하기 위한 알고리즘 기반 처리 프레임워크를 제안하였다. 중복 데이터는 네트워크 분석과 시계열 분석 등 정량 분석에서 연결 구조와 빈도 지표를 왜곡하여 수사 판단의 신뢰성을 저하시킨다. 통신 데이터에서는 발신·착신 번호와 시작일시를 핵심 식별 요소로 한 복합키 설계와 종료일시 허용 범위 검증을 통해, 계좌 거래 데이터에서는 무방향 계좌쌍 생성과 금액·시간 차이 기준을 활용한 후보군 매칭 절차를 통해 중복을 효과적으로 식별·제거할 수 있음을 확인하였다. 제안한 프레임워크는 표준화, 복합키 생성, 그룹화, 대표 레코드 선정, 시각화 및 로그 기록 일련 과정을 구조화하여, 데이터 처리의 재현성과 투명성을 동시에 확보하였다. 특히 중복 거래를 단순히 삭제하는 것이 아니라, 시각적 표기와 로그 기록을 병행함으로써 분석 과정의 검증 가능성을 높였다. 이는 법정 제출 증거나 수사 보고서 작성 시 데이터 신뢰성을 뒷받침하는 근거로 기능할 수 있다. 제안된 방법은 수사기관이 보유한 표준화 프로그램의 기능 한계를 보완하고, 대규모 다원적 데이터 환경에서 분석 정확성을 향상시키는 실무적 대안을 제시한다. 정책적으로는 표준화 프로그램(i1, A1)의 기능 개선과 법령·내규에 데이터 표준화 및 중복 식별 가이드라인을 반영하는 것이 필요하다. 이를 통해 기관 간 데이터 호환성을 높이고, 수사자료의 신뢰성을 제도적으로 보장할 수 있다. 향후 연구에서는 다양한 사업자·기관 간 데이터 연동 환경에서 알고리즘의 성능을 자동화·고도화하는 방안을 검토해야 한다. 특히 머신러닝 기반의 이상 탐지 모델과 결합하여, 중복 제거와 동시에 의심 거래나 통화 패턴을 실시간으로 식별하는 지능형 수사 지원 체계로 확장할 수 있을 것이다.

참 고 문 헌

[1] A. T. T. Wong, "Opportunities and challenges of big data analytics in crime investigation," International Annals of Criminology, pp. 1 - 15, 2025.

[2] M. S. Gal and D. L. Rubinfeld, "Data standardization," New York University Law Review, vol. 94, p. 737, 2019.

[3] L. W. Lentz and N. Sunde, "The use of historical call data records as evidence in the criminal justice system－lessons learned from the Danish telecom scandal," Digital Evidence and Electronic Signature Law Review, vol. 18, p. 1, 2021.