JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 29 No. 2, pp. 109-118, February 2024
https://doi.org/10.9708/jksci.2024.29.02.109

A Tracking Method of Same Drug Sales Accounts through Similarity Analysis of Instagram Profiles and Posts

Eun-Young Park*, Jiyeon Kim*, Chang-Hoon Kim**

*Undergraduate Student, Dept. of Computer Engineering, Daegu University, Gyeongsan, Korea
*Professor, Dept. of Computer Engineering, Daegu University, Gyeongsan, Korea
**Professor, Dept. of Information Security, Daegu University, Gyeongsan, Korea

[Abstract]

With the increasing number of social media users worldwide, cases of social media being abused to perpetrate various crimes are increasing. Specifically, drug distribution through social media is emerging as a serious social problem. Using social media channels, the curiosity of teenagers regarding drugs is stimulated through clever marketing. Further, social media easily facilitates drug purchases due to the high accessibility of drug sellers and consumers. Among various social media platforms, we focused on Instagram, which is the most used social media platform by young adults aged 19 to 24 years in South Korea. We collected four types of information, including profile photos, introductions, posts in the form of images, and posts in the form of texts on Instagram; then, we analyzed the similarity among each type of collected information. The profile photos and posts in the form of image were analyzed for similarity based on the SSIM(Structural Simplicity Index Measure), while introductions and posts in the form of text were analyzed for similarity using Jaccard and Cosine similarity techniques. Through the similarity analysis, the similarity among various accounts for each collected information type was measured, and accounts with similarity above the significance level were determined as the same drug sales account. By performing logistic regression analysis on the aforementioned information types, we confirmed that except posts in image form, profile photos, introductions, and posts in the text form were valid information for tracking the same drug sales account.

▸ **Key words**: Cyber investigation, Drug Distribution, Crawling, Social Media, Instagram

• First Author: Eun-Young Park, Corresponding Author: Jiyeon Kim
*Eun-Young Park (pey6693@daegu.ac.kr), Dept. of Computer Engineering, Daegu University
*Jiyeon Kim (jyk@daegu.ac.kr), Dept. of Computer Engineering, Daegu University
**Chang-Hoon Kim (kimch@daegu.ac.kr), Dept. of Information Security, Daegu University
• Received: 2024. 01. 05, Revised: 2024. 02. 05, Accepted: 2024. 02. 05.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

[요 약]

전 세계 소셜 미디어 사용자가 증가하면서 다양한 범죄의 수단으로 소셜 미디어가 악용되는 사례가 증가하고 있다. 특히, 소셜 미디어를 통한 마약 유통은 마약 판매자와 소비자의 높은 접근성으로 인해 청소년들의 마약 호기심을 자극하고, 구매를 용이하게 한다는 점에서 심각한 사회문제로 대두되고 있다. 본 논문에서는 다양한 소셜 미디어 중, 국내 19세에서 24세 청소년이 가장 많이 사용하는 인스타그램을 대상으로 프로필 사진, 소개글, 게시물 사진과 게시글을 수집하고, 각 정보의 유사도 분석을 통해 수집한 다수의 계정을 활용하여 마약을 유통하는 마약사범 추적 기술을 개발한다. 4개 수집 정보 중, 이미지 형태의 프로필 사진 및 게시물 사진은 SSIM(Structural Similarity Index Measure) 기반으로 유사도를 분석하고, 텍스트 형태의 소개글 및 게시글은 자카드 유사도 및 코사인 유사도 기법을 사용하여 유사도를 분석한다. 이와 같은 유사도 분석을 통해, 각 수집 정보별 계정 간의 유사도를 측정할 수 있으며 유의수준 이상의 유사성을 갖는 계정들에 대해 동일 마약 유통 계정으로 판단할 수 있다. 또한, 수집한 4개 정보에 대해 로지스틱 회귀분석을 수행하여 게시물 사진을 제외한 프로필 사진, 소개글, 게시글이 동일 마약 판매 계정을 추적하는 데에 유효한 정보임을 확인하였다.

▸**주제어**: 사이버 수사, 마약 유통, 크롤링, 소셜 미디어, 인스타그램

I. Introduction

소셜 미디어(Social Media)는 소셜네트워크를 기반으로 하여 온라인상에서 콘텐츠 공유, 상호 커뮤니케이션 및 정보 공유 등 다양한 목적을 위해 활용되는 플랫폼으로서 전 세계 인구 약 60% 이상이 사용하고 있다[1]. 소셜 미디어 플랫폼은 접근성이 매우 높아 사용자의 평균 연령도 계속 낮아지는 추세이지만 소셜 미디어의 빠른 성장과 함께 이를 통한 사이버 범죄와 위협도 늘어나고 있는 실정이다. 다양한 소셜 미디어 중에서도 인스타그램 (Instagram)은 특유의 시각적 콘텐츠 활용과 사용자들 간의 용이한 연결성으로 인해 다른 소셜 미디어보다 젊은 연령층의 사용자 수가 많지만, 이러한 인스타그램의 특성을 악용하여 인스타그램을 통한 마약 유통 또한 증가하고 있다. 인스타그램에서 마약 유통을 위한 계정이 다수 발견되고 있으며 운영자들은 복수 계정 생성, 주기적 계정 폐쇄 등을 통해 사이버 추적을 회피하는 특성을 보인다. 마약 범죄는 마약 남용 자체로도 위험하지만, 마약 남용 후 행해지는 2차 범죄로 인하여 사회적 안전에도 위협을 가하는 중대한 문제이다. 특히, 소셜 미디어를 통한 마약사범과의 접촉이 점점 용이해지기 때문에 10대 청소년을 비롯한 일반인들을 마약으로부터 보호하기 위한 집중 마약 단속 수사가 필요하다. 미국은 1973년부터 미국 법무부 산하기관인 마약단속국(Drug Enforcement Administration, 이하 DEA)을 운영하며 미국 내 마약 범죄를 집중적으로 단속하고 있으며 2022년에는 애플(Apple) 사의 에어 태그(Apple Air Tag)를 이용하여 중국에서 미국 마약 제조업체로 배송된 불법 마약을 추적한 바 있다[2-3]. 국내의 경우, 마약범죄수사대와 마약수사대 조직을 구성하고, 2019년부터 2022년까지 대검찰청과 법무부가 마약·조직범죄에 대한 전담 수사청 법안 정책을 추진한 바 있다. 최근 마약 범죄는 눈에 띄지 않는 점조직 형태의 소규모로 은밀하게 이뤄지는 경향을 보이고 있으므로 정보 수집과 즉각적인 조치가 중요하다[4]. 그러나 별도 수사청 설치가 진행되지 않아 수사권 조정의 일환으로 검찰의 직접 수사 범위가 축소되고, 인력 부족, 수사권 제한, 실적 경쟁 등 컨트롤 타워 부재로 인한 수사의 비효율성으로 인해 여전히 사각지대가 발생하고 있다. 마약 범죄는 전 세계적으로 연결되는 범죄 네트워크를 형성하고 있으므로 이러한 마약 유통 특성을 파악하여 수사 기법을 개발하는 것이 필요하다. 기존의 수사 방식은 주로 단속이나 마약 사용 현장 발견 후, 신고 및 수사관의 주도하에 이루어져 왔지만, 소셜 미디어를 활용한 마약 유통이 증가하고 있으므로 사이버 수사를 통한 마약 단속 수사 기법 개발이 필수적이다. 본 연구에서는 국내 청소년이 가장 많이 사용하는 소셜 미디어 플랫폼인 인스타그램에서[5] 정보를 수집하고, 수집된 정보의 유사도 분석을 통해 동일 마약 유통자로 의심되는 유사 계정을 추적하는 모델을 개발하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 기존에 수행된 사이버 범죄 추적 연구 및 소셜 미디어 크롤링을 수행

하는 다양한 분야의 연구를 살펴보고, 3장에서는 동일 마약 유통 계정을 추적하기 위해 인스타그램에서 수집할 4개 정보 정의 및 각 정보별 유사성 분석 방법을 제시한다. 4장에서는 크롤러 개발을 통해 실제 마약 관련 계정을 수집하고, 제안된 알고리즘을 기반으로 계정의 유사성을 분석하여 본 연구의 알고리즘 유효성을 검증한다. 마지막으로 5장에서는 결론 및 향후 연구 계획을 제시한다.

II. Related Works

2.1 Existing Cyber Crime Investigation Studies

사이버 범죄(Cyber Crime)는 정보통신망 침해 범죄와 정보통신망 이용 범죄, 불법 콘텐츠 범죄로 분류되고 있다[6]. 이러한 사이버 범죄는 일반 범죄와는 달리 단기간에 불특정 다수에게 악영향을 끼치며, 온라인 범죄라는 특성상 범인의 특정이 어렵고, 피해 사실을 인지하지 못하는 경우도 존재하여 수사의 어려움이 있다.

이러한 사이버 범죄는 2021년을 제외한 최근 5년간 우상향 추세로 나타나고 있고, 증가하는 범죄 활동에 대처하기 위해 관련된 연구가 꾸준히 진행 중이다. 보이스피싱 근절을 위해, 수집한 은어를 기반으로 웹 크롤링 기술을 추가적으로 활용하여 인출책을 모집하는 광고를 탐지하거나 관계망 분석 기술을 통하여 보이스피싱 범죄의 조직분석 및 인공지능 음성인식 기술에 은어 사전을 연결하여 수사 지원에 도움을 주는 연구[7], 데이터 마이닝과 자연어 처리 기법을 활용하여 범죄 수사를 위한 채팅 로그 분석 프레임워크를 제안하고 사이버 범죄 수사에 효과적인 도움을 주는 연구들은 국내외에서 지속적으로 진행되어 왔으며[8], 기존의 사이버 범죄 수사 연구들은 웹 크롤링을 이용하여 우선순위 도출을 위한 TF-IDF 가중치 분석으로, 아동학대와 관련된 단어를 분석하여 아동학대 수사에 활용할 수 있는 증거 기반 키워드 추출 연구와[9] 소셜네트워크 분석원리 중 2-모드 네트워크 분석원리를 활용하여 사이버 금융 범죄 중 하나인 조건만남 범죄의 수사단서로 사용되는 인터넷뱅킹 로그 자료와 메신저 접속로그 자료를 분석하는 연구[10], LECEN 웹 크롤러를 사용하여 하이퍼링크로 연결된 아동 착취 웹 사이트 집합의 데이터를 수집하고 다양한 장애 전략이 아동 착취 네트워크에 미치는 영향에 대한 연구들이 이루어졌다[11]. 추가적으로 악의적인 해킹과 같은 사이버 위협을 식별할 목적으로 다크넷과 딥넷 같은 곳에서 정보를 얻기 위한 운영 시스템을 데이터 마이닝과 NB(Naive Bayes), RF(Random Forest), SVM(Support Vector Machine) 및 LOG-REG(logistic regression)와 같은 기계 학습 기술을 사용하여 제시하는 연구 또한 계속해서 수행되고 있으며[12], 사이버 범죄 수사 자체의 도움을 주기 위한 사이버 수사에 사용되는 정보 자원의 장기적인 보존을 위해 웹 아카이버를 활용하는 정보 보존 방법 분석 연구와[13] 악성코드 유포자를 효율적으로 추적하기 위해 전통적 분석 방법과 OSINT, Intelligence 같은 최근의 방법을 융합한 차세대 악성코드 정보수집 아키텍처를 제안하는 연구가 진행되고 있다[14].

2.2 Existing SNS Crawling Studies

크롤링(Crawling)은 웹 페이지의 구조를 분석하여 정보를 추출하는 행위를 말하며 자동화된 시스템을 통하여 HTTP로부터 데이터를 수집하는 것으로, 크롤링을 수행하는 소프트웨어(Software)를 크롤러(Crawler)라 부른다[15].

크롤러를 이용한 연구는 다방면으로 계속해서 진행되고 있으며, CNN(Convolutional Neural Network)과 같은 딥러닝을 학습시키기 위해 웹 크롤링을 통하여 이미지를 수집하고, 전이 학습을 이미지 분류 모델에 접목하여 자동화된 이미지 분류 모델을 제안하는 연구가 수행된 바 있다[16].

데이터 크롤링 방식의 계약적 제한에 관한 연구도 진행되었으며[17], 이들 연구는 모두 웹 페이지에서 추출한 데이터의 유효성을 검증하였다. 또한, 사물인터넷 관련 사이버 위협 정보를 클리어 웹의 보안 웹 사이트, 소셜 웹의 보안 포럼, 다크웹의 해커 포럼/마켓플레이스에서 데이터를 투명하게 수집하기 위하여 기계 학습 기반 크롤러를 사용하고 최첨단 통계 언어 모델링 기술을 사용하는 새로운 크롤링 아키텍처를 제시하는 연구가 이루어졌다[18].

마지막으로 SNS 내에서 크롤러를 활용한 연구를 살펴보면 다음과 같다. 웹 크롤링을 통한 연구는 '자해' 관련 인스타그램 게시물의 내용을 네트워크 분석 및 텍스트 분석하여 토픽 주제를 제시하는 연구[19], 크롤링으로 수집된 소셜 미디어 데이터를 사용하여 애플리케이션 개발 방법을 제시하는 연구[20], 사이버 보안을 위한 다크웹 크롤링으로 log4j 취약점, REvil 랜섬웨어 및 WannaCry 랜섬웨어를 고려한 논의를 제안하는 연구가 진행되었고[21], SNS 포스팅을 크롤링한 후 토픽 모델링을 통한 브랜드 이미지 분석으로 브랜드 아이덴티티 전략이 효과적으로 적용되었는지 검증하는 연구 또한 진행된 바 있다[22]. 또한, 사이버 수사를 위해 텍스트 분석 기법의 결합을 중심으로 트위터 내에서 이상 계정을 식별하는 연구[23], 인스타그램 상의 마약 범죄를 추적하기 위한 크롤링 연구로서 프로필 사진의 절대오차 비교를 통해 계정의 유사도를 분석하는 연구도 진

행되었다[24].

기존에 수행된 SNS 기반 크롤링 연구들의 경우[16-22], 데이터를 수집하는 기술 개발에 초점을 두었다는 점에서 사이버 수사 목적으로 SNS 크롤링을 수행하는 본 연구와는 차이가 있다. 또한, 트위터 및 인스타그램 기반의 사이버 수사 연구의 경우에도[23-24] 게시글 또는 프로필 사진과 같이 특정한 정보만 활용한다는 점에서 프로필 사진 외의 소개글, 게시물 사진, 게시글 정보를 활용하는 본 연구와 차이가 있다. 특히, 픽셀의 절대오차 비교를 수행하는 기존 연구와 달리, 본 연구는 이미지의 구조적 차이를 분석하는 SSIM 알고리즘을 사용하였다는 점에서 차별화된다.

III. Tracking Model of Same Drug Sales Accounts

본 장에서는 Selenium, Chrome Driver 및 Python을 활용하여 인스타그램 크롤러를 개발하고, 분석에 필요한 데이터 셋을 직접 수집한다. 또한, 동일 마약 판매 계정을 추적하기 위해 수집할 정보를 정의하고, 각 정보별 유사성 분석 알고리즘을 설계한다.

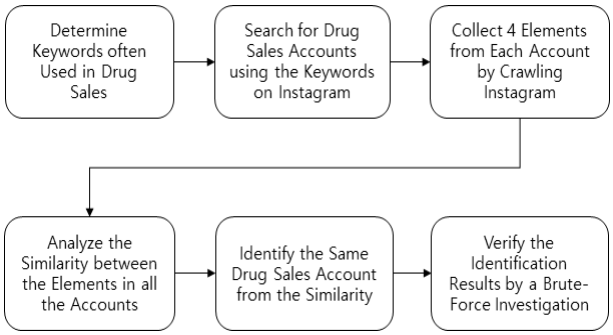

Fig. 1. Design of a Tracking Algorithm of the Same Drug Sales Accounts

본 논문이 제안하는 마약 계정 추적 과정은 Fig. 1과 같다. 먼저 인스타그램 계정에서의 사용 빈도가 높은 마약 은어를 분석하기 위해 인스타그램, 텔레그램과 같은 대표적인 소셜 미디어에서 은어를 검색하였을 때 마약과 관련된 계정 정보가 다수 나타나는 키워드를 Table 1과 같이 선정하고 이를 입력값으로 설정한다. 키워드에 따른 결괏값 중 계정 ID를 리스트로 저장하고 저장한 ID를 동적으로 바뀌는 URL에 반복적으로 대입하며 각 계정의 정보를 수집한다.

Table 1. The Number of Collected User Accounts by Keywords

Keyword	Number of Accounts	Keyword	Number of Accounts
물뽕	19	차가운술	4
시원한술	8	캔디 케이	4
아이스 가이	32	텔레 아이스	9
아이스작대기	5	텔레 얼음	2
아이스 펜타닐	1		

본 논문에서는 유사도 분석을 위한 정보를 인스타그램 내에서 수집하여 각 정보별 분석을 수행한다. 수집한 정보는 Table 2와 같으며 연구의 효율성을 위해 각 정보를 $\mathrm{F}n$으로 정의하였다.

Table 2. Definition of Instagram Elements

Element		Description
F_1	Profile Image	A unique introductory picture of a user's account
F_2	Biography	Introduction to a user's account in text form
F_3	Post image	A user's posts in image form
F_4	Post Text	A user's posts in text form

유사도 분석은 각 계정의 정보를 수집하는 알고리즘에 따른 크롤러를 실행한 후 전체 결과를 CSV(Comma Separated Values) 형태로 저장하여 코사인 유사도와 자카드 유사도, SSIM을 이용하여 비교하였다. 코사인 유사도란 벡터 간의 코사인 각도를 구하여 두 단어 간의 유사도를 연관도로 적용하는 방식으로서 수식 (1)과 같다.

$$\cos\theta = \frac{A \cdot B}{\| A \| \ \| B \|} \qquad (1)$$

자카드 유사도 수식 (2)와 같이 두 집합의 합집합 중, 교집합의 비율을 구하여 유사도를 0과 1 사이의 결괏값으로 도출해낸다.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \qquad (2)$$

자카드 유사도는 문서 간 단어 합집합의 원소 개수와 교집합 원소 개수의 비율을 고려하기 때문에, 어떤 특정 단어가 몇 번씩이나 중복되었는지는 고려하지 않으며 다른 종류의 단어가 얼마나 많이 중복되었는지만 고려한다.

본 논문에서는 수식 (1)과 수식 (2)와 같이 코사인 유사도 및 자카드 유사도를 소개글 및 게시글 유사도 분석에 사용하고, 프로필 사진 및 각 게시물 사진의 유사도 분석을 위해서는 수식 (3)과 같은 SSIM을 사용한다. SSIM은 기준 이미지와 비교 이미지의 휘도, 대비 및 구조를 그레이스케일로 비교하여 유사도를 0과 1 사이의 결괏값으로 계산한다.

$$\mathrm{SSIM}(x,y)=\frac{(2\mu_x\mu_y+c_1)(2\sigma_{xy}+c_2)}{(\mu_x^2+\mu_y^2+c_1)(\sigma_x^2+\sigma_y^2+c_2)} \quad (3)$$

IV. Experimental Results

본 장에서는 3장에서 제시한 알고리즘을 기반으로 마약 계정을 수집하고, 계정 간의 유사성 분석을 수행한다.

본 논문에서는 Table 1에 검색된 총 84개의 계정 중, 전수조사를 통해 마약 유통 계정으로 판단된 Table 3의 18개 계정을 대상으로 하여 동일 마약 판매 계정 추적을 위한 유사도 분석을 수행한다.

Table 3. Drug Sales Accounts Determined by a Brute-force Investigation

ID	User Account	ID	User Account
U_1	ai********** *****on	U_{10}	i****0
U_2	ba***03	U_{11}	mu********55
U_3	ca**********67	U_{12}	pa******84
U_4	ca*********_l	U_{13}	po*********ng
U_5	do********88	U_{14}	qq******oc
U_6	go*****25	U_{15}	si********ul
U_7	ha*******71	U_{16}	tc***un
U_8	i****4	U_{17}	te**********eu
U_9	i2***an	U_{18}	un****m5

단, Table 1에 정의된 키워드 기반 계정 수집 시, 해당 키워드가 범죄 용어가 아닌, 일상 용어로 사용되는 경우는 수동 검증을 통해 필터링하는 과정이 필요하다. 예를 들어, “아이스작대기”의 “작대기”의 경우, 마약 은어로도 사용되지만, 당구 큐대를 의미하는 단어로도 사용되므로 이러한 계정은 수동으로 검증하여 필터링하는 과정이 필요하다.

Table 2에서 정의한 각 정보의 유사성 분석을 통해 동일 마약 계정을 검출하기 위해서는 18개 계정 분석을 통하여 사전에 동일 계정을 정의하는 것이 필요하다.

본 논문에서는 18개 계정을 2개 계정씩 전수 비교하여 총 153개 조합을 수동 검증하여 23개의 동일 계정을 검출하였다. 전수조사를 통해 동일 계정을 판단하는 알고리즘은 다음과 같다. F_1은 두 계정의 프로필 사진 유사성이 SSIM 기반으로 도출한 유사도의 최댓값인 1이 아니더라도 육안으로 동일한 이미지라고 판단될 경우, 동일하다고 판단한다. F_2는 두 계정의 소개글을 구성하고 있는 중복 단어의 개수를 확인하여 50% 이상 중복될 경우, 동일하다고 판단한다. F_3은 두 계정의 전체 게시물 사진 중, 육안으로 확인하여 동일 이미지라고 판단되는 게시물의 수가 50% 이상인 경우, 동일하다고 판단하였으며 F_4는 두 계정의 게시글을 구성하는 해시태그의 중복 개수 및 문장을 구성하는 단어의 유사도를 판단하여 동일 여부를 판단하였다. 특히, F_4의 경우, 단어의 순서 또한 고려하여 중복된 단어들이 동일한 순서로 이루어진 경우에만 동일하다고 판단하였다. 최종적으로는 F_1부터 F_4가 모두 동일한 경우, 동일 계정으로 판단한다.

4.1 Similarity Analysis of Profile Photo and Biography

코사인 유사도, 자카드 유사도 및 SSIM을 활용하여 프로필 부분의 F_1, F_2를 기반으로 분석한 결과, F_1의 유사도 결과는 중복을 제거한 153개의 결괏값을 확인할 수 있으며 이 중, 프로필 사진을 등록하지 않은 기본 F_1의 유사도 결과를 제외하여 총 81개의 결괏값이 나타났다.

Fig. 2에서 볼 수 있듯이 F_1의 경우 도출된 결과에서 유사도가 높은 값은 앞서 동일 인물이라고 판별하였던 값인 Label이 1로 나타나고, 유사도가 낮으면 판별 값은 0에 분포되어 있는 것을 확인할 수 있다. Label은 동일 인물이라고 판별되는 계정은 1을, 나머지 계정은 0을 부여한다.

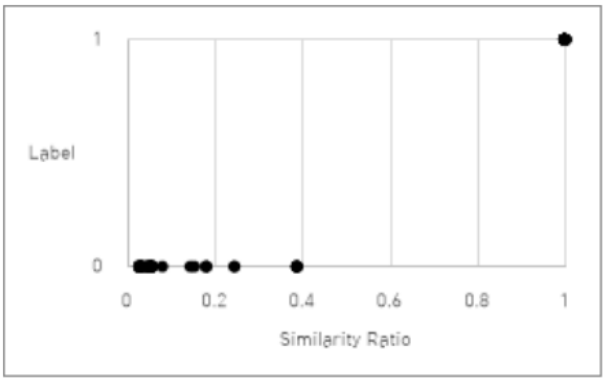

Fig. 2. Determining Same Drug Sales Accounts by Analyzing Profile Photos

각 계정의 F_2 유사도 분석 결과 또한 Table 3에서 정의하였던 U_1, U_2, … U_{17}, U_{18}의 계정을 비교하며 F_1의 결과와 마찬가지로 중복을 제거한 153개의 결과가 나타난다. F_2 같은 경우 모든 계정이 소개글을 등록하였으므로 최종 결과 개수 또한 153개이다. 이 중 유사도 값이 코사인 유사도, 자카드 유사도 두 개의 결과에서 하나라도 0의 값이 포함된 경우인 107개를 제외하여 46개의 결과가 나타났고, Fig. 3과 같이 F_2의 코사인 유사도 분석 경우 도출된 결과에서 유사도가 0.6-0.8 사이 값을 기준으로 Label 또한 0과 1로 나뉘는 것을 볼 수 있다.

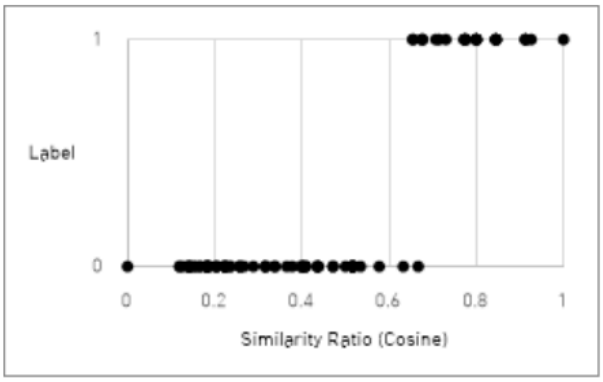

Fig. 3. Determining Same Drug Sales Accounts by Analyzing Biography using a Cosine Similarity Technique

소개글의 자카드 유사도 분석을 수행한 결과, 유사도가 낮음에도 불구하고 Label은 1인 결과가 다수 존재하므로 자카드 유사도의 동일 계정 판별 유효성은 비교적 떨어진다는 것을 Fig. 4를 통해 확인할 수 있다.

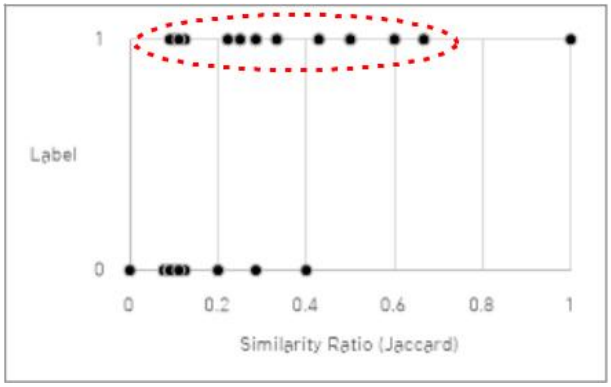

Fig. 4. Determining Same Drug Sales Accounts by Analyzing Biography using the Jaccard Similarity Technique

F_1과 F_2의 유사도 분석 후, F_1의 유사도가 1인 [U_5, U_{12}, U_{18}]의 이미지를 육안으로 확인한 결과 Fig. 5처럼 동일한 사진으로 보여지는 것을 알 수 있고, F_2의 유사도가 1인 [U_8, U_{18}] 또한 육안으로 보았을 때 Fig. 6처럼 동일한 문자열임이 확인 가능하다.

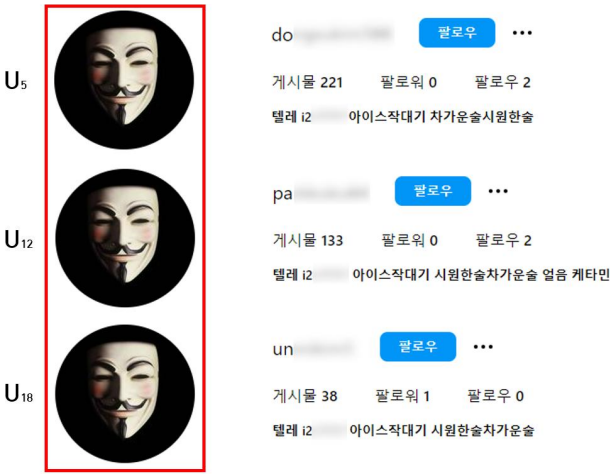

Fig. 5. Comparison of Profile Photos

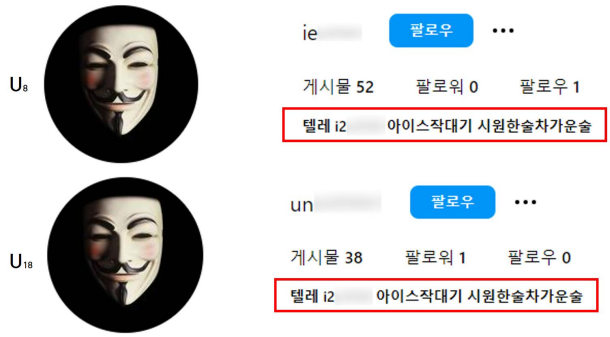

Fig. 6. Comparison of Biography

4.2 Similarity Analysis of Posts in the forms of Image and Text

본 절에서는 코사인 유사도, 자카드 유사도, SSIM을 이용하여 게시물 사진의 유사도 분석을 수행한다. 먼저 F_3에 대한 분석은 SSIM을 이용하여 게시물 사진이 존재하지 않는 경우를 제외한 유사도 분석을 수행하여 총 482,533개의 결과가 나타났으며, 비교 불가인 경우를 제외하여 총 475,851개의 결괏값이 나타났고, F_3의 유사도 분석 결과를 나타낸 그림은 Fig. 7과 같다.

게시물 사진의 경우 각 계정 조합의 유사도 평균값을 사용하여 분석한다. 게시물 사진의 유사도가 0.2-0.6 사이 값인 경우 Label 값이 불규칙적으로 분포되어 있는 것으로 보여지므로 F_3은 동일 계정 판별 유효성이 낮은 것으로 판단된다.

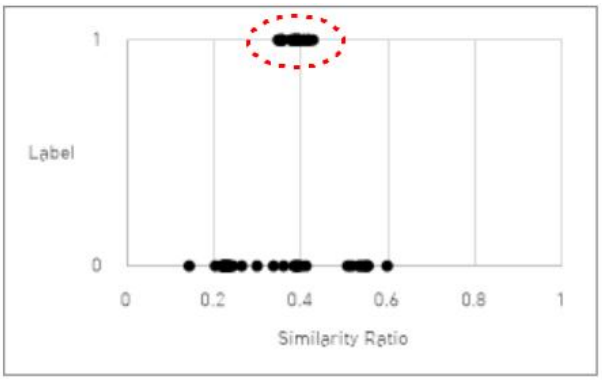

Fig. 7. Determining Same Drug Sales Accounts by Analyzing Posts in Image Form

F_4에 대해 비교 분석한 결과 총 480,290개의 결과가 보여진다. Fig. 8에서, 코사인 유사도로 분석하였을 때 유사도 값이 높을수록 Label 값도 1이 되는 것을 알 수 있고 유사도가 낮을수록 label의 값은 0으로 나타난다는 것을 확인할 수 있다. 반대로 Fig. 9를 보면 F_4를 자카드 유사도로 분석하였을 경우, 유사도가 낮은 경우에도 예측값은 1로 나타나는 것이 보여진다. F_2의 자카드 유사도 분석 결과처럼, F_4의 자카드 유사도 분석 또한 동일 계정 판별에 유효성이 떨어진다는 것을 확인할 수 있다.

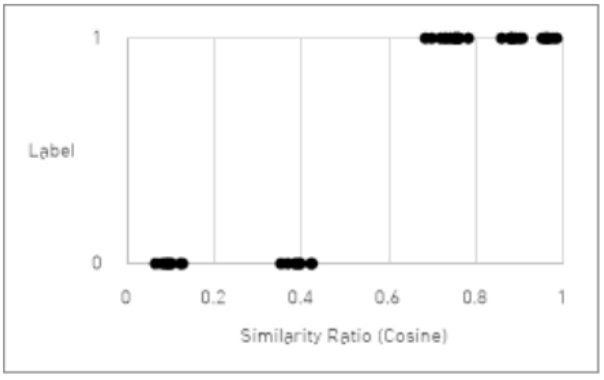

Fig. 8. Determining Same Drug Sales Accounts by Analyzing Posts in Text Form using the Cosine Similarity Technique

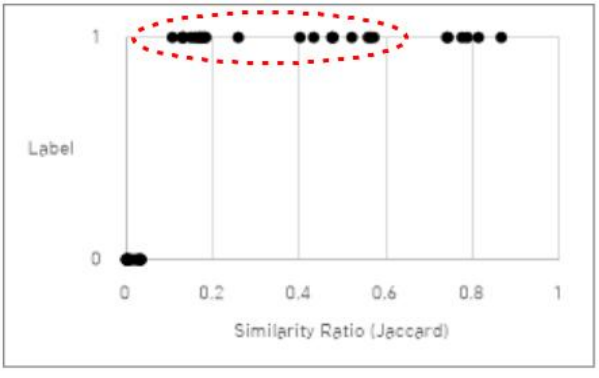

Fig. 9. Determining Same Drug Sales Accounts by Analyzing Posts in Text Form using the Jaccard Similarity Technique

게시물 사진의 유사도 분석을 수행한 후 F_3과 F_4의 유사도가 모두 1인 조합은 [U_7, U_9], [U_9, U_{12}] 두 관계가 나타난다. [U_7, U_9]의 경우 Fig. 10과 같이 게시물의 다수가 같은 사진으로 보이며, 게시글에서도 해시태그와 글자 배열이 동일한 것을 알 수 있다. [U_9, U_{12}]의 경우 또한 Fig. 11을 통해 유사도가 1인 게시물 사진과 게시글이 모두 동일함을 알 수 있다.

Fig. 10. Comparison between Posts of U_7 and U_9

Fig. 11. Comparison between Posts of U_9 and U_{12}

4.3 Importance Analysis of Instagram Elements using Logistic Regression

본 절에서는 앞서 시행한 정보별 유사도 분석 결과에 따라 F_1-F_4 중 어떤 정보가 동일 마약 유통자 판별에 효과를 미치는지 요인 분석을 시행하기 위하여 로지스틱 회귀분석을 사용한다. 로지스틱 회귀분석이란 어떠한 현상이 발생할 확률을 추정하는 것으로 통계 기법을 사용하여 독립 변수를 통해 종속 변수를 예측한다.

본 논문에서는 독립 변수를 F_1-F_4의 유사도 비교 결괏값으로 사용하였고, 종속 변수는 앞선 전수조사 결과를 기

반으로한 동일 마약 판매 계정 판별 여부를 설정한다. 로지스틱 회귀분석 결과는 Table 4와 같다.

유의 확률(P-value)은 0.05보다 작은 경우 동일 인물 판별에 유효하며 0.001보다 작은 경우에는 유효성이 더욱 높아짐을 의미한다. 계수(Coefficient)가 양수일 때의 독립변수는 유사도 값이 증가할수록 동일 마약 유통자일 확률도 증가한다는 의미이고, 계수가 음수인 경우에는 유사도 값이 증가하여도 그 값이 유효하지 않다는 것을 의미한다. 따라서 로지스틱 회귀분석 결과로 보았을 때, 프로필 사진의 SSIM 유사도와 소개글의 코사인 유사도, 자카드 유사도 및 게시글의 코사인 유사도가 유의확률 0.001 미만인 것으로, 유의미한 정보임을 알 수 있다. 그러나 앞서 시행한 유사도 분석에서 F_2, F_4의 자카드 유사도 부분 및 F_3의 SSIM 유사도에서 동일 계정 판별 유효성이 낮은 것으로 판단되었으므로, 최종적인 동일 마약 유통자 판별 정보는 F_2, F_4의 코사인 유사도, F_1의 SSIM 유사도임을 알 수 있다.

Table 4. Analysis Results of Logistic Regression

Independent variable			Coefficient	Standard error	P-value
Profile	Image		0.302724	0.038462	***
	Bio	Cosine	0.579267	0.072348	***
		Jaccard	0.637418	0.135968	***
Post	Image		1.073093	0.579443	0.068649
	Text	Cosine	1.671529	0.148139	***
		Jaccard	-0.353151	0.170491	*

*P<0.05, **P<0.01, ***P<0.001

추가적으로 인스타그램 계정의 프로필 사진만으로 마약 판매계정을 추적하는 기존 연구[24]와의 성능을 비교하기 위하여 프로필 사진을 독립 변수로 설정하여 로지스틱 회귀분석을 수행한 결과, Table 5와 같이 유의 확률이 약 8.92로 유의하지 않은 것으로 분석되었다.

Table 5. Analysis Results of Logistic Regression for Profile Images

Independent variable	Coefficient	Standard error	P-value
Profile Image	0.609306	0.054039	8.926879

이는 Table 3의 동일 마약 판매계정 검증을 위한 전수조사 시, 프로필 사진, 소개글, 게시물 사진, 게시글을 종합적으로 판단하였기 때문이며 다양한 요소를 통해 마약 판매계정을 추적하는 방법이 더욱 효과적임을 의미한다.

V. Conclusion

소셜 미디어에서 발생되고 있는 사이버 범죄가 확산됨에 따라 이를 예방하고 대응하기 위한 수사 기술 개발이 필요하다. 특히 마약 범죄는 사회적 안녕과 질서에 악영향을 미칠 수 있을 뿐 아니라, 마약 사범의 연령대가 지속적으로 낮아지고 있으므로 소셜 미디어에서 행해지는 마약 범죄 콘텐츠를 추적할 수 있어야 한다. 본 논문에서는 대중적인 소셜 미디어인 인스타그램 내에서 동일 마약 유통자를 추적하기 위한 모델을 설계하였고, 자체 개발한 크롤러를 이용하여 구축한 데이터베이스를 기반으로 마약 판매 계정 간의 유사도 분석을 수행하였다. 인스타그램의 여러 정보 중, 프로필 사진, 소개글, 게시물 사진, 게시글을 수집하여 계정 간의 유사도를 분석한 결과, SSIM 기반의 프로필 사진 유사도가 증가할수록, 동일 마약 판매 계정일 확률이 높아지는 것을 확인하였다. 또한, 프로필 소개글 및 게시글의 경우, 자카드 유사도 분석보다는 코사인 유사도 분석이 동일 마약 판매 계정 판단에 효과적이고, 게시물 사진의 경우에는 SSIM 기반 유사도 수치가 0에 가까운 것으로 보아 유효성이 낮다고 판단하였다. 이후, 동일 마약 판매 계정을 추적하는 데에 유의한 정보를 도출하기 위하여 로지스틱 회귀분석을 수행한 결과, SSIM 기반의 프로필 사진 유사도 분석, 코사인 기반의 프로필 소개글 및 게시글 유사도 분석 방법은 유의 확률 0.001 이하로 매우 유의한 분석 방법으로 검증되었다.

본 연구 결과는 인스타그램뿐 아니라, 다양한 소셜 미디어 내에서 이루어지는 마약 유통 범죄를 효과적으로 추적 및 검거하는 데에 활용될 수 있으며 불법 촬영물 유포, 무기 거래 등 다양한 사이버 범죄 추적에서도 활용될 수 있다.

향후에는 본 논문에서 선정한 인스타그램의 4개 수집 정보 외에도 팔로잉, 팔로워와 같은 유사도 분석에 유효한 추가 정보를 도출하고, 의도적으로 추적을 우회하기 위해 사진이나 글을 변형하는 경우 또한 고려하여 사이버 수사 기술의 정확성을 높이기 위한 개선연구를 수행할 계획이다.

ACKNOWLEDGEMENT

This work was supported by 'Tech. Challenge for Future Program Policing([http://www.kipot.or.kr)' /]www.kipot.or.kr)' funded by Ministry of Science and ICT(MSIT, Korea) & Korean National Police Agency(KNPA, Korea). [Project Name : Development of

Active Dark Web Information Collection, Analysis and Tracking Technology to Prevent Dark Web Cri me / Project Number : RS-2023-00244362]

REFERENCES

[1] DataReportal, Meltwater, and We Are Social. Number of internet and social media users worldwide as of October 2023 in billions, Statista, *https://www.statista.com/statistics/617136/digital-population-worldwide/*

[2] Thomas Brewster, “The DEA Quietly Turned Apple’s AirTag Into A Surveillance Tool”, Forbes, *https://www.forbes.com/sites/thomasbrewster/2023/03/23/apple-airtag-becomes-dea-surveillance-device/?sh=4b7c22f33d3d*

[3] Marco Marcelline, “DEA Uses Apple AirTag to Track Drug-Related Shipment From China”, PCMag, *https://www.pcmag.com/news/feds-used-an-apple-airtag-to-stalk-suspects*

[4] J.H Lee, "[Exclusive] The Supreme Prosecutors' Office and th e Ministry of Justice promote the establishment of a 'Drug a nd Organized Crime Investigation Agency'", legal newspapers, *https://www.lawtimes.co.kr/news/154428?serial=154428*

[5] Lee Changho, Lee Kyungsang, Kim Namdoo, “A Study on the Status of Youth Media Use and Policy Response by Target III: Late Adolescents”, Vol. 22, No. 05, pp. 114, 2022.

[6] ECRM, "Cybercrime classification", ECRM, *https://ecrm.police.go.kr/minwon/crs/quick/cyber1*

[7] Moon, Hyun Ji, Park, Hyun Min, and Kim, Gi Bum, "A Study of Slang of Voice Phishing Criminal Organization and Practical Use for Investigation," The Journal of Police Science, Vol. 21, No. 4, pp. 137-160, 2021.

[8] F. Iqbal, B. C. M. Fung, M. Debbabi, R. Batool and A. Marrington, "Wordnet-Based Criminal Networks Mining for Cybercrime Investigation," in IEEE Access, Vol. 7, pp. 22740-22755, 2019, doi: 10.1109/ACCESS.2019.2891694.

[9] Yae Eun Lee, and Jeong-hyeon Chang, "Child Abuse Analysis and Keyword Extraction through Unstructured Data Collection and TF-IDF," Korean Criminal Psychology Review, Vol. 18, No. 4, pp. 171-182, 2022.

[10] HyunChul Kim, and JiWon Yoon, "A Case of Cyber Financial Crime Investigation Through Social Network Analysis (2-Mode Concepts)," Journal of Digital Forensics , Vol. 14, No. 4, pp. 449-465, 2020.

[11] R. Allsup, E. Thomas, B. Monk, R. Frank and M. Bouchard, "Networking in child exploitation – Assessing disruption strategies using registrant information," 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Paris, France, 2015, pp. 400-407, doi: 10.1145/2808797.2809297.

[12] E. Nunes et al., "Darknet and deepnet mining for proactive cybersecurity threat intelligence," 2016 IEEE Conference on Intelligence and Security Informatics (ISI), Tucson, AZ, USA, 2016, pp. 7-12, doi: 10.1109/ISI.2016.7745435.

[13] Jiwon Jeon, Seunghwan Lee, Junghyun Lee, Hongju Kim, and Taeshik Shon, "Analysis and Development of Web Archiver for leveraging cybercrime investigations," Journal of Digital Forensics , Vol. 15, No. 3, pp. 139-149, 2021.

[14] Ho-Mook Cho, Chang-Su Bae, Jaehoon Jang, and Sang-Yong Choi, "The Next Generation Malware Information Collection Architecture for Cybercrime Investigation," Journal of the Korea Society of Computer and Information , Vol. 25, No. 11, pp. 123-129, 2020.

[15] Kang Jeong-hee, "Web Crawling Data Collection and Review from a Perspective of Competition Law - Focused on Supreme Court Decision 2021Do1533 Decided May 12, 2022," JURIS, Vol. 1, No. 61, pp. 461-500, 2022.

[16] Lee-JuHyeok, and Kim-Mi Hui, "Image Classification Model using web crawling and transfer learning," Journal of IKEEE, Vol. 26, No. 4, pp. 116-123, 2022.

[17] Kim, Junsung, "A Study on Contractual Restrictions of Data Crawling," LAW REVIEW, Vol. 63, No. 4, pp. 221-248, 2022.

[18] P. Koloveas, T. Chantzios, C. Tryfonopoulos and S. Skiadopoulos, "A Crawler Architecture for Harvesting the Clear, Social, and Dark Web for IoT-Related Cyber-Threat Intelligence," 2019 IEEE World Congress on Services (SERVICES), Milan, Italy, 2019, pp. 3-8, doi: 10.1109/SERVICES.2019.00016.

[19] Shin, Sung-Mi, and Kwon, Kyoung-In, "Text network analysis of Instagram posts with self-injury," Korea Journal of Counseling, Vol. 20, No. 6, pp. 273-295, 2019.

[20] N. S. Purohit, A. B. Angadi, M. Bhat and K. C. Gull, "Crawling through web to extract the data from Social networking site - Twitter," 2015 National Conference on Parallel Computing Technologies (PARCOMPTECH), Bengaluru, India, 2015, pp. 1-6, doi: 10.1109/PARCOMPTECH.2015.7084522.

[21] A. Dalvi, P. Kulkarni, A. Kore and S. G. Bhirud, "Dark Web Crawling for Cybersecurity: Insights into Vulnerabilities and Ransomware Discussions," 2023 2nd International Conference for Innovation in Technology (INOCON), Bangalore, India, 2023, pp. 1-6, doi: 10.1109/INOCON57975.2023.10101162.

[22] Hyun-Jin Yeo, "Mobile Commerce Brand Identity Strategy by SNS Text mining," Journal of the Korea Society of Computer and Information , Vol. 25, No. 10, pp. 255-260, 2020.

[23] A. Tundis, G. Bhatia, A. Jain and M. Mühlhäuser, "Supporting the Identification and the Assessment of Suspicious Users on Twitter Social Media," 2018 IEEE 17th International Symposium on Network Computing and Applications (NCA), Cambridge, MA, USA, 2018, pp. 1-10, doi: 10.1109/NCA.2018.8548321.

[24] Eun-Young Park, Kyeong-Hyun Cho, Jiyeon Kim, and Chang-Hoon Kim, "Tracking Drug Distribution Accounts Through Similarity Analysis of Instagram Profile Photos," Proceedings of the Korean Society of Computer Information Conference , pp. 199-201, 2023.

Authors

Eun-Young Park is an undergraduate student in the Department of Computer Engineering, Daegu University, Gyeongsan, Korea, since 2021. Her research interests include cybersecurity, internet of things, and artificial intelligence.

Jiyeon Kim received the B.S. and Ph.D. degrees in information security engineering from Seoul Women's University, Seoul, South Korea, in 2007 and 2013, respectively. Dr. Kim was a Postdoctoral Research Associate in the Department of Electrical and Computer Engineering, Carnegie Mellon University, United States, from 2014 to 2017. She is currently an Assistant professor in the Department of Computer Engineering, Daegu University, Gyeongsan, South Korea. Her research interests include cybersecurity, cybercrime investigation, cloud computing, artificial intelligence, and critical infrastructure protection.

Chang-Hoon Kim is a professor at Daegu University, Republic of Korea. He earned his bachelor of computer science engineering in 2001, a master degree of computer and information engineering in 2003 and a Ph. D in computer science engineering in 2006, both at Daegu University in the Republic of Korea. Professor Kim has published many papers in international and domestic journals and has attended many international and domestic conferences for presen-tations. His research area includes network security, system security and artificial intelligence.