KIPS Trans. Comp. and Comm. Sys.
Vol.12, No.1 pp.17~24
ISSN: 2287-5891 (Print), ISSN 2734-049X (Online)

https://doi.org/10.3745/KTCCS.2023.12.1.17

Reliability Analysis of Privacy Policies Using Android Static Analysis

Yoonkyo Jung†

ABSTRACT

Mobile apps frequently request permission to access sensitive data for user convenience. However, while using mobile applications, sensitive and personal data has been leaked even if users do not allow it. To deal with this problem, Google App Store has required developers to disclose how the mobile app handles user data in a privacy policy. However, users are not certain that the privacy policy describes all the app's behavior. They have no choice but to rely on the privacy policy to confirm how the app uses data. This study designed a system that checks the reliability of privacy policies by analyzing the privacy policy texts and mobile apps. First, the system extracts and analyzes the privacy policy texts to check which personal data the privacy policy discloses that the mobile apps can collect. After analyzing which data apps can access using android static analysis, we compare both results to analyze the reliability of privacy policies. For the experiment, we collected the APK files and metadata of about 13K android apps registered in the Google Play Store and preprocessed the apps by four conditions. According to the comparison between privacy policies and mobile app behavior, many apps can access more personal data than disclosed in the privacy policy.

Keywords : Privacy Policy, Static Analysis, Data Privacy, Android Application

안드로이드 정적 분석을 활용한 개인정보 처리방침의 신뢰성 분석

정 윤 교†

요 약

모바일 앱은 사용자의 편의를 위해 개인정보에 접근할 수 있는 권한을 자주 요청한다. 하지만 이에 따라 모바일 앱을 이용하는 동안 허용되지 않은 개인정보가 유출되는 문제가 많이 발생했다. 이러한 문제를 해결하기 위해 구글 앱스토어에 등록된 앱은 개인정보 처리방침에 사용자의 개인정보를 앱에서 어떻게 활용하는지 명시하도록 했다. 하지만 앱이 수행하는 개인정보 수집 및 처리 과정이 개인정보 처리방침에 정확히 공개되어 있는지 확인하기 어려우며, 모바일 앱 사용자가 앱이 접근할 수 있는 개인정보에 대해 알기 위해서는 개인정보 처리방침에 의존해야만 한다. 본 연구에서는 개인정보 처리방침과 모바일 앱을 분석하여 개인정보 처리방침의 신뢰성을 확인하는 시스템을 제시한다. 먼저 개인정보 처리방침의 텍스트를 추출 및 분석하여 모바일 앱이 어떤 개인정보를 이용할 수 있다고 공개하는지 확인한다. 이후 안드로이드 정적 분석을 통해 앱이 접근할 수 있는 개인정보 분류를 확인하고, 두 결과를 비교하여 개인정보 처리방침을 신뢰할 수 있는지 분석한다. 실험을 위해 구글 앱스토어에 등록된 약 13,000개 안드로이드 앱의 패키지 파일과 부가정보를 수집한 뒤 분석할 수 있는 앱을 선정하기 위해 4가지 조건에 따라 전처리를 진행했다. 선정한 앱을 대상으로 텍스트 분석과 모바일 앱 분석을 진행하고, 이를 비교하여 모바일 앱은 개인정보 처리방침에 공개한 것보다 더욱 많은 개인정보에 접근할 수 있음을 증명한다.

키워드 : 개인정보 처리방침, 정적 분석, 개인정보보호, 안드로이드 앱

1. 서 론

모바일 기기는 일상생활에 중요한 요소로 자리 잡고 있다. 특히 스마트폰이 대중화되면서 모바일 환경에서 인터넷을 사용하는 비율이 점점 높아지는 추세이다[1]. 하지만, 모바일 사용자의 증가에 따라 모바일 기기를 사용하는 동안 개인정보가 유출되는 사례가 빈번하게 일어나고 있다[2]. 모바일 앱은 사용자의 편의를 위해 개인정보를 다룰 수 있지만, 앱 서비스 제공자의 이익을 취하기 위해 특정 사용자를 대상으로 하는 맞춤형 광고 등 개인정보를 다른 목적으로 사용할 수 있다. 모바일 기기는 위치 정보 및 기기 식별자 등 개인을 식별할 수 있는 정보를 많이 포함하고 있어[3] 앱을 통해 개인정보가 유출되지 않도록 각별한 주의가 필요하며, 앱은 사용자의 개인정보보호를 위해 최소한의 권한만 요구해야 한다.

※ 이 논문은 2022년 한국정보처리학회 ASK 2022에서 "안드로이드 정적분석 기반 개인정보 처리방침의 신뢰성 분석"의 제목으로 발표된 논문을 확장한 것임.
† 종신회원 : 공군사관학교 컴퓨터과학과 조교수
Manuscript Received : August 1, 2022
First Revision : September 20, 2022
Accepted : October 11, 2022
* Corresponding Author : Yoonkyo Jung(ykjung.rokafa@gmail.com)

구글은 모바일 앱의 투명성을 확보하고 사용자의 개인정보를 보호하기 위해 사용자 데이터 정책[4]을 발표했다. 구글 데이터 정책에 따르면, 안드로이드 앱스토어에 등록한 앱의 개발자들은 개인정보 수집과 관련한 모든 앱의 활동을 개인정보 처리방침에 공개하여 앱이 사용자로부터 수집하는 정보와 활용 목적에 관해 설명해야 한다. 또한, 개인정보 처리방침에 명시한 내용은 유럽 일반 개인정보 보호법(GDPR)[5], 캘리포니아 소비자 프라이버시 보호법(CCPA)[6], 미 연방 프라이버시보호 법안(ADPPA)[7] 등 데이터 보호와 관련한 많은 규정들을 모두 만족해야 한다.

앱이 수집할 수 있는 개인정보를 제한하여 사용자를 보호하기 위해 기존에는 과도한 권한 요구를 예방하거나[8, 9] 권한 매커니즘을 분석하는 연구가 진행되었다[10, 11]. 최근 출시된 안드로이드 앱은 사용자에게 필수적인 권한만 요청하고 있으며, 앱을 처음 실행할 때 사용자가 직접 권한을 허용하거나 거부할 수 있다. 하지만 개인정보 접근 권한을 부여받은 앱이 실제로 어떤 정보에 접근하는지 확인하기 위해서는 개인정보 처리방침에 의존할 수밖에 없다. 개인정보 처리방침이란 기관 및 기업 등에서 업무를 위해 다루는 사용자 정보의 목록과 사용 목적을 공개하는 것으로, 서비스 사용자가 개인정보 활용에 관해 확인할 수 있는 유일한 수단이다. 하지만 개인정보 처리방침에 대한 투명성을 검증하기 어려워 앱이 공개하지 않은 정보수집 활동을 하더라도 사용자는 알 수 없다. 플레이스토어에 있는 앱 설명이 권한 요청에 관한 내용을 충분히 포함하고 있는지 확인하기 위해 앱 설명을 분석하여 권한과 앱 설명의 일관성을 평가하였으나[12], 앱의 권한만 조사했고 텍스트 분석 후 실제 앱의 패키지 파일 분석은 진행되지 않았다. 사용자의 개인정보보호에 위협이 되는 항목을 확인하기 위해 모바일 기기에 설치된 앱을 분석하는 연구가 있었으나[13], 개인정보 처리방침과 관련한 내용은 제외되어 데이터 수집 활동이 공개되어 있는지 확인하기 어려웠다. 따라서 모바일 사용자에게 앱의 개인정보 수집 및 활용 목적에 대한 명확한 정보를 제공하기 위해 개인정보 처리방침을 신뢰할 수 있는지 증명하기 위한 연구가 필요하다.

본 논문에서는 안드로이드 정적 분석을 바탕으로 모바일 앱 활동 결과와 개인정보 처리방침의 텍스트 분석을 통해 앱이 갖는 기능이 개인정보 처리방침에 명시되어 있는지 확인하는 시스템을 제시한다. 이를 신뢰성 분석이라고 정의하였으며, 사용자가 개인정보 처리방침을 신뢰할 수 있는지 확인하기 위해 시스템을 설계하고 실험을 진행한다. 설계한 시스템은 구글 플레이스토어에서 수집한 APK 파일을 기반으로 모바일 앱 분석을 진행하여 앱이 실제로 어떤 정보를 수집할 수 있는지 확인하고, 개인정보 처리방침의 텍스트를 추출한 뒤 내용을 분석한다. 이후 각각의 결과를 비교하여 개인정보 처리방침이 사용자에게 개인정보 수집 및 처리 활동에 대해 정확히 공개하고 있는지 입증한다. 실험을 통해 개인정보 처리방침에 공개한 내용보다 앱이 개인정보에 접근할 수 있는 기능이 더 많다는 것을 보여준다.

2. 관련 연구

본 연구의 기반이 되는 개인정보 처리방침 분석 및 모바일 앱 분석과 관련한 선행 연구를 요약한다.

개인정보 처리방침은 사용자가 앱의 데이터 수집 활동을 확인할 수 있는 중요한 자료이다. 하지만 모든 텍스트를 일일이 확인하는 것은 상당한 시간을 요구한다[14]. 이러한 문제를 해결하기 위해 기존 연구에서는 자동으로 개인정보 처리방침을 분석하는 방법을 제시했다. 텍스트를 수동으로 분류하여 말뭉치를 제작했고[15, 16], 개인정보 처리방침의 구조를 분석하는 방법을 제시했다[17]. 웹 페이지의 텍스트를 추출한 결과 개인정보 처리방침에 제 3자의 데이터 수집 활동이 명확히 공개되지 않았음을 확인했으며[18], GDPR이 규정을 바탕으로 개인정보 처리방침의 투명성을 단계별로 도출했다[19]. 하지만 개인정보 처리방침 분석에 대한 기존 연구들은 텍스트 분석 기법을 중심으로 진행되었으며, 실험을 통해 얻은 결과는 개인정보 처리방침의 내용에만 한정되어 있다. 따라서 앱의 실제 활동에 대한 분석을 포함하여 개인정보 처리방침의 신뢰성 분석을 진행하기 위해 모바일 앱 분석에 대한 사전 연구를 진행했다.

모바일 앱 분석을 위한 방법으로 크게 정적 분석과 동적 분석이 있다. 정적 분석은 앱을 직접 실행하지 않고 소스 코드만 확인하는 방법으로, 앱이 실행되었을 경우 API가 호출되어 데이터 수집이 가능함을 보여준다. 기존 연구에서는 앱의 권한과 관련한 시스템 호출을 확인하기 위해 정적 분석을 수행했으며[20, 21], 그 결과 정보 흐름에서 여러 잠재적인 정보 유출이 발견되었다[22, 23]. 또한 정적 분석의 결과를 4가지로 나누어 악성 소프트웨어 검출에 활용하는 기법도 제시되었다[24]. 이처럼 정적 분석 기법을 기반으로 진행한 연구는 소스 코드를 분석하여 앱이 수집할 수 있는 정보에 관해 확인하고, 정보의 흐름을 확인하여 잠재적인 위험이 있는 앱을 검출하는 연구가 대부분이다.

이와 달리, 동적 분석은 앱이 동작하는 동안 발생하는 네트워크 흐름이나 API를 분석하여 개인정보의 수집 활동을 추적한다. 이를 활용하여 사용자 정보의 잠재적인 유출을 동적으로 추적하고[25], 분석 도구를 개발하여 모바일 앱의 트래픽을 모니터링했다[26]. 동적 분석을 통해 많은 안드로이드 앱이 광고 목적으로 사용이 금지된 기기 식별자 정보를 주기적으로 수집하는 것을 확인했으며[27], 분석 과정에서 악성 앱의 숨겨진 컴포넌트를 검출하는 방법을 설계하는 등[28] 정적 분석으로 확인하기 어려운 앱의 활동을 검출하는 연구가 진행되었다. 동적 분석은 앱의 숨겨진 기능을 확인할 수 있다는 장점이 있으나, 실험 규모를 키우기 어렵고 분석을 위해 임의로 발생한 앱 이벤트에 따라 결과가 달라질 수 있다는 단점이 있다. 기존의 앱 분석 관련 연구 중 개인정보 처리방침과 비교한 연구는 부족했으며, 많은 앱을 대상으로 명확한 결과를 얻기 위해 정적 분석을 기반으로 연구를 진행했다.

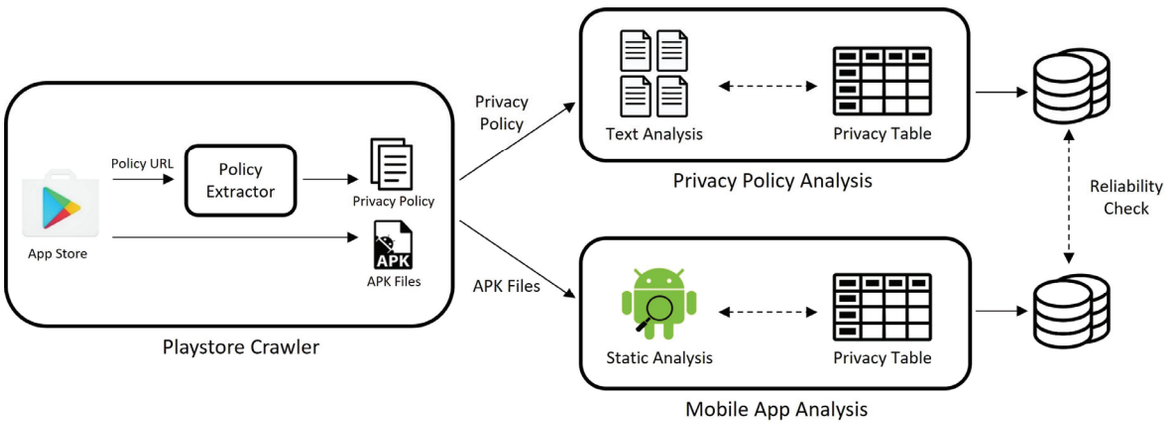


Fig. 1. System Design

3. 시스템 설계

사용자가 개인정보 처리방침을 신뢰할 수 있는지 확인하기 위해 개인정보 처리방침과 모바일 앱을 분석하여 결과를 비교하는 시스템을 설계했다. Fig. 1은 시스템의 전반적인 구조를 보여준다. 먼저 모바일 앱 스토어에 등록된 앱의 부가정보를 바탕으로 개인정보 처리방침의 텍스트와 앱의 APK 파일을 수집한다. 텍스트 분석 결과를 비교하여 개인정보 처리방침에 개인정보와 관련 있는 키워드가 포함되어 있는지 확인하고, 앱의 소스 코드에 포함된 API를 비교하여 앱이 어떤 개인정보에 접근할 수 있는지 확인한다. 텍스트 및 정적 분석 결과를 비교하기 위해서 개인정보와 관련된 키워드 및 API 목록을 조사한 개인정보 테이블[29]을 활용된다. 이후 두 결과를 비교하여 개인정보 처리방침의 신뢰성을 확인한다.

3.1 개인정보 처리방침 분석

개인정보 처리방침 분석 과정에서는 앱 스토어의 부가정보에 포함된 개인정보 처리방침 웹 주소를 활용하여 텍스트를 추출 및 분석한다. 앱마다 웹 주소의 형식이 달라 텍스트를 추출하기 어렵기 때문에, 자동으로 오류를 처리하고 자동으로 텍스트를 추출하는 텍스트 추출기를 만들었다.

Fig. 2는 개인정보 처리방침의 텍스트를 추출하기 위한 과정이다. 먼저 앱의 부가정보에 포함된 웹 주소를 확인하고 크롬 헤드리스 브라우저로 이를 실행한다. 실행된 페이지는 Readability 라이브러리[30]를 거쳐 읽기 모드로 변환 후 텍스트만 추출된다. 텍스트 추출기는 개인정보 처리방침 웹 주소에 페이지 주소가 아닌 첨부파일이 있을 경우 파일을 자동으로 실행하여 파일 형식을 확인 후 텍스트를 추출한다. 또한 웹 주소가 없거나 유효하지 않다면 예외 처리를 진행했다.

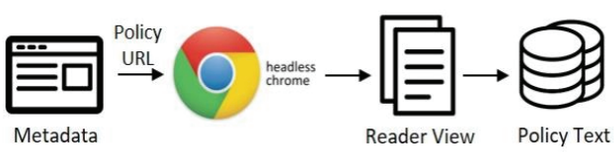


Fig. 2. Policy Extracting Phase

추출한 텍스트를 분석하기 위해 앱 스토어에 등록된 인기 앱들의 개인정보 처리방침을 참고하여 15개의 개인정보 분류를 대표하는 키워드를 선정했다. 50개 앱을 샘플링 후 정밀도, 재현율, F1 점수 측면에서 키워드의 실효성을 확인한 결과 Table 1의 결과를 얻었다.

정밀도와 재현율은 Equation (1)의 방식으로 확인했으며, F1 점수는 정밀도와 재현율의 조화 평균으로 구할 수 있다.

$$\begin{aligned} precision &= TP/(TP+FP) \\ recall &= TP/(TP+FN) \end{aligned} \tag{1}$$

Table 1. Comparison of Precision, Recall, and F1 Score for Each Privacy Category

Category	Precision	Recall	F1 score
Contact	0.625	0.769	0.690
Email	0.833	0.926	0.877
Phone Number	0.870	0.833	0.851
Unique Identifier(ID)	0.875	0.848	0.862
Cookie	0.895	0.850	0.872
Android Id	0.818	0.818	0.818
IMEI	0.923	0.857	0.889
IMSI	0.778	0.875	0.824
MAC	0.778	0.700	0.737
Mobile Carrier	0.857	0.750	0.800
SIM Serial	0.800	0.667	0.727
SSID and BSSID	0.833	0.714	0.769
Location	0.833	0.789	0.811
Cell Tower	0.833	0.714	0.769
GPS and WiFi	0.667	0.800	0.727
Average	**0.815**	**0.794**	**0.802**

True Positive(TP)는 키워드를 분석하여 확인한 개인정보 분류가 텍스트 내용에 포함될 경우, True Negative(TN)는 분석 결과 및 텍스트에 개인정보 분류가 없을 경우, False Positive(FP)는 분석 결과가 실제로도 개인정보 분류와 관련 없을 경우, False Negative(FN)는 결과에는 없지만 실제로 특정 개인정보 관련 내용이 존재할 경우이다. 선정한 키워드를 사용한 패턴 매칭을 통해 모바일 앱이 개인정보 처리방침에 어떤 개인정보를 이용할 수 있다고 공개하는지 확인했다.

3.2 모바일 앱 분석

모바일 앱 분석은 정적 분석을 통해 앱의 API 목록을 확인하여 앱이 접근할 수 있는 개인정보 분류를 확인하고, 이를 개인정보 처리방침 분석 결과와 비교한다.

안드로이드 정적 분석을 수행하기 위해 파이썬 기반의 오픈 소스 정적 분석 도구인 Androguard[31]를 활용한다. 안드로이드 공식 설명서에는 안드로이드 API의 클래스와 메소드 목록이 공개되어 있다. 정적 분석으로 앱이 가진 기능을 확인하기 전에, 안드로이드 설명서를 분석하여 개인정보를 호출하는 API 목록을 조사하고 개인정보 수집과 관련된 API를 찾아 개인정보 테이블에 목록을 저장했다.

안드로이드 설명서의 분석 결과를 바탕으로 각각의 API가 어떤 정보를 수집하는지 조사하여 Table 2와 같이 저장했다. 특정 개인정보를 호출할 수 있는 모든 API를 민감 API라고 정의했으며, 구글 사용자 데이터 정책을 바탕으로 분류한 총 15가지 개인정보 항목에 대해 31개의 민감 API 목록을 선정했다. 선정된 목록은 앱이 개인정보를 수집하는 코드를 포함하고 있는지 분석하기 위해 사용된다.

모바일 앱 분석 과정은 APK 파일을 정적 분석한 뒤 이를 민감 API 목록과 비교하는 순서이며 전반적인 과정은 Fig. 3의 예시와 같이 진행된다. 시스템은 정적 분석으로 API를 확인하고, 이를 민감 API 목록과 비교하여 앱을 통해 수집할 수 있는 개인정보 분류를 확인한다. 앱이 민감 API를 포함한다면 해당 API가 호출하는 개인정보를 수집할 수 있다고 간주한다. 앱 분석을 통해 확인한 결과를 개인정보 처리방침 분석 결과와 비교하여 앱에 포함된 기능이 개인정보 처리방침에 공개되었는지 확인하는 과정을 거쳐 신뢰성을 분석한다.

Table 2. An Example of Sensitive API Lists

Category	Sensitive API
IMEI	android.telephony.TelephonyManager.getDeviceId
	android.telephony.TelephonyManager.getImei
GPS and WiFi	FusedLocationProviderClient.getLastLocation
	android.location.LocationManager.requestLocationUpdates
	android.location.LocationManager.requestSingleUpdate
	android.location.LocationManager.getLastKnownLocation

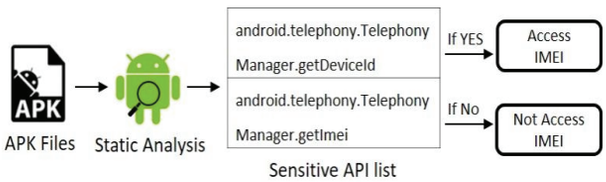


Fig. 3. A Process of Mobile App Analysis

4. 결과 분석

실험을 위해 구글 앱스토어 내 인기 순위를 바탕으로 구글 앱스토어에 등록된 13,223개 모바일 앱의 패키지(APK) 파일과 앱 부가정보를 수집했다. 데이터의 수집은 2021년 2월부터 약 한 달간 진행되었으며 수집된 APK 파일과 앱 부가정보을 Ubuntu 20.04 서버에 저장하여 실험을 진행했다. 부가정보는 NoSQL 데이터베이스 시스템인 MongoDB를 사용하여 저장했다. 수집한 부가정보 및 실험 결과에 대한 데이터는 오픈소스 페이지[29]에서 확인할 수 있다.

4.1 데이터 전처리

수집한 앱의 일부는 인터넷이 연결되지 않아 정보가 유출될 위험이 적거나 앱 부가정보에 개인정보 처리방침 웹 주소가 없는 등 분석을 수행하기 어려운 경우가 존재한다. 이처럼 모든 앱의 개인정보 처리방침과 패키지 파일을 분석하는 것은 비효율적이기에 4가지 조건에 따라 전처리를 진행했다.

Fig. 4는 4가지 예외 조건에 따른 전처리 결과이다. 수집한 13,223개의 앱 중에서 인터넷 권한이 없는 앱, 개인정보 처리방침의 웹 주소가 없어 텍스트를 분석할 수 없는 앱, 개인정보 처리방침의 텍스트가 영어로 작성되지 않은 앱, 추출한 텍스트가 개인정보 처리방침으로 보기 어려운 앱을 예외 조건으로 설정하여 전처리 과정을 통해 순서대로 제외하였다. 그 결과 전체 데이터셋의 56%에 해당하는 7,401개의 앱이 실험에 사용되었다.

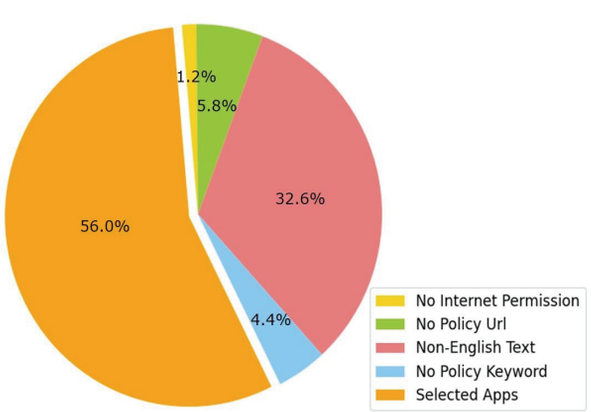


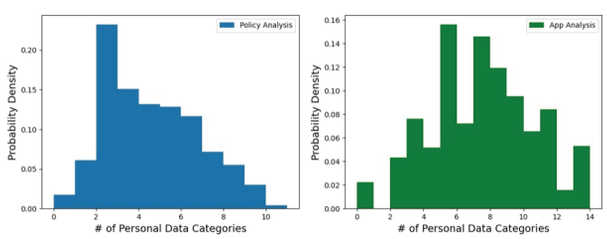
Fig. 4. A Result of Preprocessing Dataset

4.2 실험 결과

전처리한 데이터셋에 대해 개인정보 처리방침 분석과 모바일 앱 분석을 수행하여 결과를 얻고, 신뢰성 분석을 위한 비교 데이터로 사용하였다. 먼저 개인정보 처리방침 분석을 통해 각각의 앱이 개인정보 분류와 관련된 데이터를 얼마나 많이 접근한다고 명시했는지 확인하고, 모바일 앱 분석을 통해 앱이 가진 API로 접근할 수 있는 개인정보의 수를 확인했다.

Fig. 5는 개인정보 분류 수에 따라 개인정보 처리방침과 모바일 앱 분석 결과를 표현 히스토그램 그래프이다. 그래프에서 x축은 개인정보 분류 수, y축은 x축에 대한 확률 밀도를 의미한다. Fig. 5A는 각 앱이 개인정보 처리방침에 얼마나 많은 개인정보 관련 키워드를 언급했는지 확인하고 이를 히스토그램으로 나타낸 것이다. Fig. 5B는 각 앱의 코드에 포함된 API 목록이 몇 개의 개인정보 종류에 접근할 수 있는지 확인하고 각 결과에 대한 앱의 숫자를 히스토그램으로 나타낸 것이다. 그래프를 통해 모바일 앱 분석 결과가 전반적으로 더 많은 개인정보 분류와 연관이 있음을 알 수 있었고, 두 결과를 자세히 비교하기 위해 확률 밀도 함수를 그래프를 통해 차이를 확인했다.

Fig. 6은 개인정보 처리방침 분석 결과와 모바일 앱 분석 결과를 확률 밀도 함수 그래프로 나타낸 것이다. 두 그래프를 비교 및 분석한 결과 모바일 앱 분석에서 나온 개인정보 분류 수가 개인정보 처리방침 분석을 통해 얻은 수보다 더 많다는 걸 알 수 있다. 이는 앱이 개인정보 처리방침에 언급된 내용보다 더 많은 개인정보에 대해 접근할 수 있음을 의미한다.



(a) Policy Analysis (b) Mobile App Analysis

Fig. 5. Histogram Graphs for Personal Data Categories

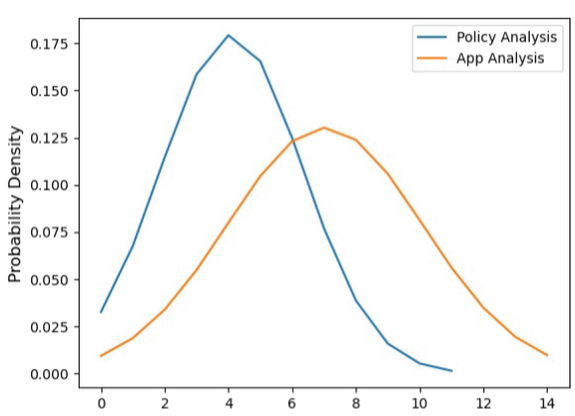


Fig. 6. PDF Graphs for Personal Data Categories

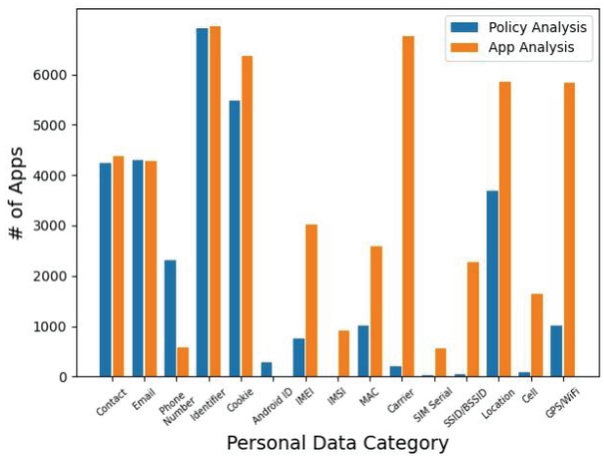


Fig. 7. Number of Apps for Each Personal Data Category

또한 개인정보 분류에 따른 결과의 차이를 비교하기 위한 분석을 진행했다. 이를 위해 실험 결과를 개인정보 분류에 따라 비교하여 실제 앱 기능과 개인정보 처리방침에 언급된 내용이 개인정보 분류에 따라 얼마나 다른지 확인했다.

Fig. 7은 개인정보 처리방침 및 모바일 앱 분석 결과를 개인정보 항목에 따라 비교한 것이다. 그래프의 Y축의 값이 클수록 패턴 매칭 결과가 높게 나타났거나 특정 개인정보와 관련한 API가 많이 확인되었음을 의미한다. 모바일 앱 분석 결과에서 개인정보 분류 관련 앱이 더 많았고, 이는 앱이 예상보다 더 많은 개인정보에 접근할 수 있다는 것을 보여준다.

다음으로, 신뢰성 확인을 위해 합집합과 교집합의 비율을 구하는 방법인 자카드 유사도를 사용했다. 자카드 유사도를 측정하는 함수 J는 Equation (2)와 같다.

$$J(A,B) = \frac{A \cap B}{A \cup B} = \frac{\text{App Analysis} \cap \text{Policy Analysis}}{\text{App Analysis} \cup \text{Policy Analysis}} \quad (2)$$

개인정보 처리방침과 모바일 앱 분석 결과에 대한 자카드 유사도를 Fig. 8과 같이 누적분포함수 그래프로 나타냈다.

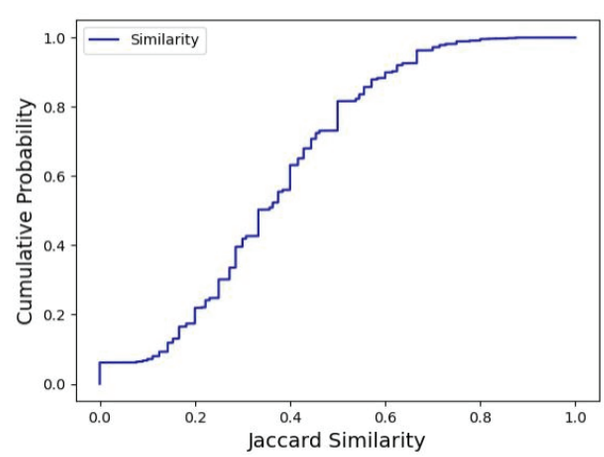


Fig. 8. Jaccard Similarity between Policy Analysis and App Analysis

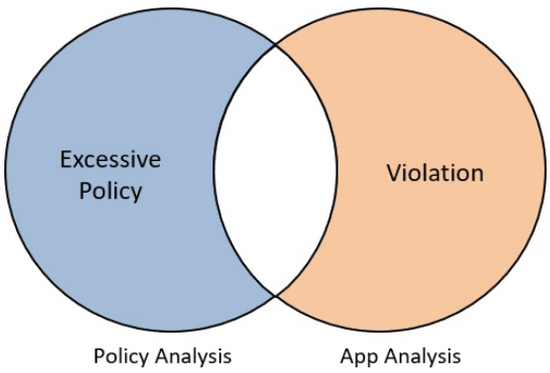


Fig. 9. Definition of Excessive Policy and Violation

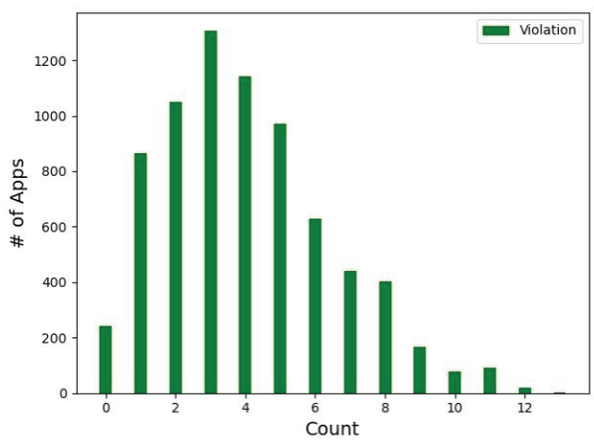


Fig. 10. Number of Apps by Violation Count

그래프에서 자카드 유사도의 x값이 0에 수렴할수록 앱의 실제 활동과 개인정보 처리방침의 내용이 일관되지 않음을 의미한다. 분석 결과에서 x값이 0.2와 0.6 사이 구간에 유사도가 많이 분포하고 있고, 0에 가까운 결과도 포함되어 일치도가 전반적으로 높지 않다는 사실을 확인할 수 있다.

또한, 개인정보 처리방침에 공개되지 않은 앱의 활동을 확인하기 위해 앱 분석 결과에는 나왔지만 개인정보 처리방침의 텍스트 분석 결과에는 포함되지 않은 개인정보 분류의 수를 확인했다. 분석을 위해 Fig. 9와 같이 앱 분석을 통해 확인한 내용이 개인정보 처리방침에 포함되지 않는 경우 정책 위반, 그 반대의 경우는 과도한 방침이라고 정의했다. 이 중에서 사용자의 개인정보보호와 특히 관련 있는 정책 위반을 중심으로 분석을 진행했다. Fig. 10은 정책 위반 횟수로 구분한 모바일 앱의 수를 나타낸다. 그래프를 통해 많은 모바일 앱에서 최소 1번 이상의 정책 위반이 확인되었음을 알 수 있다.

5. 결 론

모바일 앱을 사용하는 동안 발생하는 개인정보 유출을 방지하기 위해, 앱 제공자들은 구글 앱스토어에 앱을 등록하기 전에 앱의 개인정보 수집 및 처리 과정을 개인정보 처리방침에 모두 공개해야 한다. 하지만 개인정보 처리방침에 특정 개인정보와 관련된 앱의 실제 활동이 포함되어 있지 않더라도 사용자가 알 수 없다. 구글은 유해 앱으로부터 사용자를 보호하기 위해 앱스토어에 구글 플레이 프로텍트를 적용하여 등록된 앱에 위험 요소가 있는지 확인하고 있으나, 앱의 모든 기능을 점검하기에는 한계가 있어 사용자가 안심하고 앱을 이용하기 어렵다. 이에 따라 사용자가 직접 앱을 신뢰할 수 있는지 확인할 방안이 필요하다.

본 연구에서는 모바일 앱 사용자가 앱의 개인정보 처리방침을 신뢰할 수 있는지 확인하기 위해 안드로이드 정적 분석을 기반으로 하는 시스템을 구현했다. 실험을 위해 구글 앱스토어에 등록된 13,223개 앱의 APK 파일 및 부가정보를 수집했다. 수집한 앱의 APK 파일을 정적 분석하여 사전에 확인한 민감 API 목록과 비교하고, 앱 부가정보에서 확인한 개인정보 처리방침의 웹 주소를 이용하여 텍스트 추출 및 분석을 진행했다. 개인정보 처리방침과 모바일 앱 분석 결과를 비교하여 개인정보 처리방침에 명시한 내용보다 앱이 실제로 접근할 수 있는 개인정보가 많다는 것을 알 수 있었으며, 이는 잠재적인 정보 유출의 가능성을 시사한다. 본 연구에서 제시한 내용을 앱 사용자가 활용한다면 개인정보 처리방침의 신뢰성을 진단할 수 있고, 이를 통해 개인정보 처리방침의 투명성과 전반적인 개인정보 보호 수준을 향상할 수 있다.

하지만 본 논문은 많은 앱을 분석하고 실험 과정에서 발생하는 오류를 줄이기 위해 정적 분석을 사용했기 때문에 숨겨진 앱의 기능을 모두 확인하기 어렵다는 단점이 있다. 또한 정적 분석 도구인 Androguard는 난독화가 이루어진 일부 앱을 분석하기 어려워 실험 결과보다 실제로는 앱이 더 많은 기능을 포함할 수 있다. 이러한 한계를 개선하기 위해 프록시 서버를 사용하여 앱 사용 중에 발생하는 데이터 흐름을 파악하는 후속 연구가 필요하다. 또한, 최근 정보기술의 발전과 함께 개인정보보호의 중요성이 높아지며 여러 국가에서 정보보호를 위한 엄격한 규정들을 제정하고 있다. 또한 구글은 사용자의 개인정보를 보호하고 맞춤형 광고 제공을 목적으로 무분별하게 사용자의 개인정보를 수집하는 상황을 막기 위해 안드로이드 OS에 프라이버시 샌드박스[32]를 적용하겠다고 발표했다. 구글은 올해 말부터 새로운 기술을 적용한 시험판을 공개한 후 2024년에 이를 시행하는 것을 목표로 하고 있다. 향후 연구에는 모바일 기기의 개인정보보호와 관련된 새로운 정책들이 개인정보 처리방침에 주는 변화가 무엇인지 진단하고, 모바일 연동 측면에서 어떠한 영향이 있는지 확인하고자 한다[33].

References

[1] L. A. Mutchler, J. P. Shim, and D. Ormond, “Exploratory study on users’ behavior: smartphone usage,” in *Proceedings of Americas Conference on Information Systems*, pp.418, 2011.

[2] G. Jeon, M. Choi, S. Lee, J. H. Yi, and H. Cho, "Automated multi-layered bytecode generation for preventing sensitive information leaks from android applications," *IEEE Access*, Vol.9, pp.119578-119590, 2021.

[3] S. Kim and J. Hur, "Mobile application privacy leak detection and security enhancement research," *Journal of the Korea Institute of Information Security & Cryptology*, Vol.29, No.1, pp.195-203, 2019.

[4] Google, User data policy [Internet], https://support.google.com/googleplay/android-developer/answer/10144311.

[5] European Union, General Data Protection Regulation [Internet], https://gdpr-info.eu/.

[6] State of California Department of Justice, California Consumer Privacy Act [Internet], https://oag.ca.gov/privacy/ccpa.

[7] U.S. Congress legislation, American Data Privacy and Protection Act[Internet], https://www.congress.gov/bill/117th-congress/house-bill/8152

[8] A. P. Felt, E. Chin, S. Hanna, D. Song, and D. Wagner, "Android permissions demystified," in *Proceedings of the 18th ACM conference on Computer and communications security*, pp.627-638, 2011.

[9] K. W. Y. Au, Y. F. Zhou, Z. Huang, and D. Lie, "Pscout: Analyzing the android permission specification," in *Proceedings of the 2012 ACM conference on Computer and Communications Security*, pp.217-228, 2012.

[10] D. Arp, M. Spreitzenbarth, M. Hubner, H. Gascon, K. Rieck, and C. Siemens, "Drebin: Effective and explainable detection of android malware in your pocket," in *Proceeding of the Network and Distributed System Security Symposium*, pp.23-26, 2014.

[11] I. M. Almomani and A. A. Khayer, "A comprehensive analysis of the android permissions system," *IEEE Access*, Vol.8, pp.216671-216688, 2020.

[12] Z. Wu and S. U.-J. Lee, "Forgotten permission usages: An empirical study on app description based android app analysis," *Journal of the Korea Society of Computer and Information*, Vol.26, No.6, pp.107-113, 2021.

[13] J. Gamba, M. Rashed, A. Razaghpanah, J. Tapiador, and N. Vallina-Rodriguez, "An analysis of pre-installed android software," in *Proceeding of IEEE Symposium on Security and Privacy*, pp.1039-1055, 2020.

[14] A. M. McDonald and L. F. Cranor, "The cost of reading privacy policies," *Isjlp*, Vol.4, pp.543, 2008.

[15] S. Wilson et al., "The creation and analysis of a website privacy policy corpus," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp.1330-1340, 2016.

[16] S. Zimmeck et al., "Maps: Scaling privacy compliance analysis to a million apps," in *Proceedings on Privacy Enhancing Technologies*, pp.66-86, 2019.

[17] F. Liu, S. Wilson, P. Story, S. Zimmeck, and N. Sadeh, "Towards automatic classification of privacy policy text," *School of Computer Science Carnegie Mellon University*, 2018.

[18] T. Libert, "An automated approach to auditing disclosure of third-party data collection in website privacy policies," in *Proceedings of the 2018 World Wide Web Conference*, pp.207-216, 2018.

[19] I. Paek, J. Oh, and K. Lee, "A study on the methods for ensuring the transparency of the privacy policies in android environment: based on General Data Protection Regulation," *Journal of the Korea Institute of Information Security & Cryptology*, Vol.29, No.6, pp.1477-1489, 2019.

[20] Leontiadis. I, Efstratiou. C, Picone. M, and Mascolo. C, "Don't kill my ads! balancing privacy in an ad-supported mobile application market." in *Proceedings of the Twelfth Workshop on Mobile Computing Systems & Applications*, pp.1-6, 2012.

[21] M. Backes, S. Bugiel, and E. Derr, "Reliable third-party library detection in android and its security applications," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp.356-367, 2016.

[22] M. I. Gordon, D. Kim, J. Perkins, L. Gilham, N. Nguyen, and M. Rinard, "Information-flow analysis of android applications in droidsafe," in *Proceeding of the Network and Distributed System Security Symposium*, pp.110, 2015

[23] F. Wei, S. Roy, and X. Ou, "Amandroid: A precise and general inter-component data flow analysis framework for security vetting of android apps," *ACM Transactions on Privacy and Security*, Vol.21, No.3, pp.1-32, 2018.

[24] Y. Pan, X. Ge, C. Fang, and Y. Fan, "A systematic literature review of android malware detection using static analysis," *IEEE Access*, Vol.8, pp.116363-116379, 2020.

[25] W. Enck et al., "Taintdroid: an information-flow tracking system for realtime privacy monitoring on smartphones," *ACM Transactions on Computer Systems*, Vol.32, No.2, pp.1-29, 2014.

[26] A. Razaghpanah et al., "Haystack: A multi-purpose mobile vantage point in user space," *arXiv preprint arXiv:1510.01419*, 2015.

[27] J. Ren, M. Lindorfer, D. J. Dubois, A. Rao, D. Choffnes, and N. Vallina-Rodriguez, "Bug fixes, improvements,... and privacy leaks: A longitudinal study of pii leaks across android app versions," in *Proceeding of the Network and Distributed System Security Symposium*, 2018.

[28] J. Ahn, H. Yoon, and S. Jung, “An enhancement scheme of dynamic analysis for evasive android malware,” *Journal of the Korea Institute of Information Security & Cryptology*, Vol.29, No.3, pp.519-529, 2019.

[29] Android Privacy Analysis, Dataset of privacy policy and mobile app analysis [Internet], https://android-privacy.github.io/

[30] Mozilla, Readability.js [Internet], https://github.com/mozilla/readability.

[31] Anthony Desnos, Androguard documentation [Internet], https://androguard.readthedocs.io.

[32] Google, The Privacy Sandbox [Internet], https://privacysandbox.com.

[33] Y. Jung, “Reliability analysis of privacy policies based on android static analysis,” in *Proceedings of the Annual Spring Conference of Korea Information Processing Society Conference (KIPS)*, Vol.29, pp.221-224, 2022.

정 윤 교

https://orcid.org/0000-0003-2396-0242
e-mail : ykjung.rokafa@gmail.com
2017년 공군사관학교 시스템공학과(학사)
2021년 서울대학교 컴퓨터공학부(석사)
2021년 ~ 현 재 공군사관학교 컴퓨터과학과 조교수
관심분야 : Privacy, Mobile Security, Network, IoT, NLP