ISSN : 1976-5304
https://kdfs.jams.or.kr

구글 보관 사용자 데이터의 시각화 방안

김 현 우*, 이 상 진**
경찰청 (수사관)*. 고려대학교 정보보호대학원 (교수)**

Visualization of User Behavior With Google User Data

Hyun-Woo Kim*, Sang-Jin Lee**
Korean National Police Agency (Inspector)*, Korea University (Professor)**

요 약

각종 스마트기기의 등장과 사람들이 사용하는 디바이스가 증가함에 따라 디지털 환경이 변화하고 포렌식에 대한 인식 확대, 안티포렌식 기법의 발전으로 수사기관에서 전자정보를 확보하고 분석하는 것이 어려워졌다. 한편 IT 기업들이 사업을 확장함에 따라 수집하는 사용자 데이터가 광범위해졌는데 그 중 특히 구글은 다양한 플랫폼을 통해 수사에 활용할 수 있는 데이터들을 수집하고 있다. 그러나 구글에서 획득한 데이터는 사람이 인식할 수 없거나 양이 많기 때문에 적절한 가공 없이는 수사에 활용하는데 한계가 있다. 이에 위치정보, 어플리케이션 사용내역, 검색기록 등 구글에서 보관하고 있는 다양한 데이터들로부터 수사에 활용할 수 있는 유의미한 정보를 추출하기 위한 시각화 방안을 제시하고자 한다.

주제어 : 디지털 포렌식, 구글 사용자 데이터, 시각화, 파이썬

ABSTRACT

Digital environments have changed because various kinds of smart devices have emerged and the number of devices that people use has increased. And awareness of digital forensics has enlarged and techniques of anti-forensics have been developed. As a result, it became difficult for law enforcement agencies to obtain and analyze digital data. Meanwhile, data that IT companies collect from users have been expanded following the growth of the industry. In particular, Google collected data that can be used for investigation through various platforms. However, because some data that Google collected is unrecognizable for people and the amount of data is enormous, there is a limit to use it without proper processing. Therefore this study introduces methods to visualize data that Google collected from users such as location history, application usage or search keyword logs to utilize them for investigation.

Key Words : Digital Forensics, Google User Data, Visualization, Python

Ⅰ. 서 론

사람들은 일상생활에서 여러 디지털 기기를 사용한다. 가정에서는 노트북이나 태블릿을 사용하며, 회사에서는 회사에서 지급한 컴퓨터와 노트북을 사용하고, 항상 스마트폰을 소지하고 다닌다. '시스코 비주얼 네트워킹 인덱스 2017~2022년 전망 및 추세'에서는 2022년까지 1인당 디바이스 및 회선 평균 개수가 전세계적으로 미국(13.6개)에 이어 한국이 11.8개로 2위를 차지할 것으로 전망하고 있다[1].

한편 디지털 포렌식도 변화하는 디지털 사용 환경에 맞춰 발전하였는데 과거에는 PC 기반의 디지털 저장 매체나 운영체제, 파일시스템에 대한 연구가 활발했던 반면 모바일 기기의 사용이 증가함에 따라 모바일 기기에 대한 포렌식 연구가 급증하였다. 최근에는 IoT의 등장으로 각종 IoT 기기 내에서 수집할 수 있는 디지털 증거에 대한 연구가 진행 중에 있다[2].

2007년부터 2016년까지의 디지털 포렌식 연구동향을 보면 대상매체를 기준으로 운영체제 등 시스템 포렌식, 물리적 저장 매체와 같은 디스크 포렌식, 휴대용 기기를 포함한 모바일 포렌식 연구가 과반 이상으로 나타났다[3]. 그러나 포렌식에 대한 사회적 인식의 증가, 안티포렌식 기술 발달 등으로 물리적 저장 매체나 휴대용 기기에서 직접 데이터를 수집하는 것이

▪ Received 14 November 2019, Revised 25 November 2019, Accepted 17 December 2019
▪ 제1저자(First Author) : Hyun-Woo Kim (Email : nostalin@korea.ac.kr)
▪ 교신저자(Corresponding Author) : Sang-Jin Lee (Email : sangjin@korea.ac.kr)

어려워지고 있다. 이에 따라 기기 중심의 포렌식을 보완할 수 있는 수단으로서 사용자 계정 중심의 포렌식이 대두되고 있다. 사용자 계정 중심의 포렌식이란 대상자가 사용하는 단말기 등 물리적 기기가 아닌, 서비스를 이용하기 위하여 등록한 계정을 포렌식 대상으로 하는 것을 말하며, 한 사용자가 PC나 모바일과 같은 여러 기기를 사용하더라도 계정은 하나를 공통적으로 사용한다는 점에 착안하여 계정에 저장되는 유의미한 디지털 증거를 확보하는 포렌식 방법이다.

특히 이전에는 IT 기업들이 수집하는 사용자 정보가 한정적이었으나 산업이 고도화됨에 따라 구글이나 애플과 같은 대규모 IT 기업들이 수집하는 사용자 정보가 광범위해졌다. 이 중에서도 구글은 검색 엔진뿐만 아니라 모바일 운영체제(안드로이드), 브라우저(크롬), 메일 서비스(Gmail), 동영상 플랫폼(Youtube), 지도(Maps) 등 다양한 서비스를 제공하고 있다. 2019년 9월 기준으로 스탯카운터(statcounter)에 따르면 한국 모바일 OS 시장점유율은 [그림 1]과 같이 안드로이드가 73.85%로 1위, 브라우저 점유율은 크롬이 54.51%로 1위를 차지하고 있는 등 그 점유율도 매우 높다. 이에 구글에서 보관하는 사용자 데이터를 수사에 활용하는 아이디어를 제시하고, 구글 데이터의 각 아티팩트가 수사상 갖는 의미와 실제 사례를 소개한 연구가 있다[4].

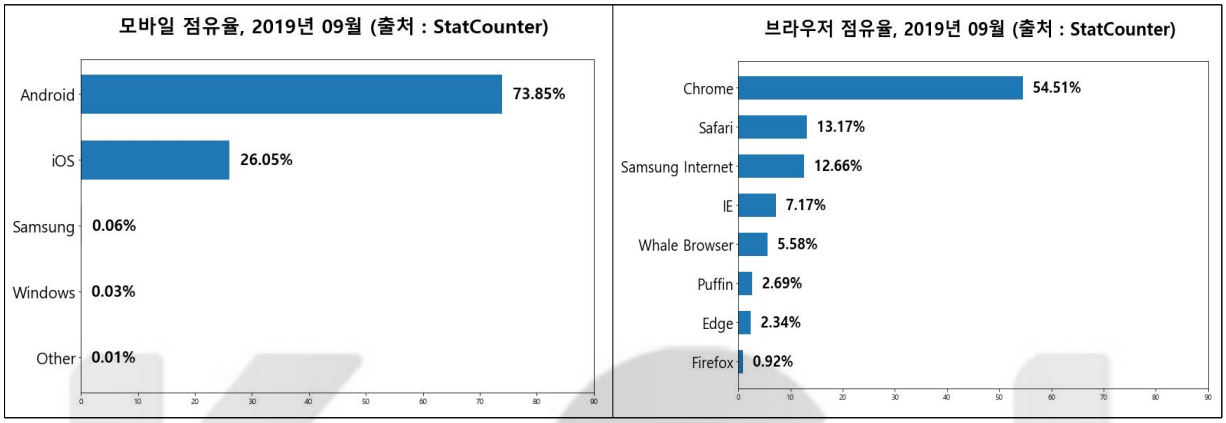

그림 1. 2019년 9월 기준 한국 모바일 운영체제 및 인터넷 브라우저 점유율
Figure 1. Mobile OS and Browser Market Share in Korea - Sep 2019

본 논문에서는 구글에서 보관하는 사용자 데이터를 수사 현장에서 유용하게 사용할 수 있도록 시각화하는 방안을 제시하고자 한다. 구글에서 보관하는 사용자 데이터 중 시각화가 가능한 데이터 종류를 확인하고, 각 데이터들을 어떠한 방식으로 시각화했을 때 수사에 도움이 될 수 있는지 살펴보고자 한다.

Ⅱ. 구글에서 보관하는 사용자 데이터 시각화의 의의

구글에서 보관하는 사용자 데이터에는 안드로이드 기기 정보, 크롬 브라우저 사용 정보, 구글 블로그 이용 내역, 구글 검색 기록, 구글 플레이스토어 설치 이력, 유튜브 사용 내역, 드라이브 내 저장된 파일, 메일함, 주소록, 위치기록 등 여러 종류가 있다. 또한 구글에서는 사용자 데이터를 json, csv, html 등 다양한 파일 형태로 추출해주는 Takeout 서비스를 제공하고 있기 때문에 이를 비교적 쉽게 획득하여 수사에 활용할 수 있다[4].

다만 인간의 인지능력에는 한계가 있기 때문에 산재해있는 데이터들 중에서 수사에 활용할 수 있는 유의미한 정보를 추출하기 위해 많은 노력을 기울여야 한다. 특히 구글에서 수집한 사용자 데이터는 그 자체로는 어떠한 의미를 갖지 않는 단편적인 사실의 나열이므로, 데이터를 가공하여 목적에 맞게 의미를 도출하여 유의미한 정보를 추출하는 것이 매우 중요하다[5]. 즉, 방대한 구글 사용자 데이터 중에서 수사관이 수사 목적에 따라 필요한 정보를 추출해내는 작업은 수사의 성패를 좌우한다고 할 수 있다.

한편 구글에서 보관하는 데이터는 위치기록, 브라우저 사용 기록, 활동내역 등에 대해 사용자가 구글 계정으로 로그인한 모바일, 브라우저 등 다양한 플랫폼에서 수집된다. 이처럼 구글에서 수집하는 사용자 정보는 규모가 크고 종류가 여러 가지이며 그 데이터 중에는 직관적으로 이해할 수 없는 위경도, 타임스탬프(timestamp)와 같이 사람이 이해하기 위해 변환이 필요한 데이터도 존재한다. 그 결과 수사관이 모든 데이터를 하나씩 직접 보고 이해하는 것은 한계가 있으며, 이 속에서 수사와 관련된 유의미한 정보를 찾아내는 것은 더 어렵다. 따라서 이를 수사 목적에 따라 효율적으로 보여줄 수 있는 시각화 방법론을 제시하는 것은 의미가 있다[6].

Ⅲ. 시각화 관련 선행연구

정보보안과 디지털 포렌식 분야에서는 시각화 대상과 목적에 따라 보안 로그로 위험을 탐지하거나 수집된 증거를 활용하여 수사단서를 찾기 위하여 로그나 증거들을 시각화한 연구들이 있다.

보안로그 시각화는 주로 보안장비에서 수집되는 로그들을 바탕으로 침입 등 보안 이슈를 예측하거나 분석하는 방법 중 하나로 제시되었으며, RGB Palette 도구를 이용하여 색상, 크기 등으로 보안 로그의 중요 요소인 출발지·목적지 IP, 목적지 포트, 로그 발생량을 표현하고 Port Scanning, DDoS Attack과 같은 보안 이슈에 따른 패턴을 탐지하는 시각화 방안을 제시한 연구가 있다[7]. 또한 오픈소스인 Elastic Stack을 이용하여 네트워크 보안 관제 시각화 방안을 제시하고자 한 연구가 있었으며, 구체적으로 침입 탐지 시스템(IDS) 중 하나인 오픈소스 Bro의 수집 데이터를 바 차트, 파이 차트, 히트맵 등으로 표현한 보안 관제 대시보드를 제시하기도 하였다[8].

또한 수집된 증거를 시각화하여 범죄수사에 활용할 수 있는 방법에 대한 연구로는 사회연결망 분석원리를 이용하여 통화기록과 거래내역으로 대상자들을 연결 짓고, 이를 정보분석 프로그램 i2(information image)를 활용하여 범인을 특정하거나 공범을 찾아내는 방법을 제시한 연구가 있다[9].

이 외에 지리적 데이터 시각화와 관련하여 구글이나 네이버, 다음과 같은 포털사이트에서 자체 지도 API를 제공하고 있으며, 정부나 지자체, 공공기관에서는 정책목적으로 수집한 데이터들을 민간에 공개하여 활용하도록 장려하고 있다. 이에 구글 지도 API를 이용하여 서울시 등에서 제공하는 공공데이터를 시각화하는 연구도 있다[10].

Ⅳ. 연구방법 및 시각화 방안

1. 연구방법

시각화를 위한 개발환경은 Python 3.7.1(Anaconda 4.7.5)을 이용하였으며, 주된 라이브러리로는 데이터 처리를 위한 Numpy, Pandas, 지도 데이터 표현을 위한 Folium, 그래프 시각화를 위한 Matplotlib, 문자열 빈도수를 표현하기 위한 Word Cloud를 사용하였다. 시각화 대상은 2013. 12. 31.부터 2019. 8. 15.까지 Android 운영체제가 설치된 모바일 기기와 Windows가 설치된 PC와 노트북에서 크롬 브라우저와 구글 검색엔진 등을 이용하여 직접 수집한 구글 사용자 데이터이다.

앞에서 설명했듯이 구글에서 보관하는 사용자 데이터는 그 종류가 여러 가지이다. 이 중 수사의 활용도가 높으며 특히 그 데이터가 방대하고 사람이 식별하기 어려운 정보들을 보관하고 있는 정보를 표 1과 같이 선별하였다. 각 데이터는 시간, 공간 그리고 활동량을 중심으로 시각화가 가능한 대상들이다. 이 데이터를 이용하여 시간의 흐름에 따른 공간이나 활동량의 변화를 시각화할 수 있다. 또한 수사 활용도가 높은 텍스트 기반의 데이터인 검색어도 시각화 대상으로 선정하였다.

표 1. 구글에서 보관하는 사용자 데이터 중 선정한 시각화 대상
Table 1. Visualized Data Selected from Google User Data

구분	위치	데이터
위치 기록	- Takeout\위치 기록	- 타임스탬프(Timestamp) - 위도 - 경도
안드로이드 활동내역	- Takeout\내 활동\Android	- 시간 - 사용한 어플리케이션
검색기록	- Takeout\Chrome - Takeout\내 활동\검색	- 시간 - 검색한 단어

2. 시각화 방안

우선 사용자의 위치기록을 수사에 활용할 수 있는 시각화 방안을 제시하고자 한다. 위치기록을 이용하여 대상자가 범죄현장과 같은 특정 장소를 방문한 사실이 있는지를 확인하는 시각화 방안을 소개하고자 한다. 이는 범행시각 특정, 대상자의 알리바이 입증 등 수사에 활용할 수 있는 단서를 확보하는 수단이 될 수 있다. 또한 특정 시간 동안 대상자의 방문장소를 확인하고 예상되는 이동 경로를 추정해보는 시각화 방안을 제시하고자 한다. 이를 통해 구글의 위치기록에서는 확인되지 않지만 대상자가 경유한 장소를 추정해볼 수 있다. 또한 하루 단위로 대상자의 위치를 비교하면 대상자의 행동패턴을 확인할 수 있다. 대상자가 특정 장소에 머무르는 시간 등을 확인한다면 대상자의 거주지, 사무실 위치 등을 알아낼 수 있으며, 더 나아가 평소와 다른 특이한 패턴을 보이는 것도 감지해낼 수 있다.

다음으로 대상자의 안드로이드 어플리케이션 사용기록을 활용하여 대상자가 주로 사용하는 어플리케이션과 시간을 확인하는 시각화 방안을 제시하고자 한다. 그리고 특정 어플리케이션이 범죄와 관련되었다고 할 때, 시간 흐름에 따른 어플리케이션의 사용 시간을 보여주는 타임라인을 이용하여 수사에 활용할 수 있는 방안을 제안하고자 한다. 또한 어플리케이션 사용 내역을 통해 대상자의 생활패턴을 분석하는 방안을 제시하고자 한다.

또한 검색어와 관련하여 대상자의 관심사, 감정상태 등을 확인하기 위한 방안으로 검색어 빈도수에 따른 시각화 방안, 시

간의 흐름에 따른 검색어 변화 등을 분석하도록 한다.

마지막으로 시각화 결과들을 도출한 소스코드는 모두 Github에 공개하였으며, 각 시각화 결과가 도출된 과정은 GIthub 페이지(https://github.com/briron/google-user-data-visualization)를 통해 확인할 수 있다.

2.1. 위치정보를 이용한 시공간 데이터 시각화

구글 위치정보는 "Takeout\위치 기록" 경로에 "위치 기록.json"이라는 JSON(Javascript Object Notation) 형태로 저장된다. 구체적으로 "위치 기록.json" 파일에는 "locations"라는 키(key)에 배열 형태로 각 위치정보가 저장된다. 개별 위치정보에는 "timestampMS", "latitudeE7", "longitudeE7", "accuracy", "activity" 등 키가 있는데, 각 키의 의미는 표 2와 같다.

표 2. '위치 기록.json' 파일에 저장된 값들의 의미
Table 2. Meaning of Values Stored in 'LocationHistory.json' File

키	의미	예시
timestampMS	측정된 시간을 밀리초을 밀리초(millisecond, 천 분의 1초) 단위의 유닉스 시간으로 표현한 것	1388481387673
latitudeE7	측정된 위도값에 10^7을 곱한 값	349714321
longitudeE7	측정된 경도값에 10^7을 곱한 값	1267069866
accuracy	측정된 값의 정확도 (낮은 값일수록 높은 정확도를 의미)	14
activity	구글에서 추정하고 있는 사용자의 움직임 (차로 이동하는지, 걷고 있는지 판단)	[{ "type" : "STILL", "confidence" : 100 }]

위 데이터 중 실제 사용자가 특정 시간에 있었던 위치를 확인하기 위해서는 시간과 위도, 경도의 세 가지 데이터를 추출할 필요가 있다. 이에 해당 데이터를 담고 있는 "timestampMS", "latitudeE7", "longitudeE7" 값을 각각 한국시간(UTC+9), 위도와 경도로 변환하였으며, 그 결과를 확인하면 [그림 2]와 같다.

	latitude	longitude	datetime
126768	37.564192	126.966817	2019-08-15 01:17:46.813000+09:00
126769	37.564192	126.966817	2019-08-15 01:19:47.422000+09:00
126770	37.564192	126.966817	2019-08-15 01:21:47.452000+09:00
126771	37.564192	126.966817	2019-08-15 01:23:47.465000+09:00
126772	37.562445	126.966231	2019-08-15 01:25:16.221000+09:00
126773	37.563025	126.966579	2019-08-15 01:25:40.469000+09:00

그림 2. '위치 기록.json'에 저장된 데이터를 추출하여 변환한 결과
Figure 2. Result of Extracting and Converting Data Stored in 'LocationHistory.json' file

위치기록은 특정 시간에 대상자의 위치(위도, 경도)를 알 수 있다는 점에서 유용하다. 따라서 위치기록의 시각화는 범행장소와 관련하여 대상자의 범행장소 방문여부와 예상 경로를 추정할 수 있도록 고안하였다. 또한 하루 단위나 특정 시간대의 대상자의 위치를 표시함으로써 대상자의 행동패턴을 확인할 수 있는 시각화 방안을 연구하였다.

2.1.1. 특정장소 방문 확인을 위한 시각화

용의자와 범죄장소는 확인되었지만 범행 시간을 특정하지 못한 경우나 여러 용의자 중 실제 용의자를 찾아야 하는 상황에서 이들 중 범죄현장에 단 한 번도 가본 적이 없는 용의자를 가려낼 필요가 있다. 위도와 경도는 그 데이터 자체로는 사람이 인식하기 어렵고 지도에 표시되어야 의미를 알 수 있기 때문에 지도를 이용하여 시각화하였다.

구체적으로 지도상에서 특정 장소를 표시해놓고 대상자가 있었던 위치들 중 특정 장소와 가장 가까웠던 위치들을 지도상

에 표시하였다. 특정 장소와 가장 가까웠던 위치의 데이터를 구하기 위하여 대상자의 위치들과 특정 장소 사이의 거리를 각각 계산한 이후, 계산된 거리들 중 가장 작은 10개의 데이터만 추출하였다.

〔그림 3〕은 특정 장소를 방문한 적이 있는 경우와 방문한 적이 없는 경우에 지도상에 표시되는 결과를 비교한 것이다. 아이콘은 확인하고자 하는 특정 장소를 표시하고 있으며, 원형 점은 실제 대상자가 특정 장소와 가장 가까운 곳에 있었던 위치를 표시하고 있다. 특정 장소를 방문한 적이 있는 왼쪽 그림과 같은 경우 아이콘과 매우 가깝게 대상자의 위치가 표시된 것을 확인할 수 있는 반면, 방문한 적이 없는 오른쪽 그림의 경우 아이콘과 가까운 위치에는 대상자의 위치가 표시되지 않고 먼 곳에서 원형 점이 표시되는 것을 확인할 수 있다.

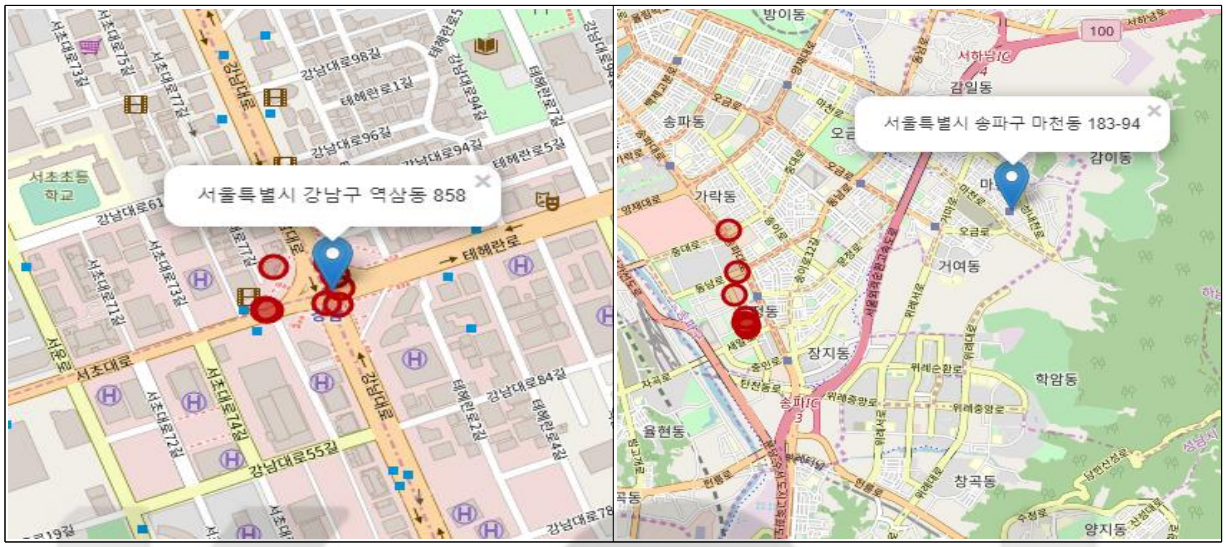

그림 3. 특정 장소를 방문한 적이 있는 경우(왼쪽)와 없는 경우(오른쪽) 시각화 결과
Figure 3. Visualization Results if Visited a Specific Location (Left) and if Not (Right)

2.1.2. 일정 기간 위치 이동 변화 시각화

다음으로는 위와 반대로 시간을 중심으로 대상자의 장소 변화를 시각화하였다. 범행시간이 특정되었다고 할 때, 〔그림 4〕는 특정기간 동안 대상자의 위치를 표시한 것으로, 점으로 표시된 것이 대상자의 위치를 의미한다.

가령 살인 사건의 경우 피의자가 살인 이후 사체를 은닉하였으나 시신이 발견되지 않은 경우 살인 직후 피의자의 위치를 확인한다면 은닉한 사체의 위치를 추론해볼 수 있다. 또한 피의자가 물건을 절도한 이후 장물을 매각/은닉한 경우에도 범행 시각 이후 피의자의 위치를 확인하면 매각하거나 은닉한 장소를 확인할 수 있다.

〔그림 3〕은 장소를 기준으로 지도상에 대상자의 위치를 표시했다면, 〔그림 4〕는 시간을 중심으로 지도상에 대상자의 위치를 시각화하였다는 점에서 그 차이가 있다.

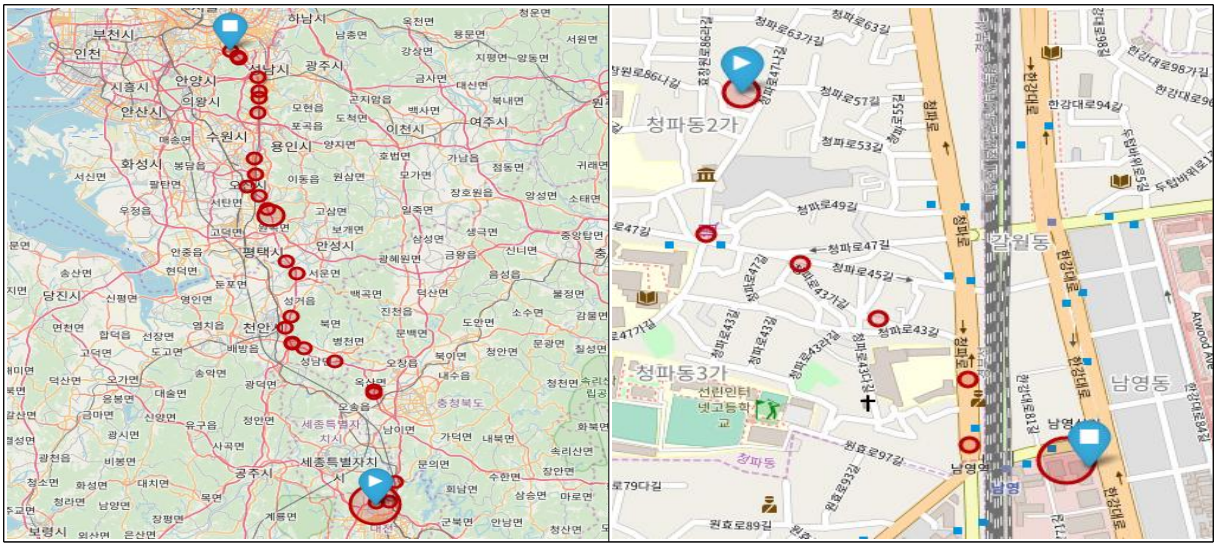

그림 4. 일정 기간 동안 위치 이동을 보여주는 시각화 결과
Figure 4. Visualization Results Showing Position Movement over a Period of Time

2.1.3. 예상 이동경로 시각화

〔그림 4〕와 같이 범행 전후 대상자의 동선을 확인하기 위하여 시간 흐름에 따라 대상자의 위치를 지도상에 표시하는 것 외에도 대상자가 표시된 위치 사이에 이동한 이동경로를 확인할 필요가 있다. 이를 위해 대상자가 표시된 위치 사이를 단순히 직선으로 표현한다면 통과할 수 없는 건물이나 산과 같은 지형물을 통과하는 것처럼 표현되기 때문에 실제 이동경로와 상당한 차이가 발생한다. 따라서 실제와 가장 가까운 경로를 예상하기 위해서는 실제 사람이 이동할 수 있는 경로를 표시할 필요가 있다.

이를 위해 〔그림 5〕와 같이 네비게이션 API를 활용하여 대상자의 출발지와 목적지 사이에 예상 경로를 확인하여 지도상에 표시하였다. 원형의 점은 〔그림 4〕와 같이 실제 대상자가 있었던 위치이며, 점 사이를 잇는 선은 네비게이션을 이용하여 대상자의 예상 이동경로를 표시한 것이다.

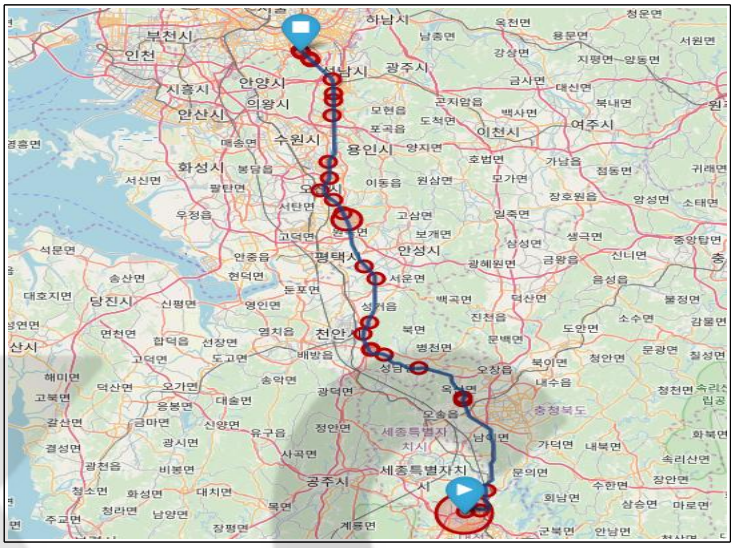

그림 5. 일정 기간 동안 위치 이동을 보여주는 시각화 결과 (차로 이동)
Figure 5. Visualization Result Showing Position Movement over a Period of Time (By Car)

다만 네비게이션에서 제공하는 최적 경로는 시간에 따라 달라질 수 있고 또한 대상자가 여러 경로 중 항상 최적화된 길로 통행한다고 볼 수는 없으므로, 실제 경로와 예상 경로 사이에 오차가 발생할 수 있다. 일반적으로 출발지와 목적지 사이의 거리가 멀수록 가능한 경로의 수는 더 많아지므로 예상 경로의 정확도는 낮아진다. 따라서 이를 보완하기 위하여 출발지와 목적지 사이에 대상자가 방문한 경유지를 예상 이동경로에 포함하도록 하였으며, 그 결과 〔그림 6〕과 같이 경유지를 포함한 오른쪽 그림에서 정확도가 향상된 것을 확인할 수 있다.

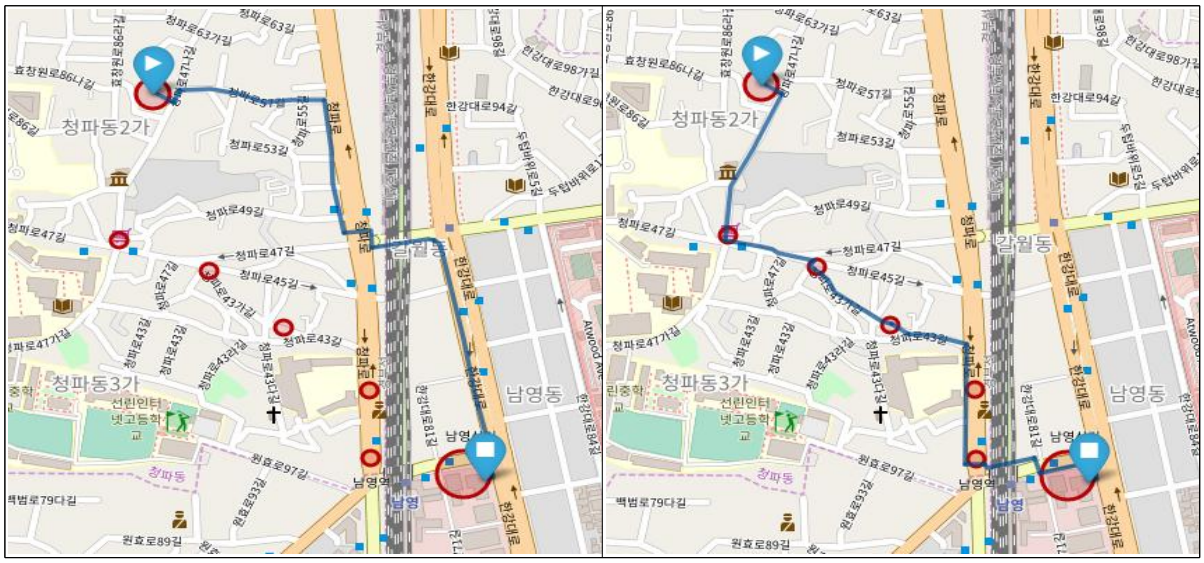

그림 6. 일정 기간 동안 위치 이동을 보여주는 시각화 결과 (도보)
Figure 6. Visualization Results Showing Position Movement over a Period of Time (On Foot)

2.1.4. 주된 활동지역 확인 시각화

앞에서는 특정 장소나 시간에 대하여 대상자의 실제 위치를 확인하였다면, 오랜 기간 동안 수집된 대상자의 위치 정보를 이용한다면 대상자의 주된 활동지역을 확인할 수도 있다. 수사 현장에서 수사관들은 집이나 사무실과 같이 대상자가 주로 머무르는 곳을 확인해야 할 필요가 있을 때가 있다. 특히 현장에 대한 압수수색을 통해 증거를 확보하는 경우, 대상자가 자주 머무르는 곳에 대상자와 관련된 자료가 많이 남아있을 가능성이 높기 때문에 압수수색의 장소를 선정하기 위해서는 대상자의 주된 활동지역을 알 수 있다면 수사에 큰 도움이 된다.

〔그림 7〕은 한 대상자에 대하여 4일 간 낮 시간(12시부터 18시까지) 동안의 대상자 위치기록을 추출한 후, 이를 히트맵(heatmap)으로 표현한 그림이다. 진하게 표시된 지역일수록 해당 시간 동안 그 지역에 위치가 많이 기록됐다는 것을 의미한다. 〔그림 7〕의 왼쪽을 보면 서울 지역 중 서대문구 주변이 진하게 표시되는 것을 확인할 수 있으며, 이를 확대한 오른쪽을 보면 서대문구 중에서도 특히 경찰청 주변이 진하게 표시된 것을 확인할 수 있다. 이를 통해 대상자가 주로 활동하는 지역을 추론해볼 수 있다.

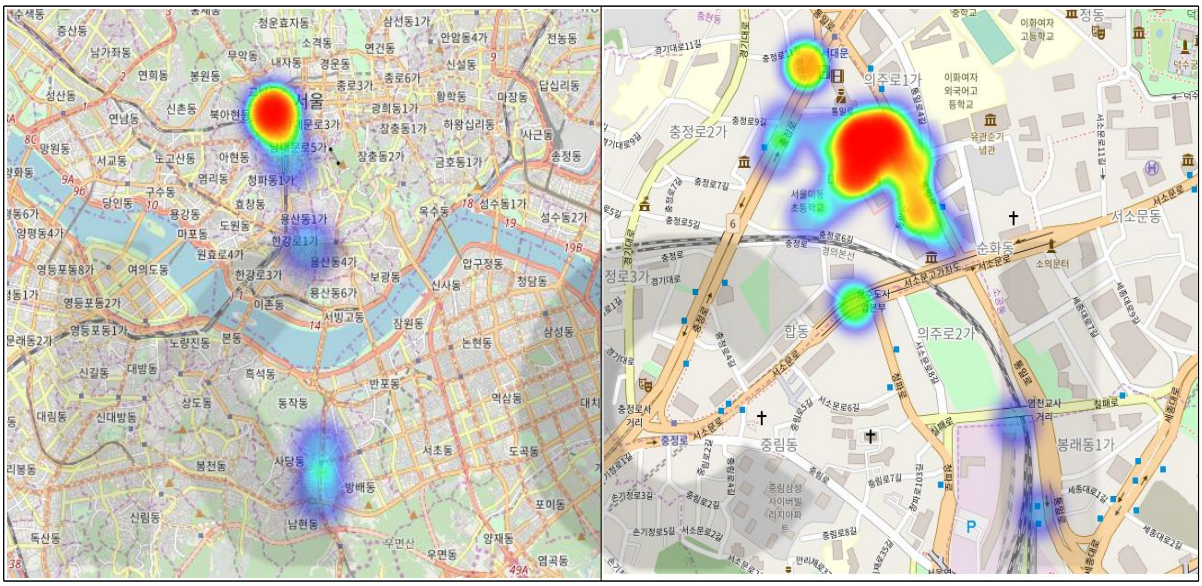

그림 7. 시간대별 대상자의 주된 활동 지역을 보여주는 시각화 결과 (오른쪽은 왼쪽을 확대한 그림)
Figure 7. Visualization Results Showing the User's Primary Activity Area by Time

2.1.5. 범죄현장에 근접한 타임라인 시각화

용의자가 범죄현장에 일회성으로 방문한 것이 아니라 지속적으로 방문한 경우, 용의자가 범죄현장을 방문한 시간을 타임라인으로 확인할 필요가 있다. 또한 범죄현장뿐만 아니라 대상자가 사무실이나 집에 머물렀던 시간대를 확인해야 하는 경우도 있다. 예를 들어 대상자가 과로로 근로자가 사망한 경우, 사망 원인을 규명하는 과정에서 과도한 업무가 사망에 영향을 미쳤는지 확인하기 위하여 근로자가 사무실에 있었던 시간을 알아야 하는 경우가 있을 것이다. 만약 사업장에서 출퇴근 시간이 기록되지 않고 근로자의 근무시간을 입증할 만한 다른 수단이 없더라도 〔그림 8〕과 같이 구글 데이터를 이용하여 타임라인으로 시각화한 결과를 이용한다면 사무실에 출퇴근한 시간을 알 수 있다.

〔그림 8〕은 날짜를 X축으로 하고 시간을 Y축으로 하여 대상자가 특정장소에 근접(가령 100m 이내)한 경우 점으로 표시한 그래프이다. 대상자가 특정장소에 근접한 시간 데이터는 ① 전체 데이터 중 대상이 되는 기간의 데이터만 남긴 후, ② 특정 장소의 위경도와 데이터의 위경도 사이의 거리를 계산하여 근접하였다고 볼 수 있는 특정 값 내의 데이터만 남기는 방법으로 추출하였다.

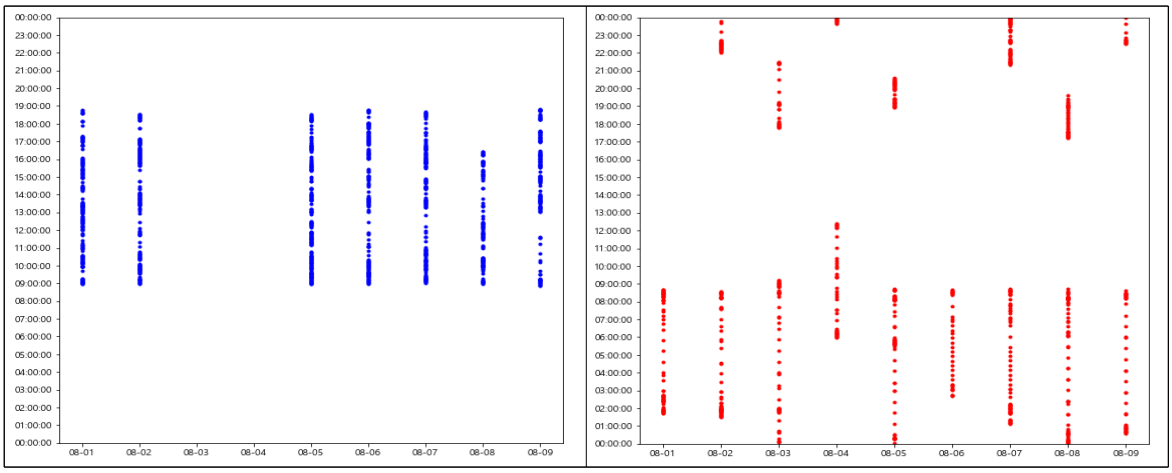

그림 8. 특정장소에 근접했던 시간을 타임라인으로 보여주는 시각화 결과 (왼쪽은 사무실, 오른쪽은 집에 근접한 시간)
Figure 8. Visualization Results That Show the Time Line That Was Close to a Specific Location (Left : Office, Right : Home)

2.2. 어플리케이션 사용내역을 이용한 행동패턴 시각화

어플리케이션 사용내역은 "Takeout\내 활동\Android" 경로에 "내활동.html"이라는 HTML 형태로 저장되어 있다. "내활동.html"을 인터넷 브라우저로 확인하면 [그림 9]와 같이 안드로이드 운영체제가 설치된 모바일 기기에서 대상자가 사용한 어플리케이션 이름, 사용 일시를 확인할 수 있다.

그림 9. 안드로이드 활동내역 중 일부 내역
Figure 9. Part of Android Activity

어플리케이션 활동내역은 특별한 변환을 거치지 않더라도 데이터를 사용할 수 있으나 사용자가 하루에 사용하는 어플리케이션이 여러 가지인 것을 고려한다면 효율적인 시각화 방안이 필요하다. 실제로 앱 데이터 플랫폼 앱애니(App Annie)에서 발표한 '소비자 앱 사용량 집중 탐구' 보고서에 따르면 2017년 1분기 기준으로 한국을 포함한 전세계 주요 국가의 사용자들은 하루 평균 9개 이상의 어플리케이션을 사용하는 것으로 조사됐다[11]. 또한 보고서는 한국 사용자가 평균적으로 하루에 3시간이 넘는 시간 동안 어플리케이션을 사용하며 조사 대상 국가 중 가장 많은 시간동안 어플리케이션을 사용한다고 밝혔다. 따라서 수백, 수천 건의 어플리케이션 사용 내역 중 수사에 필요한 데이터만 추출하고 이를 시각화 하는 방법에 대해서 살펴보고자 한다.

2.2.1. 어플리케이션 사용내역 통계

데이터들의 최소값과 최대값, 평균 등을 확인하는 것은 데이터 분석의 시작점이라고 할 수 있다. [그림 10]의 왼쪽 그림은 안드로이드 활동내역 중 상위 9개 어플리케이션의 통계를 보여주는 원형 그래프이다. [그림 10]의 원형 그래프에 사용된 데이터는 특정 기간 동안 대상자의 안드로이드 활동내역에서 각 어플리케이션별 빈도수를 계산한 이후 상위 9개를 추출하고 9개의 빈도수를 비율로 환산하여 원형 그래프로 표현한 것이다. 이 그래프를 보면 대상자가 메신저로는 주로 카카오톡을 사용하고 있으며 모바일 브라우저는 Samsumg Internet, SNS로는 instagram을 사용하는 것을 확인할 수 있다. 이처럼 상위 어플리케이션 활동내역을 확인하면 대상자가 주로 사용하는 메신저, 인터넷 브라우저, SNS, 메일 서비스 등을

확인할 수 있으며, 이는 압수수색 대상을 선정하거나 모바일 기기에 대한 분석 시에 유용한 참고자료로 활용할 수 있다.

또한 평소에는 사용하지 않으나 범죄 상황과 같이 특수한 상황에서 사용하는 어플리케이션을 확인할 필요도 있다. 따라서 상대적으로 사용횟수가 적은 어플리케이션을 보여줄 필요가 있는데 전체 이용 내역 중 1~2회와 같이 너무 적은 횟수로 사용한 어플리케이션은 그 종류가 많아 구별하기 어렵기 때문에 유의미한 사용(가령 50회 이상 사용) 내역 중 하위 9개를 추출한 결과는 〔그림 10〕의 오른쪽과 같다.

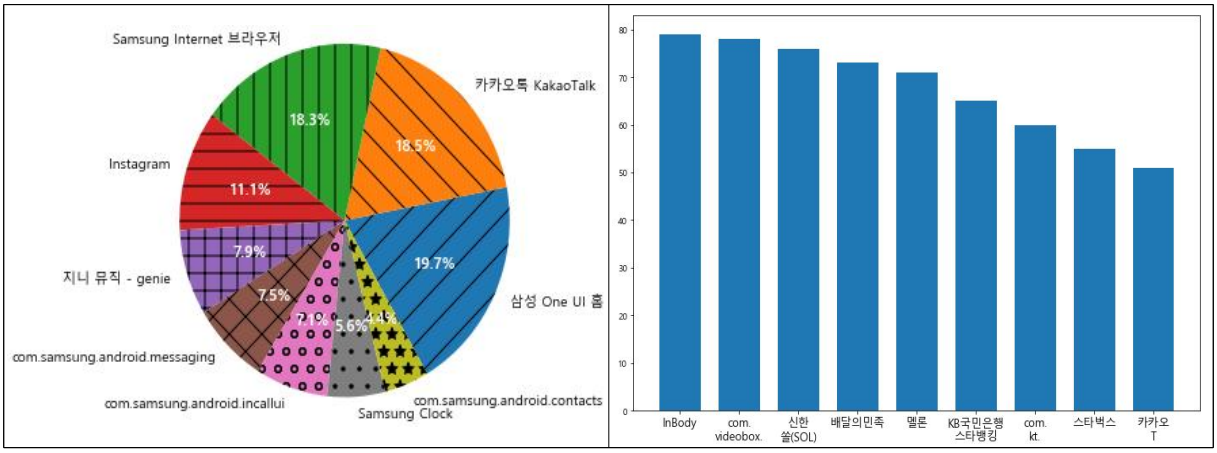

그림 10. 대상자의 안드로이드 활동내역 중 상위(왼쪽), 하위(오른쪽) 어플리케이션의 사용 비율
Figure 10. Percentage of Top (Left) and Bottom (Right) Applications in Android Activity

〔그림 11〕은 특정 기간 동안 시간대별로 대상자의 안드로이드 활동내역의 빈도수를 표현한 막대 그래프이다. X축은 0시부터 24시까지 시간대를 표현하고 있으며 Y축은 안드로이드 어플리케이션 실행 횟수를 의미하며, 그래프 상의 선은 시간대별 실행 횟수의 평균값이다. 그래프를 보면 02시부터 08시까지 어플리케이션 실행 횟수가 다른 시간대에 비해 적은 것을 확인할 수 있다. 이를 통해 대상자가 주로 활동하는 시간을 추론해볼 수 있다.

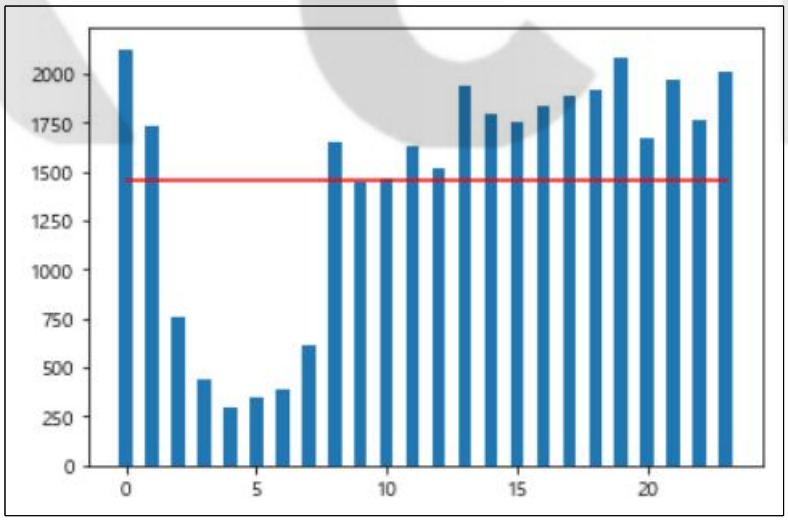

그림 11. 대상자의 시간대별 활동내역 수
Figure 11. The Number of Activities over Time

2.2.2. 특정 어플리케이션을 사용한 타임라인

온라인상에서 발생하는 많은 범죄들은 모바일 어플리케이션을 이용하여 발생한다. 이와 같은 범죄 유형에서 모바일 어플리케이션의 사용기록을 확보하는 것은 범죄수사에서 중요한 단계인데 해외 사업자가 운영하는 어플리케이션의 경우 국내에서 영장을 받는다고 하더라도 사업자로부터 접속기록 등의 자료를 확보하기가 어렵다. 국내 사업자가 운영하는 어플리케이션의 경우 상대적으로 자료를 확보하기 수월하지만, 이마저도 기간이 오래된 경우 자료가 폐기되어 증거를 확보할 수 없는 경우가 많다.

피의자가 텔레그램, 텀블러, 인스타그램 등을 이용하여 음란물을 유포했다고 가정하자. 피의자에게 해당 어플리케이션을 사용했는지 추궁하였으나 피의자는 사용한 적이 없다고 부인하거나 사용했다고 하더라도 범죄가 발생한 그 시간에는 사용한

적이 없다고 진술한다. 이와 같은 피의자의 진술을 반박하기 위하여 수사관은 어플리케이션 사용기록을 확보하려고 하지만 텔레그램이나 텀블러 등은 모두 해외 사업자가 운영하기 때문에 사용자 정보나 접속 기록 등을 확보하는 것이 쉽지 않다. 이와 같은 경우 구글에서 보관하는 사용자 데이터를 시각화하면 범죄 발생 기간 동안 피의자가 특정 어플리케이션을 사용한 타임라인을 확인할 수 있다.

〔그림 12〕는 특정 어플리케이션을 사용한 타임라인을 보여주는 시각화 결과이다. X축은 날짜, Y축은 시간으로 하고 어플리케이션을 사용한 날짜와 시간에 점을 표시하였다. 〔그림 12〕에서는 별 모양의 점은 텔레그램을 사용한 시간, 사각형 점은 인스타그램을 사용한 시간을 표시한 결과이다.

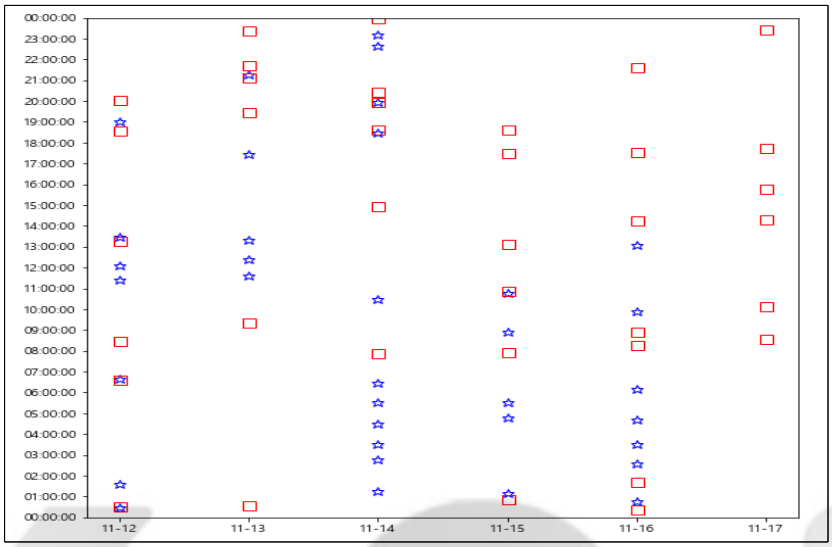

그림 12. 특정 어플리케이션을 사용한 시간을 보여주는 시각화 (별모양 : 텔레그램, 사각형 : 인스타그램)
Figure 12. Visualization Showing the Time that a Specific Application was used

2.2.3. 어플리케이션 사용내역을 이용한 생활패턴 시각화

사람들은 기상할 때부터 취침할 때까지 모바일 단말기를 소지하면서 어플리케이션을 사용한다. 즉, 어플리케이션이 사용된 시간을 대상자가 깨어있는 시간이라고 추정할 수 있기 때문에 어플리케이션 사용 시간들을 그래프 상에 표시한다면 대상자가 주로 언제 취침하고 기상하는지 등 생활패턴을 알 수 있다. 다만 이 경우에는 〔그림 12〕와 같이 특정 어플리케이션의 사용 내역을 통해 특정 시간을 알아내는 것이 목적이 아니라 생활패턴과 같은 전체적인 추세를 파악하는 것이 목적이므로 밀도 그래프(density plot)를 이용하는 것이 더 효과적이다. 〔그림 13〕은 어플리케이션 사용 시간을 이용하여 밀도 그래프를 그린 시각화 결과이다. X축은 날짜, Y축은 시간을 의미하고, 어플리케이션 사용 내역이 많은 일시일수록 더 진하게 표시되며, 사용 내역이 없다면 흰 배경으로 남게 된다. 〔그림 13〕을 보면 대상자가 일정하게 기상하고 취침하는 패턴이 뚜렷이 나타나는 것을 알 수 있다.

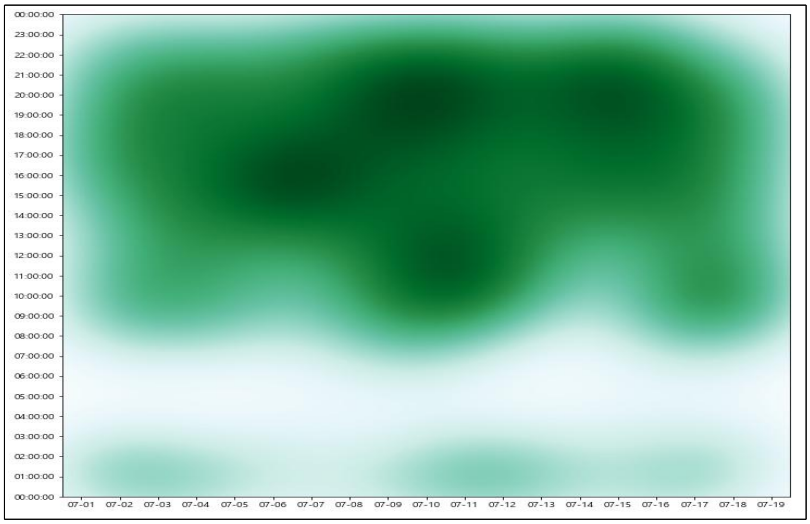

그림 13. 대상자의 시간대별 활동내역 수를 밀도 그래프로 표현한 결과
Figure 13. Density Graph Showing the Number of Activities over Time

2.2.4. 검색내역 시각화

검색내역은 "Takeout\내 활동\검색" 경로에 "내활동.html"이라는 HTML 형태로 저장되어 있다. "내활동.html"을 인터넷 브라우저로 확인하면 [그림 14]와 같이 검색어, 검색한 시간 등을 확인할 수 있다.

그림 14. 검색 활동내역 중 일부 내역
Figure 14. Part of Search History

다만 대상자의 검색어가 양이 많은 경우 웹 브라우저를 이용하여 대용량의 HTML 파일을 웹 브라우저를 이용하여 열람하기 어렵고, 실제 파일을 보더라도 [그림 14]과 같이 시간대(timezone)가 맞지 않고 불필요한 정보가 많아 한 눈에 필요한 정보를 찾기 어려운 문제가 있다. 이를 간결하게 표현하는 것도 시각화의 한 방법으로 검색 내역에서 필요한 정보는 시간과 검색어이므로 이 두 데이터만 추출하고 시간대를 실제 사용자의 시간대로 변환하였다. 그 결과 [그림 15]와 같이 훨씬 가시적으로 대상자의 검색어를 확인할 수 있다.

datetime	activity
2019-06-19 19:26:58+09:00	heidisql text import data truncated
2019-06-19 19:23:58+09:00	heidisql text import
2019-06-19 19:23:57+09:00	heidisql txt import
2019-06-19 19:19:20+09:00	heidisql 데이터 import
2019-06-19 19:19:06+09:00	heidisql 데이터 넣기
2019-06-19 19:01:42+09:00	cmd concat 명령어
2019-06-19 19:01:23+09:00	type 명령어

그림 15. 대상자의 검색어와 시간만을 표시한 결과
Figure 15. Results Showing only the Subject's Search Term and Time

검색어는 대상자의 관심사, 심리상태 등을 확인할 수 있는 유용한 수단이다. 가령 살인 사건의 피의자가 사건 발생 전에 살인 방법, 범행도구 또는 사체 은닉 방법 등을 검색하였다는 사실은 피의자의 범죄가 사전에 치밀하게 계획된 범죄임을 입증하는 증거로 사용하기도 한다. 또한 포렌식에 대한 사회적 인식이 높아짐에 따라 수사기관의 압수수색에 대비하여 안티포렌식 도구를 이용하여 PC나 모바일에 저장된 데이터들을 완전 삭제하는 사례도 있다. 이 때 검색내역에서 '안티포렌식', '완전삭제 프로그램'과 같은 검색어가 발견된다면 대상자의 증거인멸을 입증할 수 있는 하나의 증거로 사용될 수 있을 것이다.

우선 대상자의 주된 관심사를 확인하기 위하여 [그림 16]과 같이 빈도수에 따른 검색내역을 시각화하였다. 구체적으로 대상자의 검색내역을 모두 어절(語節)로 나누고 각 어절의 빈도수를 계산하였다. 계산된 빈도수에 따라 높은 빈도수일수록 검색어의 크기를 크도록 하고 작은 빈도수일 경우 크기를 작도록 시각화하였다. [그림 16]을 보면 대상자의 검색어에는 '파이썬', 'pandas' 등 프로그래밍과 관련된 단어가 많은 것으로 볼 때 프로그래밍에 관심이 많다는 것을 추론해볼 수 있다.

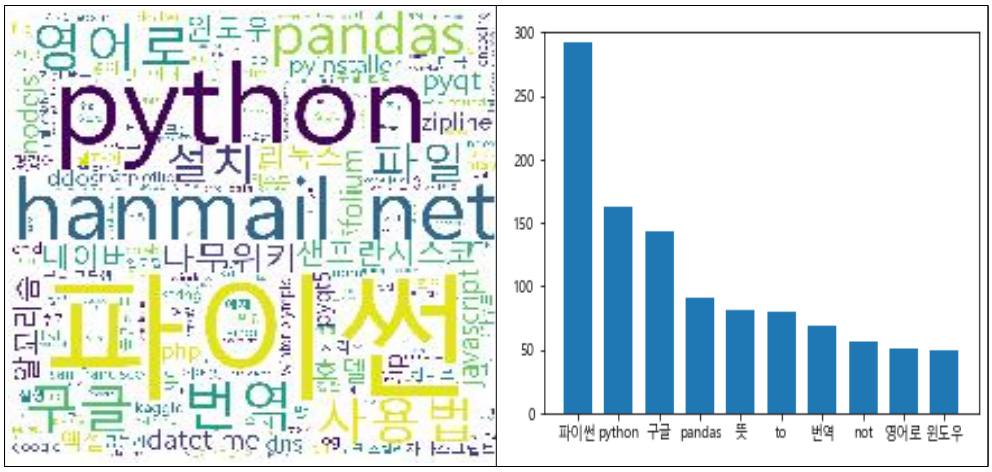

그림 16. 빈도수에 따른 검색내역 시각화
Figure 16. Visualization of Search History by Frequency

또한 시간의 흐름에 따라 대상자가 무엇을 검색했는지 확인하기 위해 타임라인에 대상자의 검색어를 직접 보여주는 시각화도 가능하다. [그림 17]은 시간 흐름에 따른 대상자의 검색어를 보여주는 시각화 결과이다. X축은 날짜, Y축은 시간을 의미하고, 좌표 상에 검색어를 텍스트 형태로 표현하였다. 이 같은 시각화 방법은 대상자가 무엇을 검색했는지 한 눈에 볼 수 있어 가시성이 높을 뿐만 아니라, 대상자의 관심사나 심리상태가 시간에 따라 어떻게 변화하는지 확인할 수 있다는 점에서 유용하다.

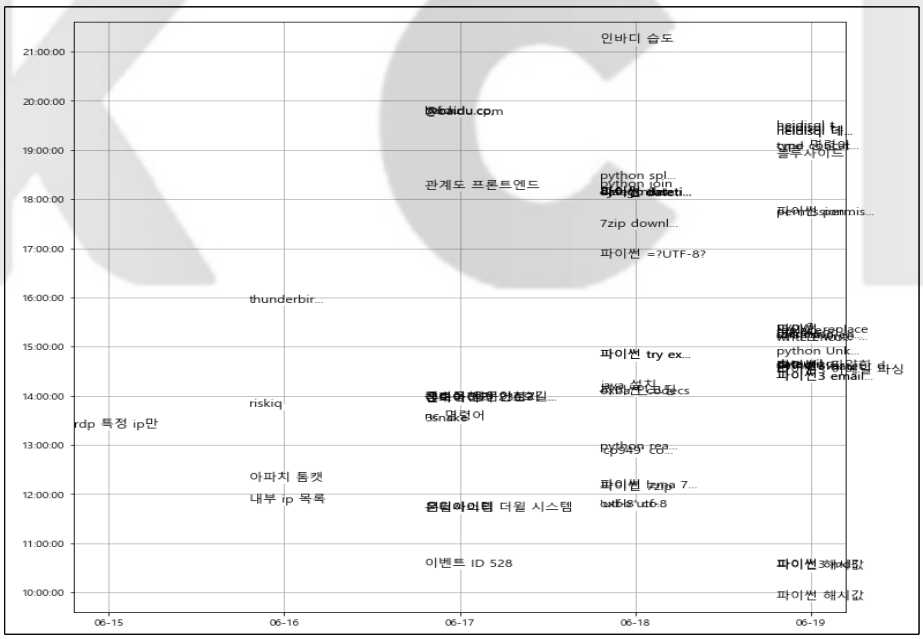

그림 17. 타임라인에 따른 검색내역 시각화
Figure 17. Visualization of Search History along the Timeline

Ⅴ. 결 론

시각화는 인간이 가지고 있는 인지능력의 한계를 극복할 수 있는 좋은 수단으로 많은 정보를 한 눈에 볼 수 있게 하여 데이터의 의미를 이해할 수 있게 해준다는 점과 흩어져 있는 여러 데이터들 사이의 상관관계나 연관성을 보여주고 이로 인해 통찰할 수 있는 중요한 단서를 제공한다는 점에서 그 의의가 있다.

특히 수사와 같이 관련 대상자와 주변 상황에 대한 충분한 이해를 바탕으로 실체적 진실을 밝혀야 하는 작업에서는 산재해있는 복잡한 데이터 사이에서 신속하게 유의미한 정보를 찾아내는 것이 매우 중요하다. 그러나 수사 현실은 녹록치 않은데, 실제로 2018년 기준으로 전체 사이버범죄는 149,604건으로 2017년 대비 약 14% 증가한 반면[12], 사이버 사건에 있어 일선 수사관의 업무량은 여전히 과중하다[13]. 또한 사이버 사건 외에도 일반 경제, 지능, 강력 사건 등에서도 디지털 포렌식 기법이 활용되고 사건 해결에 결정적인 역할을 하는 등 수사에 있어서 디지털포렌식의 수요는 계속 증가하고 있다.

이러한 상황에서 구글 사용자 데이터를 수사에 활용할 수 있도록 시각화한 이번 연구는 아래와 같은 이점을 얻을 수 있다.

첫째, 광범위한 위치 정보와 어플리케이션 사용내역, 검색어 등을 지도와 그래프로 보여줌으로써 범행시간과 장소, 대상자의 동선, 생활패턴 등 수사에 중요 단서가 될 수 있는 많은 정보를 획득할 수 있다. 과거 연구에서는 구글 사용자 데이터를 획득하여 수사에 활용할 수 있는 사례를 제시하였다면[4], 이번 연구에서는 한발 더 나아가 구글 데이터를 활용할 수 있는 구체적인 방법을 제시하였다는 점에서 의미가 있다.

둘째, 데이터의 특성상 일반 수사관들이 접근하거나 정제하기 어려운 이유로 데이터 활용도가 떨어지는 한계가 있는데, 이를 극복할 수 있는 방법을 제시하였으며 그 결과 데이터를 효율적으로 수사에 활용할 수 있도록 하였다. 특히 대량의 데이터 중 필요한 자료만 추출하여 효율적으로 처리하여 수사관들이 별도의 전처리 과정 없이 수사 현장에서 바로 사용할 수 있는 방법을 제시하였으며, 향후 이것이 수사관들이 이용할 수 있는 도구로서 기능할 수 있도록 발판을 마련하였다. 실제 시각화에 사용된 소스코드를 Github 페이지(https://github.com/briron/google-user-data-visualization)에 공개하여 누구든지 이를 사용할 수 있다.

마지막으로 데이터는 가공, 결합하는 방법과 이를 보여주는 방식에 따라 여러 가지 의미를 도출해 낼 수 있는데, 이 연구에서 시각화 모델을 한 가지 제시함으로써 추후에 이것을 변형한 새로운 시각화 모델을 만들어 낼 수 있는 재료가 되었다는 점에서 발전가능성이 있다. 무엇보다 사용 내역 등의 기초 통계를 보여주는 것뿐만 아니라 예상 경로, 생활 패턴, 활동 지역, 상태 변화 등 대상자에 대한 복합적이고 동적인 정보를 보여주었으며, 이를 통해 CCTV 자료 확보, 압수수색 대상 선정, 진술의 진위 판단, 범행 도구 및 방법 확인, 다른 사건과의 연관성 확인 등 유의미한 수사 단서를 확보할 수 있는 새로운 수사기법을 제시하였다.

IT 산업이 고도화되고 자율주행, 인공지능, 사물인터넷이 발전함에 따라 앞으로 기업들이 수집하는 사용자 데이터는 확대될 것이 분명하다. 이번 연구에서는 구글 사용자 데이터만을 대상으로 하여 시각화 방안을 제시하였지만 애플, 마이크로소프트, 네이버 등에 산재한 데이터로 그 연구대상이 확대될 것으로 기대한다.

참 고 문 헌 (References)

〔1〕 Cisco, "Cisco Visual Networking Index: Forecast and Trends", 2017 - 2022 White Paper, pp.6, 2019.

〔2〕 Terrence Nemayire, Alex Ogbole, Sungmi Park, Keecheol Kim, Yeonseok Jeong and Yunsik Jang, "A 2018 Samsung Smart TV Data Acquisition Method Analysis". Journal of Digital Forensics, vol.13, No.3. pp.205-218, 2019.

〔3〕 Kwak, Na-Yeon, Choong C. Lee, Maeng, Yun-Ho, Cho, Bang-Ho and Lee, Sang-Eun, "A Meta Study on Research Trend of Digital Forensic in Korea". Informatization Policy, vol. 24, pp.91-107, 2017.

〔4〕 Dongho Kim and Sangjin Lee, "A Study on the Usage of Investigation of Google Cloud Data (Smartphone user-oriented)". Journal of Digital Forensics, vol.12, No.3, pp.107-118. 2018.

〔5〕 Seokhyun Jang, Jooyoup Lee and Kyungwon Lee, "Study of Representation Methodology by Comparative Analysis between Information Visualization and Knowledge Visualization". In Proceedings of the HCI Society of Korea, pp.1242-1248, 2008.

〔6〕 Gwang-Seon Choe, Yeong-Gyeong Ham and Seon-Ho Kim, "Big Data Visualization", Korea Society of Computer Information Review, vol.21. No.1. pp.33-43, 2013.

〔7〕 Dong-gun Lee, Huy Kang Kim and Eunjin Kim, "Study on security log visualization and security threat detection using RGB Palette". Journal of The Korea Institute of Information Security & Crypto logy. vol.25, No.1, pp. 61-73, 2015.

〔8〕 Woo-Jin Joe, Hyo-Jeong Shin and Hyong-Shik Kim, "A log visualization method for network security monitoring". Smart Media Journal, Vol.7 No.4. pp.62-70, 2018.

〔9〕 Ji-On Kim, "A Study on the Application of Social Network Analysis Principles to Criminal Investigation". Journal of Digital Forensics, Vol.13. No.2. pp.87-107, 2019.

〔10〕 Youngok Kang and Hyeon Deok Kim, "A Study on Geographic Visualization of Public Data Using Google API in Cloud Computing Environment". Journal of the Korean Cartographic Association, Vol.14. No.1. pp.1-15. 2014.

〔11〕 App Annie, "Spotlight on Consumer App Usage", pp.10. 2017.

〔12〕 Cyber Bureau, Korean National Police Agency, "2018 Cyber Threat Report", pp.4. 2019.

〔13〕 Jeong Woong, "Analysis on the Workload of Cyber Investigation Team in the Police Station". The Journal of Police Policies, vol.31, no.3, pp.27-60. 2017.

저 자 소 개

김 현 우 (HyunWoo Kim)
준회원
2014년 3월 : 경찰대학 행정학과 졸업
2018년 1월 ~ 현재 : 경찰청 사이버안전국 사이버수사과
2018년 8월 ~ 현재 : 고려대학교 정보보호대학원 석사과정
관심분야 : 디지털 포렌식

이 상 진 (SangJin Lee)
정회원
1989년 10월 ~ 1999년 2월 : ETRI 선임연구원
1999년 3월 ~ 2001년 8월 : 고려대학교 자연과학대학 조교수
2001년 9월 ~ 현재 : 고려대학교 정보보호대학원 교수
2008년 3월 ~ 현재 : 고려대학교 디지털포렌식연구센터 센터장
2017년 3월 ~ 현재 : 고려대학교 정보보호대학원 원장
관심분야 : 디지털 포렌식, 심층암호, 해쉬함수