

웹스크래핑과 공간 클러스터링 기술을 활용한 전화금융사기(보이스피싱) 수사기법 연구

김혜진

To cite this article : 김혜진 (2020) 웹스크래핑과 공간 클러스터링 기술을 활용한 전화금융사기(보이스피싱) 수사기법 연구, 한국경찰연구, 19:3, 45-62

① earticle에서 제공하는 모든 저작물의 저작권은 원저작자에게 있으며, 학술교육원은 각 저작물의 내용을 보증하거나 책임을 지지 않습니다.

② earticle에서 제공하는 콘텐츠를 무단 복제, 전송, 배포, 기타 저작권법에 위반되는 방법으로 이용할 경우, 관련 법령에 따라 민, 형사상의 책임을 질 수 있습니다.

www.earticle.net

웹스크래핑과 공간 클러스터링 기술을 활용한 전화금융사기(보이스피싱) 수사기법 연구

김 혜 진*

차 례

- I. 서론
- II. 이론적 배경 및 선행연구
- III. 연구방법
- IV. 분석결과
- V. 결론

〈국 문 초 록〉

본 연구는 빅데이터 분석 기법인 웹스크래핑과 공간클러스터링 기술을 활용하여 전화금융사기(보이스피싱) 수사에 활용 될 수 있는 지리 기술 개발을 목적으로 수행되었다. 연구 결과, 온라인상에 공개되어 있는 금융회사의 지점정보를 웹스크래핑 기술을 통해 자동 수집하여 압수수색영장 집행 시 제공되는 지점코드를 지점정보(지점명, 지점주소, 지점 연락처 등)로 자동 변환하고, 인출 지점들의 좌표정보를 공간클러스터링화(Hierarchical clustering) 함으로서 해당 지점들의 중심지를 파악하여 인출책의 근거지 추정에 사용될 수 있는 정보를 도출하는 알고리즘을 생성하였다. 최근 다양한 방식으로 개발·확산되고 있는 빅데이터 분석 기법들이 실제 수사현장에서 어떠한 방식으로 효율성 있게 사용될 수 있는 지에 대한 구체적인 연구사례로 해당 연구가 활용될 것을 기대한다.

주제어: 전화금융사기, 보이스피싱, 웹스크래핑, 공간클러스터링, 수사기법 개발

* 경찰대학 치안정책연구소 연구관

I. 서론

2020년 2월 전화금융사기(보이스피싱)에 속아 430만원의 피해를 당한 20대 취업준비생이 고통 끝에 자살하는 사건이 발생했다. 검사를 사칭한 범인에게 통화를 중단하면 공무집행방해로 2년 이하의 징역에 처할 수 있다는 협박을 받은 피해자는 무려 11시간 동안 전화를 끊지 못하고 압박에 시달렸으며, 유서에도 고의로 수사를 방해한 것이 아니라는 호소를 남겼다. 두 달 후 같은 수법에 5억원의 사기피해를 당한 대한상공회의소 직원도 절망 속에서 극단적인 선택을 했다. 이렇듯 막대한 재산피해를 넘어 치명적인 인명손실까지 발생시킬 수 있는 보이스피싱은 최근 5년간 누적 발생빈도와 피해액 모두에서 역대 최대를 기록했다. 경찰청 통계¹⁾에 따르면 2019년 상반기 기준 5년 간 합산된 전화금융사기의 발생건수는 직전 5년 대비 18%가 증가한 11만 7000여건에 달했는데, 특히 학생, 주부, 노인 등 경제 취약 계층을 대상으로 한 사기가 전체의 77%에 달했다. 사회적 약자를 대상으로 한 지능범죄의 이러한 급속한 증가는 매년 수법이 첨단화 되고 있는 전화금융사기 수사에 보다 과학적이고 신속한 수사기법의 개발 및 적용이 필수적인 시점이라는 점을 강력하게 시사한다고 할 수 있다.

전화금융사기 사건 수사 과정에 있어 사기계좌의 인출이 이루어진 은행지점 정보를 신속하게 확보하는 일은 매우 중요하다. 물리적 범죄 발생지를 특정하기 어려운 보이스피싱 사건의 경우, 사기계좌에서 인출이 이루어진 지점을 범죄 발생지로 가늠하여 관할 경찰서를 특정 할 수 있으며, 지급정지와 기소 전 몰수 보전에 용이하고, 무엇보다도 현금 수송 업무를 맡는 인출책의 검거에 필수적인 CCTV(Closed Circuit Television) 기록을 해당 지점에 빠르게 요청²⁾할 수 있기 때문이다. 또한 사기계좌를 제외한 단서가 전무한 사건의 경우, 인출책이 거점지(집이나 직장 근처 등) 인근에서 인출을 했을 가능성이 존재하기 때문에 사후 인출책의 근거지를 추정하는 것에도 해당 정보가 활용될 수 있다.

문제는 수사기관에서 금융계좌추적용 압수수색영장을 발부받아 금융기관에 집행할 경우, 금융기관에서는 사기계좌의 거래내역을 회신해 주는데 인출이 발생한 지점정보를 문

1) 국정감사 김영호 의원의 경찰청 보이스피싱 통계 보도자료 참조(검색일 2020.07.02.)

<http://www.safetimes.co.kr/news/articleView.html?idxno=77454>

2) 사기계좌의 인출기록은 최소 십 여건부터 최대 몇 천건으로 사건별로 다양하기 때문에 이들의 출금 지점을 일일이 웹상에서 검색하는 것은 신속한 수사를 방해하는 최대 요인 중 하나라 할 수 있다. 특히, 금융기관이 CCTV 기록을 보관하는 기간이 통상 길지 않아 지점정보가 늦게 취득될 경우 인출책 검거에 가장 결정적인 증거인 CCTV 기록을 소실할 가능성이 있다.

자가 아닌 숫자형태의 코드로 제공한다는 점이다. 지점코드를 지점정보(지점명, 지점주소, 지점 연락처 등)로 일괄 변환하여 신속하게 수사업무를 보조할 수 있는 분석 프로그램이 전무하기 때문에, 현재는 수사관들이 금융결제원 웹사이트(www.kftc.or.kr)의 금융회사코드 조회 메뉴 또는 각 은행 웹사이트의 지점정보 검색 메뉴에 개별 조회하는 방식이 사용되고 있다. 그러나 이러한 건별 개별 조회의 경우, 시간 소모가 커 신속한 수사 진행을 방해할 뿐만 아니라, 지점들의 분포나 인근 CCTV들의 위치 등에 대한 공간 정보를 제공하고 있지 못해 인출책의 근거지에 대한 수사관의 직관적인 추론을 돕지 못한다는 한계점이 있다.

따라서 본 연구는 빅데이터 분석기법들을 활용하여 이러한 전화금융사기 사건 수사에서 사용될 수 있는 분석 프로그램을 개발하고자 한다. 상세 개발사항에 있어 첫째, 온라인상에 공개되어 있는 금융회사의 지점정보(지점코드를 포함)를 웹스크래핑 기술을 통해 자동 수집하여 지점코드를 지점정보(지점명, 지점주소, 지점 연락처 등)로 자동 변환하고, 둘째, 변환된 은행 지점의 주소지를 좌표정보로 변환(Geocoding)하여 지도상에 시각화하고, 마지막으로 이를 다시 공간 클러스터링화 함으로서 해당 지점들의 중심지를 파악하여 인출책의 근거지 추정에 사용될 수 있는 유용한 정보를 제공하고자 한다.

II. 이론적 배경 및 선행연구

1. 웹스크래핑 기법을 활용한 치안과학연구

웹스크래핑(Web scraping)은 웹(http)상에서 필요한 데이터를 컴퓨터 프로그램으로 하여금 자동으로 추출하여 수집하는 기술로, 웹크롤링(Web crawling)이라는 용어와 교차 사용되기도 한다. 웹사이트의 구조 및 프로그래밍 언어에 대한 공학적 이해를 필요로 하기 때문에 사회과학 방법론을 주로 활용하는 치안과학 분야에서는 2010년 이후부터 범죄의 예방과 수사 기법 개발을 목적으로 조금씩 사용되기 시작했다. 예를 들어 개인정보 침해 예방 프로세스에 대해 연구한 배영효와 이금녀(2013)는 인터넷 검색 사이트 구글(www.google.com)에서 개인정보 판매글 30개를 표본 분석하여 개인정보 판매와 관련된 주요 키워드(디비 팝니다, 대출 디비 등) 199개를 추출한 후, 웹크롤링 프로그램을 통해 해당

단어들을 사용한 판매 게시글을 찾아내 범죄 관련성이 높은 해당 URL들을 방송통신심의 위원회에서 자동 차단하는 예방 방안을 제안함으로써 웹스크래핑 기술을 활용한 수사 기법에 대한 실무적 예시를 최초로 제공했다.

장종욱 외(2018)는 온라인 상 중고거래 구매자들을 실제 거래 사이트와 유사한 위·변조 웹사이트로 유인하여 로그인 정보와 결제 금액을 가로채는 범죄행위를 막기 위해 피싱 사이트를 구별, 탐지하는 기술을 개발했다. 위조된 사이트의 HTTP 헤더 정보와 소스코드를 웹스크래핑을 통해 수집한 후 이를 기존의 정상적인 웹사이트의 구성 정보와 비교함으로써 문제가 있는 사이트를 자동 판별하는 모델을 구축하고 위·변조 웹사이트로 인한 개인 정보 유출과 금전적 피해를 최소화 하는 성과를 남겼다.

웹스크래핑 기법은 또한 언론에 의해 집중도 있게 다루어진 치안 이슈의 효율적인 파악을 위해서도 사용되었다. 조주연과 조경원(2018)은 1993년부터 2018년에 이르기 까지 26년간 미디어에 의해 다루어진 청소년 문제를 탐색하기 위해 주요 언론(조선일보, 동아일보, 중앙일보)의 웹사이트에서 12,946건의 뉴스 기사를 자동 수집하였다. 청소년 관련 기사의 화제가 환경, 정책, 문화 등의 변화와 변화함을 토픽 모델링을 통해 확인하였으며 특히 2018년의 청소년 관련 기사들의 가장 큰 주제어는 성범죄임을 확인하였다.

비교적 최근에 이루어진 김혜진(2020)의 연구 또한 형사사법기관의 공식 범죄 통계에서 포착하고 있지 못한 신종 성범죄에 관한 치안 수요를 보다 시의성 있게 확인하려는 목적에서 수행 되었으며, 2018년~2019년 사이 네이버 포털에 게재된 성범죄 뉴스기사 3,764건을 웹크롤러를 통해 자동수집 하였다. 자연어처리(Natural Language Processing: NLP) 기법인 네트워크 분석과 토픽 분석을 시행한 한 결과, 뉴스 기사들이 6가지 주요 유형의 성범죄(디지털 성범죄, 사회 고위층 성상납, 직장 내 성폭력, 청소년 대상 지능형 성범죄, 해외 유명인사 성범죄)에 관해 집중적으로 다루고 있음을 발견해 이에 대한 치안정책 수립의 필요성에 대해 강조하였다.

2. 공간군집분석 기법을 활용한 치안과학연구

군집분석(Clustering analysis)은 대표적인 비지도학습(Unsupervised learning) 기법 중 하나로 입력변수(X)와 출력변수(Y)의 관계에 대해 모델링 하는 지도학습(Supervised learning)과 달리 입력변수(X)들 간의 관계성에 대해 모델링하여 유사한 특성을 가진 데이터들 끼리 묶어 그룹화를 시켜주는 직관적 분석 구조를 가지고 있다. 따라서 군집분석 모

델을 공간 데이터에 적용할 경우, 이벤트(본 연구의 경우 인출이 이루어진 은행지점 위치)가 발생한 특정 지역에서의 군집 유무와 군집별 중심점 위치에 관한 유용한 정보를 얻을 수 있게 된다.

활용 역사가 짧은 웹스크래핑과 달리, 공간 군집분석은 치안과학 분야에서 비교적 활발히 사용되어온 분석방법이라 할 수 있다. 예를 들어, 현행 지구대가 출동시간 단축이라는 측면에서 적절한 위치에 자리 잡고 있는지를 검증하기 위해 황초희와 서용철(2008)은 2006년 4월부터 2007년 6월 사이에 수집된 부산 경찰관서 관할 내 지구대의 112 신고 자료 150만 건을 대상으로 K-Means 군집분석 모델을 시행하였다. 그 결과 관내에 7개의 범죄 발생 밀집지(군집)가 존재한다는 것을 확인하였으며, 해당 군집들이 지구대 관할별로 균일하게 존재하는 것이 아니라, 특정지구대(가야지구대, 당감지구대) 근처에서 몰려있어 부산 진구 지구대들의 위치는 적극적 범죄 예방을 위해 신설 내지는 이전이 필요하다는 결론을 제시했다.

김주환(2014)은 범죄예방환경설계(Crime Prevention Through Environmental Design: CEPTED)에 기반 한 안전 시스템과 장비를 도입하기 적절한 장소를 탐색하려는 목적에서 안양시에서 2011년과 2012년 사이에 발생한 5대 범죄 및 112 신고접수(10,863건) 자료를 활용하여 공간 군집분석(Nearest Neighbor Hierarchical Spatial Clustering)을 수행하였고, 그 결과 다양한 범죄 유형 중 특히 절도범죄, 폭력범죄, 성폭력범죄가 특정한 지역에서 패턴을 형성하며 집중적으로 발생함을 확인하였으며, 특히 연구 지역 내 전체 105개의 위험 지역이 존재한다고 보고하였다.

이전학 외(2016)는 여성가족부의 성범죄자알림e 사이트의 정보를 활용하여 2002년에서 2015년 사이 전국에서 발생한 성범죄(N=4,500)의 공간적 분포 특성을 탐색했다. 성범죄 핫스팟 지역을 보다 정교하게 분류하기 위해 Ward Method에 기반 한 계층적 군집 분석을 실시한 결과, 지역 환경요인(인구학적요인, 사회경제요인, 공간구조요인, 방어기제요인)에 따라 성범죄 다발지역이 주거지형, 도심형, 공장지형, 일반도시형, 주변도시형의 5가지 그룹으로 나누어지는 것을 확인하였다. 이들 연구자들은 주거지형은 대도시 내 주거 밀집도가 높은 지역을, 도심형은 상업지역의 면적이 높고 녹지 면적이 드문 지역을, 공장지형은 외국인 인구 밀도가 높은 기계공장이 밀집한 지역을, 일반도시형은 인구밀도와 여성 1인 가구 비율이 높은 지역을, 주변도시형은 주간인구 및 인구밀도는 높은 반면 단독가구의 비율은 낮은 지역을 의미한다고 부연하였다.

비교적 최근에 이루어진 연구에서 노기윤과 이창배(2018)는 2015년과 2016년 사이 서울시 동작구에서 발생한 폭력범죄(N=2,005)의 발생 위치 정보를 기반으로, 폭력 범죄 다발

지 주변의 특성을 커널 밀도 함수와 최근린 지수에 기반 해 군집화 하는 시도를 했다. 그 결과 동작구의 특정지역(노량진역, 신대방삼거리역, 흑석역 등)에서 특히 폭력 범죄가 집중적으로 발생하며, 주점, 지하철역, 식당 등 지역의 상업적 특성 조합에 따라 대부분의 폭력 범죄의 발생을 설명할 수 있다고 주장함으로써 이를 근거로 한 지리적 프로파일링 시스템 및 예방 활동의 구축을 제안하였다.

이밖에도 범죄가 공간 특성에 따라 불균등하게 분포, 발생한다는 사실을 확인함에 있어 토지용도 특성과 강력범죄(강도, 절도)사이의 관계성을 탐색 하거나(김걸 & 김병선, 2009), 보행자 무단 횡단 핫스팟과 해당 지역의 특성(지하철역, 버스정류장, 재래시장과 같은 상업시설 근처)을 파악하거나(조정윤 외, 2018), 연쇄 강도 강간범의 거점을 예측하는(이문국, 2015) 등, 다수의 국내문헌들이 다양한 방식으로 공간 군집분석 기법을 활용하였다. 하지만 대부분의 선행연구들이 범죄 발생 밀집지(Hotspot)의 지정학적 위치 및 이들의 지역사회 환경 특성을 파악하는 것에 국한 되었을 뿐, 해당 기술들이 실무 수사 과정에서 활용된 사례는 없었다. 따라서 본 연구는 웹스크래핑과 공간 군집분석을 접목한 빅데이터 분석 기술을 수사 목적에 맞추어 새롭게 개발하는 방법론을 제안하고자 한다. 구체적으로 파이썬(Python) 언어에 기반 한 프로그래밍 작업을 통해 금융기관의 지점정보를 자동 수집하고, 이들을 금융기관에서 제공한 사기계좌 거래내역의 인출 지점코드와 연결함으로써 인출책의 신속한 검거를 돕고, 더 나아가 사기계좌 인출지점들의 밀집지 및 중심지를 특정하여 인출책의 근거지를 추정하는 공간 분석을 수행하도록 한다.

Ⅲ. 연구방법

1. 전화금융사기 계좌의 인출 지점코드 데이터

앞서 설명한 바와 같이, 수사관은 보이스피싱 사기계좌에 대한 영장 집행을 통해 이들의 거래내역, 특히 인출지점(일명 지점코드)에 대한 정보를 금융기관으로부터 회신 받는다. 회신 받는 정보들의 예시는 <그림 1>과 같다. 하단에서 “거래점”으로 표기된 은행 “지점코드”는 1~7자리의 숫자로 이루어진 값들로 각 은행권 마다 고유한 번호체계를 가지고 있다. 이러한 지점코드를 지점정보(지점명, 지점주소, 지점연락처 등)에 연결하여 수사와

정에서 사용될 수 있는 범죄정보를 생성하는 분석 프로그램의 개발을 위해 2020년 5월 실제 전화금융사기에 관여 되었던 계좌들의 지점코드(5,200건)를 일선 경찰관서에서 제공 받았다. 수사정보의 보호를 위해 숫자로 이루어진 은행의 지점코드 외, 그 어떤 사건 관련 정보도 수령하지 않았음을 기록하도록 한다.

〈그림 1〉 금융기관의 사기계좌 입출금 내역 회신자료 예시

| 지역농축협 금융거래 내역 | | | | | | | | | |
|---------------|----------|-----------|------------|----|--------|--------|------|------|----------------|
| 2020년 03월02일 | | | | | | | | | |
| 거래일자 | 거래시간 | 거래금액<액> | 잔액<원> | 구분 | 거래점 | 거래내용 | 입출금명 | 은행코드 | 계좌번호 |
| 2020-01-01 | 3:04:24 | 3,800,000 | 11,429,130 | 입금 | 737031 | G-우리은행 | 김기영 | 020 | 10026754297234 |
| 2020-01-01 | 8:02:17 | 410,500 | 11,018,630 | 지급 | 000303 | E-기업은행 | 이윤열 | 003 | 01089341831 |
| 2020-01-01 | 14:53:23 | 1,100,000 | 9,918,630 | 지급 | 001459 | G-국민은행 | 한상수 | 004 | 30699194284127 |

빅데이터 분석 기술이 수사 과정에 어떻게 적용될 수 있는지 방법론을 제안하는 것이 본 연구의 목적이기 때문에 은행 지점 별로 동일한 과정을 반복하고 개별 설명하는 대신, 금융기관 1곳을 선택하여 기법을 개발하는 상세한 과정을 설명하는 방식을 선택했다³⁾. 금융기관들 중 특히 은행 브랜드 평판이 높아 사용자가 많을 뿐⁴⁾만 아니라, 수사관들의 주요 검색 플랫폼인 금융결제원 사이트(<http://m.kftc.or.kr/mobile/data/MobileBankingByCode.do>)에서 명확한 정보를 제공하지 않아 웹크롤링을 통한 정보 수집이 가장 시급한 농협 사이트(http://nonghyup.tritops.co.kr/list_branch.jsp)를 대상으로 웹스크래핑을 시행⁵⁾하였다. 경찰관서에서 제공받아 단일사건에서 수사대상이 된 사기계좌들 중 농협(NH농협은행, 농·축협지점)에서 인출된 계좌는 전체 82건 이었다.

- 3) 각 은행별 사이트마다 DOM(Document Object Model)이 상이하여 웹스크래핑 프로그래밍 코드가 달라지기 때문에 해당 연구에서 모든 은행권 사이트별 연구과정을 개별 설명할 수 없어, 금융회사 1곳을 특정하여 연구를 진행한 점을 설명하도록 한다.
- 4) 한국기업평판 연구소의 은행 브랜드 평판 온라인 기사 참조(2020.08.24 검색)
<http://www.futurekorea.co.kr/news/articleView.html?idxno=138397>.
- 5) 예를 들어 은행 지점번호 '903824'을 검색할 경우 금융결제원 사이트의 지점코드 메뉴는 결과값을 찾을 수 없다고 응답하는 반면, 농협 사이트의 영업점 메뉴는 정확한 결과 값(제주감귤농협 위미지소이며, 주소는 제주특별자치도 서귀포시 남원읍, 대위로, 123-3에 해당)을 리턴해 주는 것을 알 수 있다.

2. 웹스크래핑(Web Scraping Method)

보이스피싱 사기계좌의 인출 지점코드를 지점정보로 일괄 변환하기 위해 웹스크래핑 기술을 활용해 농협's 모든 지점정보를 자동 수집하였다. 웹스크래핑 프로그래밍을 위한 라이브러리(모듈)가 풍부한 파이썬(Python)을 활용, Selenium 모듈의 Webdriver() 함수와 bs4 모듈의 BeautifulSoup() 함수를 적용하여 자동 수집을 하고자 하는 농협 웹사이트의 인터페이스 DOM(Document Object Model)을 분석하였다. <그림 2>에서 제시된 바와 일반검색과 지도검색의 2가지 옵션이 있는데, 전국 농협은행과 농·축협 지점들을 일괄로 수집하기 위해 지도검색 메뉴의 영업점 정보 게시판 선택했다. 게시판의 순서별 이동에 따라 URL이 함께 변하는 정적(Static) 페이지가 아닌, URL에 관계없이 웹페이지 위에 영업점 정보를 불러오는 동적(Dynamic) 페이지였기 때문에 Css_selector 옵션을 통해 게시판의 순번이 자동 이동할 수 있도록 루프를 지정해 주고, 지점정보가 포함되어 있는 <div class = "boardlist"> 태그 속 'tr'을 findAll() 메소드를 호출하여 수집하였다. 페이지 이동 반복문에는 sleep을 3초씩 걸어서 JS 코드가 실행되는 것을 멈추었다. 최종적으로 게시판에서 지점명, 지점구분, 지점코드, 주소, 전화번호를 자동 수집하면서 수집한 결과물을 컬럼별 리스트(List) 형태로 임시 저장한 후, Pandas() 함수의 데이터 프레임으로 변환하여 엑셀 CSV 형식으로 저장하였다.

〈그림 2〉 금융회사 웹사이트의 영업점 안내 웹페이지 구조

은행지점 검색 메뉴 웹페이지

```

<div class="boardlist">
  <table width="100%" border="0" cellspacing="0" cellpadding="0" summary="지점명, 구분, 지점코드, 주소, 전화번호를 보여줍니다.">
    <thead>
      <tr>
        <td class="listtt_center">지점명</td>
        <td class="listtt_center">구분</td>
        <td class="listtt_center">지점코드</td>
        <td class="listtt_center">주소</td>
        <td class="listtt_center">전화번호</td>
      </tr>
    </thead>
    <tbody>
      <tr>
        <td class="listtt_center">NH농협은행</td>
        <td class="listtt_center">NH농협은행</td>
        <td class="listtt_center">1087</td>
        <td class="listtt_center">서울특별시 서초구 강남대로 27</td>
        <td class="listtt_center">02-5140-7330</td>
      </tr>
      <tr>
        <td class="listtt_center">NH농협은행</td>
        <td class="listtt_center">NH농협은행</td>
        <td class="listtt_center">1409</td>
        <td class="listtt_center">서울특별시 서초구 강남대로 140</td>
        <td class="listtt_center">02-5140-2531</td>
      </tr>
      <tr>
        <td class="listtt_center">NH농협은행</td>
        <td class="listtt_center">NH농협은행</td>
        <td class="listtt_center">1404</td>
        <td class="listtt_center">서울특별시 강남구 삼성로 155</td>
        <td class="listtt_center">02-5140-6881</td>
      </tr>
      <tr>
        <td class="listtt_center">NH농협은행</td>
        <td class="listtt_center">NH농협은행</td>
        <td class="listtt_center">1402</td>
        <td class="listtt_center">서울 서초구 강남대로 140</td>
        <td class="listtt_center">02-5140-9910</td>
      </tr>
      <tr>
        <td class="listtt_center">NH농협은행</td>
        <td class="listtt_center">NH농협은행</td>
        <td class="listtt_center">2031</td>
        <td class="listtt_center">서울 강남구 테헤란로 90</td>
        <td class="listtt_center">02-5175-0851</td>
      </tr>
      <tr>
        <td class="listtt_center">NH농협은행</td>
        <td class="listtt_center">NH농협은행</td>
        <td class="listtt_center">610</td>
        <td class="listtt_center">서울특별시 송파구 올림픽로 90</td>
        <td class="listtt_center">02-5401-9641</td>
      </tr>
      <tr>
        <td class="listtt_center">NH농협은행</td>
        <td class="listtt_center">NH농협은행</td>
        <td class="listtt_center">42</td>
        <td class="listtt_center">서울특별시 송파구 올림픽로 90</td>
        <td class="listtt_center">02-4006-9903</td>
      </tr>
      <tr>
        <td class="listtt_center">NH농협은행</td>
        <td class="listtt_center">NH농협은행</td>
        <td class="listtt_center">16</td>
        <td class="listtt_center">서울특별시 송파구 올림픽로 90</td>
        <td class="listtt_center">02-4006-6812</td>
      </tr>
    </tbody>
  </table>

```

지점정보 게시판의 DOM

3. 공간군집분석(Geo-Clustering Analysis)

웹스크래핑을 통해 얻어진 전국 농협지점 정보(지점코드, 지점명, 지점구분, 지점주소, 지점연락처)를 연구대상인 82개의 인출 지점코드에 매칭함으로써 각 인출지점에 대한 상세정보를 얻었다. 이후 이들 지점의 주소를 좌표 변환도구인 Geocoder-Xr을 활용하여 지오코딩(Geocoding)함으로써 위도와 경도 값들을 산출하였다. 마지막으로 인출지점들의 공간정보 분석을 통해 인출책의 근거지를 특정하기 위해 Sklearn.cluster() 패키지의 군집분석(Geo-Clustering Analysis) 모델링을 시행하였는데 최적의 모형을 찾기 위해 3가지 상이한 모델을 차례로 적용했다. 첫 번째로 시행한 K-Means 군집분석은 각 군집에 할당된 포인트들의 평균 좌표를 이용해 중심점을 찾아내는 기법이다. 방식이 직관적이고 해석이 용이하여 가장 대중적으로 사용되는 분석 방법이나, 연구자가 하이퍼파라미터(Hyperparameter) k(군집수)를 지정해야 하며, 노이즈(Noise)와 이상치(Outlier)에 취약하다는 제한점을 갖는다(Kanungo et al., 2002). 점과 점 사이의 거리를 측정 하기위해 최단 거리를 찾는 Euclidean Distance Method⁶⁾를 사용하였고, k값을 찾기 위해서는 비율의 한계비용(Marginal cost)이 줄어드는 최적의 지점을 나타내는 Elbow Method⁷⁾를 사용했다. 두 번째로 시행한 계층적 군집분석(Hierarchical clustering)은 나무 모양의 계층 구조를 기초로 서로 유사한 개체들을 가까운 집단부터 순차적, 계층적으로 묶어나가는 기법이다(Navarro et al., 1997). 유사한 포인트들이 결합되어 시각적으로 제공되는 덴드로그램(Dendrogram)을 통해 그룹핑 지점을 알 수 있으며, K-Means 군집분석과 달리 사전에 k를 지정하지 않아도 된다는 장점을 갖는다. 군집간의 거리 측정을 위해서는 두 개의 군집이 병합되었을 때 증가되는 변동성의 양을 나타내는 Ward Method⁸⁾를 적용하였다. 마지막으로 밀도 기반 희소 군집(Density-Based Spatial Clustering of Applications with Noise: DBSCAN) 모델을 데이터에 적용하였다(Birant & Kut, 2007). DBSCAN은 한 데이터를 eps-neighbors(하나의 데이터를 중심으로 epsilon 거리 이내의 데이터들을 하나의 군

$$6) d_e(x, y) = [\sum_{j=1}^m |x_j - y_j|^2]^{1/2}$$

$$7) \text{Ratio} = \frac{TSS - WSS}{TSS}, \text{ WSS (within cluster sum of squares)} = \sum_{j=1}^k \sum_{i \in c_j} d(x_i, c_j)^2$$

$$8) \text{Ward Distance} = \sum_{i \in A \cup B} \|x_i - m_{A \cup B}\|^2 - \sum_{i \in A} \|x_i - m_A\|^2 + \sum_{i \in B} \|x_i - m_B\|^2$$

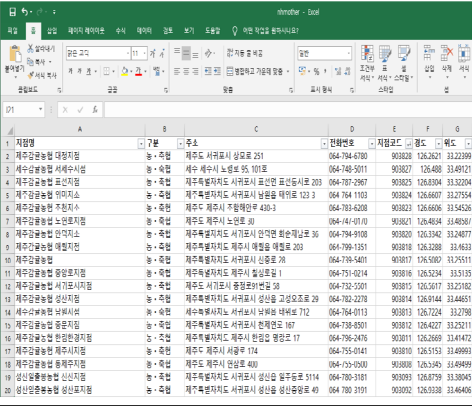
집으로 묶음)와 MinPts(한 군집은 MinPts 보다 많거나 같은 수의 데이터로 구성)를 사용하여 군집을 구성하는 방법이다. 계층적 군집 분석과 마찬가지로, 사전에 하이퍼파라미터 k 를 설정할 필요 없이 군집화 할 수 있는 기법으로 특히 다양한 모양의 군집 패턴 분석에 용이하다는 강점을 갖는다. 앞서 설명한 바와 같이 본 연구에서는 위의 세 가지 공간 군집 분석 기법을 모두 시행한 후 보이스피싱 인출책의 근거지를 보다 정확하게 추정할 수 있도록 최적의 군집화 패턴 및 중심값(Centroid)을 제공하는 모형을 선택하여 실무 수사에서 사용하도록 제안하도록 한다.

IV. 분석결과

1. 웹스크래핑을 통한 금융지점정보 자동 수집

연구 대상인 금융기관의 지점정보를 자동수집 하는 프로그램 및 해당 프로그램을 통해 얻어진 결과물의 예시가 <그림3>에 제시되어 있다. 전체 5,883개의 지점정보를 수집하였으며, 정보의 종류는 지점코드, 지점명, 지점구분, 지점주소, 지점코드, 전화번호를 포함하였다. 수집한 지점주소는 도로명 주소 형식이나, 불필요한 건물 정보가 있는 332건은 자연어처리(Natural Language Processing)의 정규표현식(Regular Expression)을 통해 일괄 삭제했다. 예를 들어 특정 지점의 주소가 서울특별시 강남구 삼성로 155 대치퍼스트 1,2층이라고 표기된 경우 “대치퍼스트 1,2층” 부분을 삭제해야 Geocoder-Xr안에서 올바른 위경도 변환이 이루어 질 수 있기 때문에 해당 부분을 자동으로 찾아내 일괄 삭제하는 코드를 작성해 시행하였다.

〈그림 3〉 금융회사 지점정보 자동 수집 프로그래밍 코드 및 수집 결과

| | |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------|
| <pre>def crawling_site(err = lambda e: print('%s' % e, datetime.now(), file = sys.stderr)): results = [] browser = webdriver.Chrome(r"C:\Users\user\data\chromedriver") browser.get('http://nonhyup.tritops.co.kr/list_branch.jsp') time.sleep(3) h1 = browser.find_element_by_css_selector('h1') browser.find_element_by_css_selector(h1).click() time.sleep(3) c1 = browser.find_element_by_css_selector('div.tabM > ul > li.li_R > a') browser.find_element_by_css_selector(c1).click() # 1 through 17 time.sleep(3) c2 = browser.find_element_by_css_selector('div.tabM > ul > li.li_R > a') browser.find_element_by_css_selector(c2).click() for page in count(start = 1): script = 'getPage(%d)' % page browser.execute_script(script) time.sleep(1) html = browser.page_source try: bs = BeautifulSoup(html, 'html.parser') tag_body = bs.find('div', attrs={'class': 'boardList'}) tags_tr = tag_body.findAll('tr') if len(tags_tr) == 1: break for tag_tr in tags_tr: strings = list(tag_tr.strings) print(strings) store = strings[2] btype = strings[4] code = strings[5] address = strings[7] phone = strings[8] results.append((store, btype, code, address, phone)) except AttributeError as e: err(e) except AttributeError as e: err(e)</pre> |  |
| 웹스크래핑 프로그래밍 코드 | 영업지점 정보 수집 결과물 |

이후 정제된 지점 주소를 바탕으로 지오코딩한 결과값인 지점의 위도와 경도 정보가 추가 되었다. 최종적으로 분석대상인 82개의 지점코드와 웹스크래핑을 통해 수집된 지점정보를 매칭(Matching)함으로서, 기존의 농협 사이트에서 전별 검색을 하는 수작업 방식 대신 자동으로 농협 지점정보와 좌표를 일괄 확인하는 프로그램을 완성하였다. 〈그림 4〉는 개발된 프로그램을 통해 금융기관 사기계좌 입출금 내역 회신자료의 지점코드와 웹스크래핑으로 수집한 지점정보를 매칭한 결과물의 예시이다.

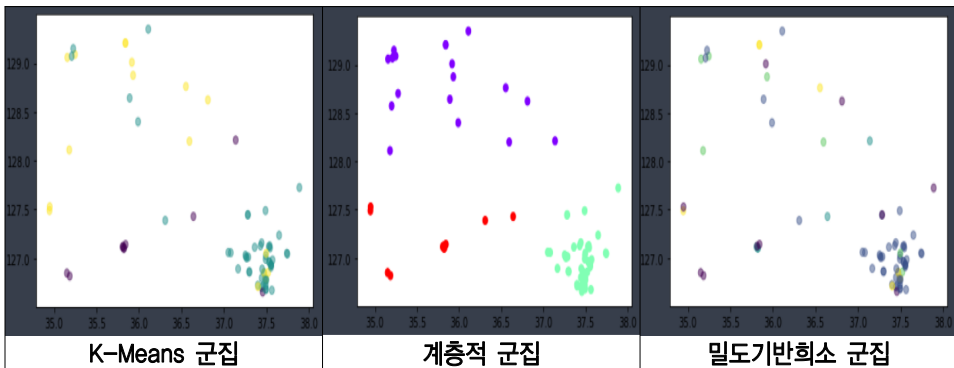
〈그림 4〉 사기계좌 입출금 내역 기록의 지점코드와 지점정보 매칭 결과물

| 지역농축협 금융거래 내역 | | | | | | | |
|---------------|----------|-----------|------------|----|--------|-----------------|---------------------|
| 2020년 03월 02일 | | | | | | | |
| 거래일자 | 거래시간 | 거래금액<액> | 잔액<원> | 구분 | 거래점 | 지점명 | 주소 |
| 2020-01-01 | 3:04:24 | 3,800,000 | 11,429,130 | 입금 | 737031 | 감천농협 부곡지점 | 경북 김천시 김천로 9 102 |
| 2020-01-01 | 8:02:17 | 410,500 | 11,018,630 | 지급 | 000303 | 청주시 지부 농협은행 | 충청북도 청주시 상당구 중앙로 19 |
| 2020-01-01 | 14:53:23 | 1,100,000 | 9,918,630 | 지급 | 001459 | NH금융PLUS 광화문역센터 | 서울 종로구 세종대로 149 |
| | | | | | | | 연락처 |
| | | | | | | | 054-432-6048 |
| | | | | | | | 043-222-5101 |
| | | | | | | | 02-3210-2531 |

2. 군집분석을 통한 전화금융사기 인출액 근거지 추정 모델

사기계좌의 인출 지점을 공간 시각화하고 이들을 근거지로 추정되는 중심점을 추정하는 최적의 모델 개발을 위해 3가지 유형의 공간 군집분석을 차례로 시행하였다. 먼저 k-means 군집분석의 시행에 있어 가장 적절한 군집 수 k를 찾는 Elbow method를 적용한 결과, 3 그룹이 가장 적절한 것으로 나타나⁹⁾ k값을 3으로 지정한 후 군집을 구성하였다. 계층적 군집분석의 경우, k-means 군집분석과 달리 k(군집 수)를 사전에 지정할 필요는 없으나, 유사한 개체들이 결합되는 모양을 나타내는 덴드로그램을 통해 시각화 했을 때 <그림 6>에서와 같이 3 그룹이 최적인 것으로 나타났다. 밀도 기반 희소 군집분석은 이와는 상반되게 5가지 유형의 군집 분류가 최적인 것으로 나타났다. 하단의 <그림 5>는 각 모형의 군집 분류 결과물을 Seaborn() 함수를 기반으로 위·경도 좌표 상에서 시각화한 것이다. 비교 결과, 계층적 군집, k-means 군집, 그리고 밀도 기반 희소 군집 순으로 정확하게 공간 군집이 이루어진 것을 확인 할 수 있다.

〈그림 5〉 모형별 군집분석 분류 결과 값의 공간(좌표 상) 시각화

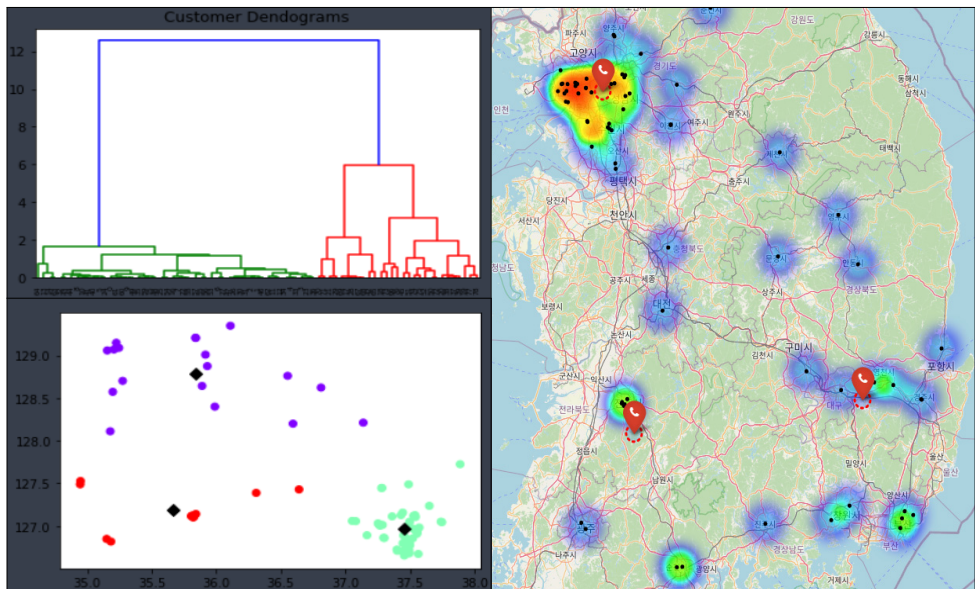


검증한 공간 군집분석 기법 중 유사도 행렬 계산을 통해 거리가 인접한 관측치들(인출이 이루어진 은행 지점)끼리의 군집을 형성해 주는 계층적 군집이 가장 명확한 것으로 나

9) 비율의 한계비용이 줄어드는 지점을 표기하는 WSS(Within cluster sum of square)값이 k가 1일 때 113.20, 2일 때 33.06, 3일 때 15.50, 4일 때 10.96인 것으로 나타나 감소폭이 급격하게 줄어드는 3을 K로 지정했다.

타나, 계층적 군집을 통해 분류된 공간 그룹들의 중심점(Centroid)들을 Folium의 HeatMap()과 plugins() 함수를 통해 HTML 형식으로 공간 시각화 한 결과는 <그림6>과 같다. 그림의 좌측은 그룹 분류의 기준이 된 덴드로그램과 그룹핑 공간 시각화 결과물 및 중심점(검은 다이아모양 도형)을 나타낸다. 그림의 우측은 사기계좌 인출지점들을 실제 지도에 표기하고 그 밀집지를 핫스팟(Hotspot)의 형태로 레이어링 한 후 인출책의 근거지로 추정되는 중심점¹⁰⁾들을 전화기 모양의 아이콘으로 표기한 최종 결과물을 나타낸다. 연구 대상이 된 보이스피싱 사건의 인출지점들이 크게 3개의 지역(서울 인근 지역, 대구 인근 지역, 전주 인근 지역)에 분포하고 있으며 인출책의 근거지로 추정되는 중심점이 각각 3군데 존재한다는 것을 하단의 지도를 통해 확인 할 수 있다.

<그림 6> 사기계좌 인출지점의 군집 및 인출책 근거지 추정 지도



10) 계층적 군집분석 결과 분류된 3 군집의 중심점 위·경도는 각각[35.839802, 128.790626], [37.456147, 126.966547], [35.663468, 127.190427] 이다.

V. 결론

본 연구는 빅데이터 분석기법을 활용하여 전화금융사기 실무 수사에서 활용될 수 있는 분석툴(Tool)의 개발에 관한 연구방법론을 제안하였다. 구체적으로 온라인상에 공개되어 있는 금융회사의 지점정보를 웹스크래핑 기술을 통해 자동 수집하여 지점코드를 지점정보(지점명, 지점주소, 지점 연락처 등)와 매칭하고, 변환된 은행 지점의 주소지를 좌표정보로 변환(Geocoding)하여 지도상에 시각화하고, 최종적으로 이를 다시 공간 클러스터링화(Hierarchical Clustering) 함으로서 해당 지점들의 중심지를 파악하여 인출책의 근거지 추정에 사용될 수 있는 정보를 도출하였다. 이러한 수사기법의 개발은 현장에서 보다 신속하게 보이스포싱 수사 정보를 분석하고 활용하는데 도움을 주어 민생 치안 안정에 보탬이 될 뿐만 아니라, 최근 다양한 방식으로 개발·확산되고 있는 빅데이터 분석 기술이 현장에서 어떻게 적용될 수 있는지에 대한 구체적인 사례로 이용되는 등 다양한 학술적 함의를 갖는다.

다만 보이스포싱 인출책의 신속한 검거라는 실무상의 목적 달성의 위해 몇 가지 보완 또는 확장 개발 되어야 하는 사항들이 분명히 존재한다. 첫 번째로, 본 연구에서는 연구방법론의 제안이라는 목적에 맞게 구체적인 웹스크래핑 과정을 상세 설명하기 위해 특정 금융회사의 웹사이트를 지정하여 분석을 시행하였으나, 실무 적용을 위해서는 최소 10개 이상의 주요 금융회사의 지점정보가 같은 방식으로 수집되어야 한다. 두 번째로, 효과적인 인출책의 도주경로 추적을 위해 인출지점 근처의 CCTV 위치가 지도상에 함께 표기되어야 한다. 예방적 활동이 아닌, 인출이 이미 이루어진 사후에 수사가 이루어질 경우 인출책의 이동 경로에 대한 영상증거를 확보하는 일이 필수적이기 때문에 코드 조건 값으로 인출 지점 근처 반경 1 km 이내 설치되어 있는 CCTV의 위치주소, 설치목적, 관리주체, 기록보관기간 등의 정보가 동시 제공되도록 프로그램이 확장될 필요가 있다. 같은 맥락에서 도주경로 추적을 위해 지하철 및 버스 노선에 대한 정보 또한 추가 되어야 할 것이다. 시간에 따른 인출책의 이동경로 추적을 위해 인출 지점의 좌표를 인출시간을 기준으로 동영상 형태로 순차적으로 시각화하는 기능 또한 함께 활용되는 것이 효과적일 것으로 예상된다. 마지막으로 해당 프로그램은 효율성을 위해 플랫폼에 독립적이며 인터프리터식 대화형 언어인 파이썬(Python) 프로그래밍을 기반으로 이루어졌으나, 해당 언어는 기반지식(공학, 수학, 프로그래밍 등)이 있는 전문가들이 주로 사용하는 것으로, 현장의 실무 수사관이 사용하기 위해서는 보다 직관적인 형태의 인터페이스를 갖춘 별도의 소프트웨어가 추가 개

발 되어야 한다. 이러한 제한점을 개선한 후속 연구의 진행이 활발히 이루어져야 할 것이며, 그 시작점이 된 본 연구가 지역사회 시민들의 안전을 위해 많은 노력을 하고 있는 현장 수사관들에게 작게나마 도움이 되는 것을 기대해 본다.

《참 고 문 헌》

- 김 절·김병선. (2009). "토지용도별 범죄의 시· 공간적 분포패턴 사례연구". 「한국도시지리학 회지」, 12(3): 83-96.
- 김주환. (2014). "탐색적 공간데이터 분석을 통한 CPTED장비 위치설계 방안에 대한 연구". 박사학위논문, 한성대학교 대학원.
- 김혜진. (2020). "뉴스 비정형 데이터의 수집과 토픽 분석 (LDA) 을 통한 성범죄 치안 이슈의 효율적 탐색". 「한국범죄학」, 14(1): 5-20
- 노기윤·이창배. (2018). "접합 분석(Conjunctive Analysis)을 적용한 폭력범죄발생의 지리적 프로파일링". 「한국공안행정학회」, 27(4): 227-248.
- 배영호·이금녀. (2013). "구글 검색을 통한 불법정보 유통차단 기법". 「디지털포렌식연구」, 7(1): 77-92.
- 이건학·진찬우·김지우·김원희. (2016). "성폭력 범죄의 공간적 분포 특성에 관한 연구: 환경범죄학에 기반한 공간 분석". 「대한지리학회지」, 51(6): 853-871.
- 신지용·조지호·이 한·김정민. (2016). "이미지를 이용한 웹사이트 위· 변조 탐지 기법 연구". 「융합보안논문지」, 16(1): 81-87.
- 이문국. (2015). "지리적 프로파일링 시스템 (GeoPros)을 활용한 범인 거점 예측: 연쇄 강도강간범을 중심으로". 「한국웹테드학회지」, 6(1): 105-124.
- 장종욱·정준호·손윤식. (2018). "웹 크롤러를 활용한 범죄 정보 수집 시스템 연구." KIIT 학회 발표.
- 조정운·심수진·서한별·이민식. (2018) "공간분석을 활용한 서울시 무단횡단 사고분석과 대응방안". 「한국웹테드학회지」, 19(1): 264-296.
- 조주연·조경원. (2018). "텍스트 마이닝을 이용한 청소년 문제 토픽 모델링". 「한국정보통신학회논문지」, 22(12): 1589-1595.
- 황초희·서용철. (2008). "범죄 GIS를 통한 지구대의 위치분석". 「한국방재학」, 13(2): 357-360.
- 김은선·조윤식. (2020). "잠재디리슬레할당 기반 군집화를 통한 유사 범죄코드 발굴과 범죄예측". 「정보과학회논문지」, 47(1): 45-51.

Birant, D., and Kut, A. (2007). "ST-DBSCAN: An algorithm for clustering spatial - temporal data. Data knowledge engineering." 「Data & Knowledge Engineering」, 61(3-4): 264-282.

60(1): 208-221.

Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., and Silverman, R., Wu.
(2002). "An efficient k-means clustering algorithm: Analysis and implementation,"
「IEEE Transactions on Pattern Analysis and Machine Intelligence」, 24(7):
881-892.

Navarro, J. F., Frenk, C. S., and White, S. J. (1997). "A universal density profile from
hierarchical clustering," 「Astrophys」, 490(2): 493.

Abstract

Phone Scam: Developing an Investigative Technique Through Web Scraping and Geo-Clustering Analysis

Kim, Hye-Jin*

This study aims to develop an investigative technique that can plausibly be utilized in the investigation of criminal phone fraud (A.K.A Voice Phishing) based on web scraping and spatial clustering methods. The web scraping is designed to mass collect the open-source bank branch information as criminal investigators need to convert bank branch codes into written branch information(i.e., branch name, branch address, branch contact information, etc.) after executing the seizure and search warrant. Hierarchical clustering algorithm is generated to identify the center of the corresponding points of the bank branch, in which represents the criminal intelligence for possible homeground(i.e., house, workplace) of the phone scammer. It is expected that this research would be recognized as a concrete evidence on how big data analytics is efficiently employed in the field of Criminal Justice.

Key words : Phone Scam, Voice Phishing, Web Scraping, Geo-Clustering, Investigative Technique, Big data Analytics

논문 접수일 : 2020년 8월 21일
심사 완료일 : 2020년 9월 10일
게재 확정일 : 2020년 9월 12일

* Researcher, Police Science Institute at Korean National Police University