안드로이드 정적분석 기반 개인정보 처리방침의 신뢰성 분석

정윤교
공군사관학교 컴퓨터과학과
ykjung.rokafa@gmail.com

Reliability Analysis of Privacy Policies Based on Android Static Analysis

Yoonkyo Jung
Dept. of Computer Science, Republic of Korea Air Force Academy

요 약

모바일 사용자가 증가함에 따라 모바일 앱에서 사용자가 허용하지 않은 개인정보가 유출되는 프라이버시 문제가 많아졌다. 이를 해결하기 위해 구글은 앱스토어에 등록된 앱이 사용자의 개인정보를 어떻게 활용하는지 개인정보 처리방침에 명시하도록 했다. 하지만 개인정보 처리방침이 실제로 앱의 개인정보 수집 및 처리 과정을 정확히 공개하는지 확인할 수 있는 해결책이 없으며, 사용자는 앱이 개인정보를 어떻게 활용하는지 알기 위해 개인정보 처리방침에 의존해야만 한다. 본 연구에서는 안드로이드 정적 분석으로 앱이 접근할 수 있는 데이터를 확인하고, 개인정보 처리방침의 텍스트를 추출 및 분석한 뒤 결과를 비교하여 개인정보 처리방침의 신뢰성을 분석한다. 실험을 위해 구글 앱스토어에 등록된 13,223개 앱의 패키지 파일과 부가정보를 수집했고 전처리 과정을 거쳐 분석 가능한 앱을 선정했다. 선정한 앱의 모바일 앱 분석 결과와 텍스트 분석 결과를 비교하여 모바일 앱이 개인정보 처리방침에 명시된 것보다 더 많은 개인정보에 접근할 수 있음을 입증한다.

1. 서론

스마트폰이 대중화되면서 모바일 환경에서 인터넷을 사용하는 비율이 점점 높아지고 있다. 하지만, 모바일 사용자의 증가와 함께 프라이버시 문제도 많이 발생하고 있다. 앱은 사용자의 편의를 위해 개인정보를 다룰 수 있지만, 특정 사용자를 대상으로 하는 광고나 마케팅 등 개인정보를 다른 목적으로 사용할 수 있다. 모바일 기기는 위치 정보나 기기 식별자 등 개인을 식별할 수 있는 민감한 정보를 많이 포함하고 있어 앱을 통해 개인정보가 유출되지 않도록 주의가 필요하다.

구글은 모바일 사용자의 개인정보를 보호하기 위해 사용자 데이터 정책[1]을 제정했다. 구글 데이터 정책에 따르면, 앱 제공자들은 개인정보 처리방침에 모든 앱의 활동을 공개하여 구글 앱스토어에 등록한 앱이 사용자로부터 어떤 정보를 수집하고 이를 어떻게 활용하는지 설명해야 한다. 또한, 공개한 내용은 유럽 일반 개인정보 보호법(GDPR), 캘리포니아 소비자 프라이버시 보호법(CCPA) 등 데이터 보호와 관련한 여러 규정을 만족해야 한다.

앱이 수집하는 개인정보를 제한하여 사용자를 보호하기 위해 기존 연구에서는 과도한 권한 요구를 예방하는 방법을 제시했다[2, 3, 4]. 최근 출시된 안드로이드 앱은 필수적인 권한만 요구하고 있으며, 앱을 실행할 때 사용자에게 권한을 요청하여 사용자가 직접 권한을 허용하거나 거부할 수 있다. 하지만 앱이 실제로 어떤 정보에 접근하는지 확인하기 위해서는 개인정보 처리방침에 의존해야 한다. 사용자가 개인정보 처리방침의 투명성을 확인하기 어려워 공개하지 않은 앱의 활동이 있더라도 사용자는 알 수 없다. 따라서 모바일 사용자들에게 개인정보의 수집 및 처리에 관한 정확한 정보를 제공하기 위해 개인정보 처리방침을 신뢰할 수 있는지 증명해야 한다.

본 연구에서는 안드로이드 정적 분석을 바탕으로 개인정보 처리방침의 신뢰성을 분석하기 위한 시스템을 제시한다. 설계한 시스템은 모바일 앱 분석을 통해 실제로 앱이 어떤 정보를 수집할 수 있는지 확인하고, 개인정보 처리방침을 분석한 뒤 각각의 결과를 비교하여 사용자에게 개인정보 처리 활동에 대해 명확히 공개하고 있는지 증명한다.

2. 시스템 설계

실험을 위해 앱스토어 인기 순위를 기반으로 구글 앱스토어에 등록된 13,223개 안드로이드 앱의 패키지(APK) 파일과 앱 부가정보를 수집하여 모바일 앱 분석 및 개인정보 처리방침 분석에 사용했다.

3.1 모바일 앱 분석

기존 연구에서는 앱의 권한과 관련한 시스템 호출을 조사했으며[5, 6], 정적 분석 결과를 바탕으로 정보의 흐름을 분석하여 개인정보 유출 가능성을 제시했다[7, 8]. 본 연구에서는 정적 분석으로 앱의 API 목록을 확인하여 앱이 어떠한 개인정보에 접근할 수 있는지 확인한다.

모바일 앱 분석은 파이썬으로 작성된 오픈 소스 정적 분석 도구인 Androguard[9]를 기반으로 한다. 정적 분석을 통해 앱의 활동을 확인하기 전에, 안드로이드 공식 설명서를 활용하여 개인정보를 호출하는 API 목록을 조사했다.

<표 1> 개인정보를 호출하는 API 목록 예시

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

조사 결과를 바탕으로 각각의 API가 어떤 정보를 수집하는지 확인하여 <표 1>의 형태로 저장했다. 개인정보 분류와 관련한 모든 API를 민감 API라고 정의했으며, 총 15가지 개인정보 분류에 대해 31개의 민감 API를 선정하였다.

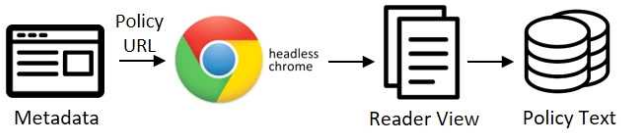
모바일 앱 분석은 APK 파일을 정적 분석 후 이를 민감 API 목록과 비교하는 순서로 진행되며 전반적인 과정은 (그림 1)과 같다. 시스템은 정적 분석으로 앱의 API를 확인하고, 이를 민감 API 목록과 비교하여 앱이 수집할 수 있는 개인정보를 확인한다. 앱이 민감 API를 포함한다면 해당 API에서 호출하는 개인정보를 수집할 수 있다고 고려한다.

(그림 1) 모바일 앱 분석 과정

3.2 개인정보 처리방침 분석

개인정보 처리방침을 분석하기 위한 선행 연구로 텍스트를 직접 분류하여 말뭉치를 제작하거나[10], 페이지 구조를 분석하는 방법이 있었다[11, 12]. 본 연구에서는 앱스토어의 부가정보에 포함된 개인정보 처리방침 웹 주소를 활용하여 텍스트를 추출 및 분석하고 이를 모바일 앱 분석 결과와 비교한다.

개인정보 처리방침의 웹 주소는 앱마다 형식이 달라 단순한 방법으로 텍스트를 추출하기 어렵다. 따라서 웹 주소에서 발생하는 오류를 처리하고 자동으로 텍스트를 추출하는 텍스트 추출기를 만들었다.



(그림 2) 개인정보 처리방침의 텍스트 추출 과정

(그림 2)는 텍스트 추출기를 사용한 텍스트 추출 과정이다. 먼저 앱 부가정보에서 개인정보 처리방침 웹 주소를 확인하고 헤드리스 브라우저로 실행한다. 실행한 페이지를 readability 라이브러리[13]를 사용하여 읽기 전용으로 변환한 뒤 텍스트만 추출한다. 텍스트 추출기는 웹 주소에 페이지가 아닌 첨부파일만 존재하는 경우 파일을 자동으로 실행하여 형식을 확인 후 텍스트를 저장한다. 또한, 웹 주소가 없거나 유효하지 않을 시 예외 처리를 진행했다.

추출한 텍스트를 분석하기 위해 각각의 개인정보 분류를 대표하는 키워드를 선정했으며, 50개 앱을 샘플링 후 정밀도, 재현율, F1 점수 측면에서 키워드의 실효성을 확인하여 <표 2>의 결과를 얻었다.

<표 2> 키워드의 실효성 검증을 위한 샘플링 결과

Precision	Recall	F1 score
0.81	0.78	0.79

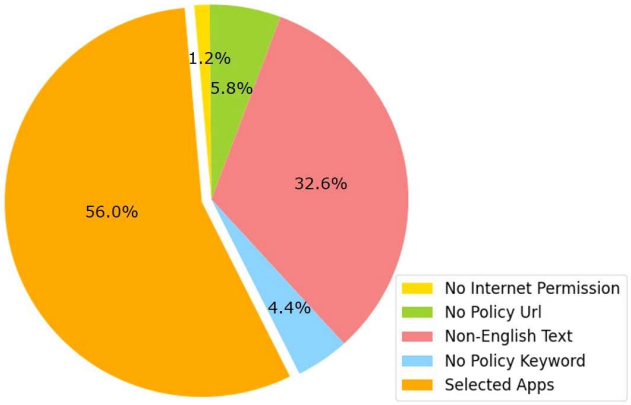
정밀도와 재현율을 구하는 식은 (수식 1)과 같으며 F1 점수는 정밀도와 재현율의 조화 평균이다.

$$precision = TP/(TP+FP)$$
$$recall = TP/(TP+FN) \quad \text{(수식 1)}$$

True Positive(TP)는 키워드로 분석한 개인정보 분류가 실제로 텍스트에 포함된 내용인 경우, True Negative(TN)는 분석 결과가 없고 실제로 관련 없는 경우, False Positive(FP)는 분석을 통해 나온 결과가 실제로 관련 없는 경우, False Negative(FN)는 분석 결과가 없지만 실제로 개인정보 관련 내용이 존재하는 경우이다. 선정한 키워드를 바탕으로 패턴 매칭을 진행하여 개인정보 처리방침에서 앱이 어떤 개인정보를 다룰 수 있다고 설명하는지 확인했다.

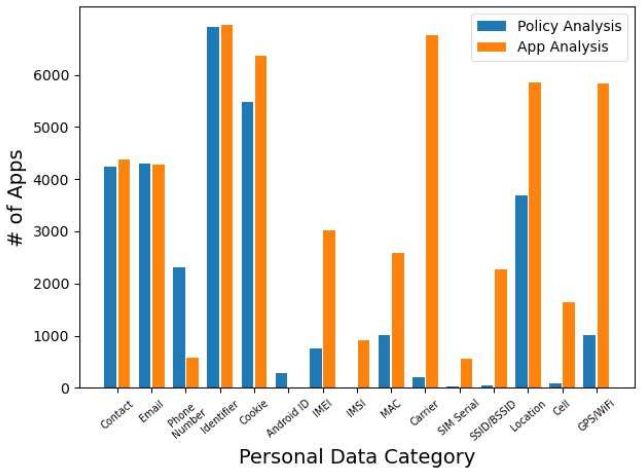
4. 결과 분석

수집한 앱의 일부는 정보 유출의 가능성이 작거나 개인정보 처리방침 웹 주소가 없어 분석을 수행할 수 없다. 이처럼 모든 앱을 분석하는 것은 비효율적이기에 4가지 조건에 따라 전처리를 진행했다.



(그림 3) 예외 조건을 적용한 앱 전처리 결과

(그림 3)은 예외 조건에 따른 전처리 결과이다. 인터넷 권한이 없는 앱, 개인정보 처리방침의 웹 주소가 없는 앱, 텍스트가 영어가 아닌 앱, 추출한 텍스트가 개인정보 처리방침이라고 보기 힘든 앱을 제외하고 총 7,401개의 앱이 실험에 사용되었다.



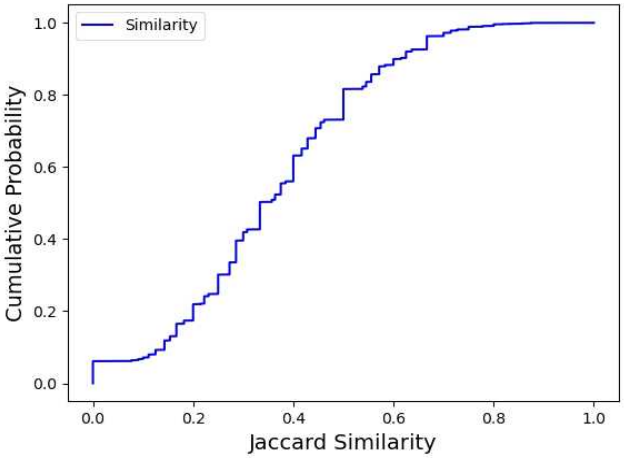
(그림 4) 개인정보 분류에 따른 분석 결과 비교

(그림 4)는 각 분석 결과를 개인정보 분류에 따라 비교한 것이다. Y축의 값이 클수록 개인정보 분류와 관련한 API가 많이 확인되었거나 패턴 매칭 결과가 높게 나타남을 의미한다. 그래프를 통해 앱 분석 결과에서 개인정보 분류와 관련한 앱이 더 많음을 알 수 있었고, 모바일 앱이 예상한 것보다 많은 개인정보에 접근할 수 있다는 점을 확인했다.

다음으로 신뢰성을 확인하기 위해 합집합과 교집합의 비율을 구하는 자카드 유사도를 사용했다. 자카드 유사도를 구하는 함수 J는 (수식 2)와 같다.

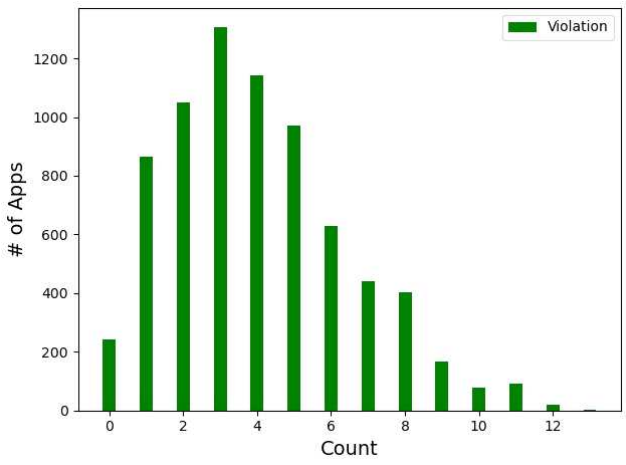
$$J(A,B) = \frac{A \cap B}{A \cup B} = \frac{\text{App Analysis} \cap \text{Policy Analysis}}{\text{App Analysis} \cup \text{Policy Analysis}} \quad \text{(수식 2)}$$

모바일 앱과 개인정보 처리방침의 텍스트 분석 결과에 대한 자카드 유사도를 (그림 5)와 같이 누적 분포함수 그래프로 나타냈다.



(그림 5) 모바일 앱과 텍스트 분석 결과의 자카도 유사도

그래프에서 x값이 0에 가까울수록 앱의 실제 활동과 개인정보 처리방침의 내용이 일치하지 않음을 의미한다. 분석 결과 0.2와 0.6인 구간에 유사도가 분포하고 있고 0에 가까운 결과도 포함되어 있다.



(그림 6) 각 앱에서 개인정보 처리방침을 위반한 횟수

또한, 개인정보 처리방침에 앱 분석을 통해 확인한 민감 API에 대한 설명이 없는 경우 정책을 위반했다고 고려하여 (그림 6)과 같은 결과를 얻었다. 그래프를 분석해보면 많은 모바일 앱이 최소 1번 이상 개인정보 처리방침을 위반한 점을 알 수 있다.

5. 결론

모바일 환경에서의 개인정보 유출을 방지하기 위해 앱 제공자들은 안드로이드 마켓에 앱을 등록할 때 개인정보 처리방침에 앱의 개인정보 수집 및 처리 과정을 공개해야 한다. 하지만 개인정보 처리방침에 앱의 활동이 일부 누락되어 있어도 사용자가 알 수 없다. 구글은 유해 앱으로부터 기기를 보호하기 위해 구글 플레이 프로텍트 기능을 구현하여 위험 요소가 있는지 확인하고 있으나 앱의 모든 기능을 점검하기 어렵다. 따라서 사용자가 앱을 신뢰할 수 있는지 직접 확인하는 방안이 필요하다.

본 연구에서는 구글 앱스토어에 등록된 앱의 개인정보 처리방침을 사용자가 신뢰할 수 있는지 진단하기 위해 안드로이드 정적 분석을 기반으로 하는 시스템을 제작했다. 실험을 위해 구글 앱스토어에서 13,223개 앱의 APK 파일과 부가정보를 수집했다. APK 파일을 정적 분석하여 사전에 조사한 민감 API 목록과 비교하고, 부가정보에서 가져온 개인정보 처리방침 웹 주소를 사용하여 텍스트 추출 및 분석을 진행했다. 두 결과를 비교한 결과 개인정보 처리방침에 명시한 내용에 비해 앱이 접근할 수 있는 개인정보가 많다는 것을 확인할 수 있었으며, 이는 잠재적인 정보 유출의 가능성을 시사한다. 본 연구에서 제안한 시스템을 모바일 앱 사용자가 활용한다면 개인정보 처리방침을 신뢰성을 판단할 수 있고, 이를 통해 개인정보 처리방침의 투명성과 전반적인 개인정보 보호 수준을 향상할 수 있다.

참고문헌

[1] "User data policy." https://support.google.com/googleplay/android-developer/answer/10144311/. Accessed: 2022-05-06.

[2] 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.

[3] 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.

[4] 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 NDSS, vol. 14, pp. 23-26, 2014.

[5] I. Leontiadis, C. Efstratiou, M. Picone, and C. Mascolo, "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.

[6] 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.

[7] M. I. Gordon, D. Kim, J. H. Perkins, L. Gilham, N. Nguyen, and M. C. Rinard, "Information flow analysis of android applications in droidsafe.," in NDSS, vol. 15, p. 110, 2015.

[8] F. Wei, S. Roy, and X. Ou, "Amandroid: A precise and general intercomponent data ow analysis framework for security vetting of android apps," in Proceedings of the 2014 ACM SIGSAC conference on computer and communications security, pp. 1329-1341, 2014.

[9] "Androguard documentation." https://androguard.readthedocs.io/. Accessed: 2022-05-06.

[10] R. Ramanath, F. Liu, N. Sadeh, and N. A. Smith, "Unsupervised alignment of privacy policies using hidden markov models," in Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pp. 605-610, 2014.

[11] 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.

[12] 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.

[13] "Readability.js." https://github.com/mozilla/readability/. Accessed: 2022-05-06.