Journal of Computer and Communications, 2019, 7, 147-156
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Predicting Users Mobile App Privacy Preferences

Aziz Alshehri[1], Fayez Alotaibi[2]

[1]Computer Science and Information Technology College, Umm Al-Qura University, Makkah, KSA
[2]Computer Science and Information Technology College, Shaqra University, Riyadh, KSA
Email: aaashehri@uqu.edu.sa, fayezotb@su.edu.sa

How to cite this paper: Alshehri, A. and Alotaibi, F. (2019) Predicting Users Mobile App Privacy Preferences. *Journal of Computer and Communications*, **7**, 147-156.
https://doi.org/10.4236/jcc.2019.710014

Received: September 8, 2019
Accepted: September 30, 2019
Published: October 22, 2019

Abstract

Home users are using a wide and increasing range of different technologies, devices, platforms, applications and services every day. In parallel, home users are also installing and using an enormous number of apps, which collect and share a large amount of data. Users are also often unaware of what information apps collect about them, which is really valuable and sensitive for them. Therefore, users are becoming increasingly concerned about their personal information that is stored in these apps. While most mobile operating systems such as Android and iOS provide some privacy safeguards for users, it is unrealistic to manage and control a large volume of data. Accordingly, there is a need for a new technique, which has the ability to predict many of a user's mobile app privacy preferences. A major contribution of this work is to utilise different machine learning techniques for assigning users to the privacy profiles that most closely capture their privacy preferences. Applying privacy profiles as default settings for initial interfaces could significantly reduce the burden and frustration of the user. The result shows that it's possible to reduce the user's burden from 46 to 10 questions by achieving 86% accuracy, which indicates that it's possible to predict many of a user's mobile app privacy preferences by asking the user a small number of questions.

Keywords

Applications, Privacy in Mobile Computing, Context-Aware Privacy Control

1. Introduction

With the rapid growth of the devices, activities, services and information, the enormous amount of private and personal information that is stored has also increased. Therefore, users are becoming increasingly concerned about their personal information, how it is used, by whom and where it is stored [1]. For instance, a Consumer Report found that 92% of British and US Internet users are concerned about their privacy online [2]. When users became aware of online privacy issues, they were asked what made them most worried about their online privacy: 45% of British Internet users stated that it is personal information being shared between companies [3]. In addition, 89% of users avoided these companies because they believed that companies do not protect their privacy. It was also found that 76% of Internet users limited their online activity in the last 12 months due to these concerns [4]. This evidence indicates that users are sufficiently worried about their online privacy.

Due to the concerns of the users about privacy protection, most mobile operating systems such as Android and iOS provide some privacy safeguards for users [4]. Despite these provisions, it is unrealistic to manage and control a large volume of data. Prior study shows that users, on average, have to make over one hundred permission decisions [5]. This approach could significantly increase the burden on and frustration of the user.

A noticed focus has been given to the development of policies, procedures and tools that aid an end-user in managing and understanding their privacy-related information. However, these approaches assume that users can correctly configure all resulting settings and they have uniform privacy requirements. In reality, users do have different privacy concerns and requirements as they have heterogeneous privacy attitudes and expectations [6]. For example, some users consider personal information such as age, address and gender in their profile on a social network being more sensitive than others [7]. Furthermore, in practice it is unrealistic to assume homogeneous privacy requirements across a whole population [8].

In order to overcome the burden related to controlling a large number of privacy settings, prior study suggests that—despite users' privacy preferences are very diverse, it is possible to assign users to the privacy profiles that most closely capture their privacy preferences [9] [10]. Despite these approaches, shows the diversity of users' privacy preferences and cluster users into different profiles, they do not show how to assign users to the most closely profile using machine learning by asking the user a small number of questions.

Accordingly, there is a need for a solution in order to assign the user to the privacy profiles that most closely capture their privacy preferences. This study is an extension for the prior research which shows users are diverse and it is possible to divide users into small groups according to the users' privacy preferences [11]. The result indicates the optimal number of clusters based on the k-means method is four clusters. However, the prior study does not show how to predict the user's mobile app privacy preferences by asking the user a small number of questions. Applying privacy profiles as default settings for initial interfaces could significantly reduce the burden and frustration of the user.

This paper is organised into five sections. Section 2 presents an analysis of background literature. Section 3 shows how the data collected. Section 4 described the machine learning approach that applied in order to predict the user's mobile app privacy preferences. The conclusions are presented in Section 5.

2. Background Literature

Numerous techniques have been proposed to monitor personal information [6] [12] [13]. The majority of existing techniques has focused upon the technical aspect to protect the privacy of users. They have shown that is possible to monitor sensitive information for users in real time. Some of the tools used a dynamic approach to monitor personal information for users. Whilst, a few studies used a network approach to detect information leakage in the mobile as shown in Table 1.

Table 1. A review of monitoring and privacy controls.

Author	Privacy Tool	Methods	Control Type
[6]	Protect MyPrivacy	Developing a crowdsourcing system to help the user to make informed decisions	Static control
[7]	PrivacyGuard	Detecting the leakage of multiple types of sensitive data and modifying the leaked information	Finer granularity
[9]	Reconciling Mobile App Privacy	Analyzing people's privacy preferences when it comes to granting permissions	No
[12]	Taintdroid	Dynamic approach to track the data through four levels	No
[13]	Little Brothers Watching You	Investigation users' understanding when the data is shared.	No
[14]	AntMonitor	Analysing actual network traffic of Android using VPNService API to intercept traffic.	Static control
[15]	TISSA	Providing users with empty or bogus options.	Finer granularity
[16]	AppFence	Providing users with two privacy controls to protect sensitive resources (shadowing and blocking)	Finer granularity
[17]	PiOS	Using static analysis to detect apps leak	No
[18]	Styx	Providing the user with more meaningful privacy information based on the actual behaviour of apps	No
[19]	AppIntent	AppIntent determines if transmission is user intended or not	Static control
[20]	Turtle Guard	TurtleGuard helps users to vary their privacy preferences based on a few selected contextual circumstances.	Finer granularity
[21]	SmarPer	Predicting permission decisions at runtime.	Finer granularity
[22]	Ask On First Use	Contextually-aware permission system that performs permission denial dynamically	Static control

A number of research prototypes have not only monitored sensitive information for users but also provided user control over the personal information such as AntMonitor [14], ProtectMyPrivacy [6] and TISSA [15]. However, most current privacy controls support only binary and static privacy controls. A few studies such as TISSA [15] and AppFence [16] provided users with multiple levels of control. TISSA provides users with empty or bogus options for personal information that may be requested by the app. Whilst, AppFence provides users with two privacy controls to protect sensitive resources: shadowing and blocking. However, the tools do not allow users to limit the disclosure of their private information in multiple levels taking factors like the level of user's knowledge to make the right choice in order to reduce the burden on users.

A few studies such as [20] [21] and [22] used machine learning to predict user preferences. Tsai *et al.* designed TurtleGuard that automatically make privacy decisions on behalf of the user. Olejnik *et al.* also designed a system that predicts permission decisions at runtime. These studies have shown that is possible to predict user's preferences. In order to design initial interface, Lin et al divided users into a small number of privacy profiles, which collectively go a long way in capturing the diverse preferences of the entire population [9]. However, they do not elicit user's privacy preferences in a context where they are not just about the permissions requested by an app but also about current knowledge and their desires to control different aspects of privacy and good usability design to maximise use.

However, the aforementioned privacy solutions also assume that users are the same in the context of how to use the privacy system and how to control the large volume of personal information. Whilst the current research and available literature have highlighted differences between the expert and novice knowledge in the context of using the system and the knowledge in the domain [23] [24]. Therefore, it is difficult for the novice user to configure a lot of settings correctly while the expert user has the ability to manage it because he has knowledge in the domain.

3. Data Collection

Participants were recruited through different platforms such as Email, social network, and some communities' centre. Prior to displaying the survey questions, its aims and structure were briefed confirming that the respondents should be 18 years or older and they are free to withdraw up until the final submission of their responses. In total, 407 completed responses and the total responses are within the range of other surveys in the research domain and close to the expected and targeted figure.

Demographic information was collected including questions related to gender, age, education, and the level of knowledge in order to analyse the data, though the age ratio or any other demographic composition of the participants were not specifically controlled. Among these participants, 30% of them were male; 70 of them were female. Regarding the age, almost half of the participants were between 25 and 34 which represent 47% of participants. The second large age group was between 35 - 44 which represent 35%. The vast majority of the population of the participants were aged between 24 - 44 years old.

4. Predicting the User's Mobile App Privacy Preferences

This section examines how to predict many of a user's mobile app privacy preferences which could significantly reduce user burden with minimum questions. Accordingly, the machine learning technique was utilised to assign users to the privacy profiles that most closely capture their privacy preferences.

4.1. Interpreting the Resulting Privacy Profiles

As mentioned in the previous section, this paper is an extension of a previous study, which divided users into different profiles by utilising k-means method to determine the number of profiles. However, in order to determine the optimal number of clusters in the dataset, three methods were performed: statistical testing methods, visual exploration, and precision of predicting users' preferences, which is already performed in our prior work [11].

- Statistical testing methods such as the Gap statistic to compare the total within intra-cluster variation for different values of k.
- Visual exploration relies on how the cluster is easy to interpret and meaningful.
- Precision of predicting users' preferences: multi classifications were performed to determine the optimal number of clusters. When the classifier has a high prediction of determining each user with their cluster and has high accuracy, this indicates that the number of clusters is good to utilize. However, the result shows that the optimal number of clusters is four clusters.

The red chart in Figure 1 indicates a higher level of concern while the green chart indicates a lower level of concern. Cluster 2 represents the conservative group. However, it is clear from Figure 1 that cluster 2 is the largest cluster. Whilst the green chart indicates a lower level of concern who comfort to disclose their data (indicate of unconcerned) and represents 26% of users. The rest of the participants in the remaining clusters such as cluster 4 and 3 are covered in orange or yellow colures which indicates participants are moderately concerned or somewhat to share privacy-related information to the apps.

When the centroid of each cluster was computed by averaging the feature vectors of instances within the cluster, cluster 4 and 3 indicate somewhat concerned (cluster 4: $\mu = 2$, cluster 3 $\mu = 3$). Whilst the centroid of cluster 2 and 1 were $\mu = 1$ and $\mu = 4$ respectively. The characteristics of the clusters reveal a significant difference between each privacy profile, which means each cluster, has a unique profile.

4.2. Feature Selection

The machine learning approach is applied in order to assign users to these four privacy profiles. R program supports different machine algorithms such as Support Vector Machines (SVM), Random Forest (RF) and K-Nearest Neighbors (kNN). Before using classifier, it is important to determine which questions are the most important to ask users in order to minimise the questions. Therefore, the following techniques were used to minimise the 46 questions and help the models perform better and efficiently as shown in Figure 2.

Machine learning assigns a weight to each question about privacy-related information. Hence, it could be easy to minimise the number of questions, which in turn could actually help reduce user burden. Table 2 shows the features ranked according to the ranking algorithm with their weight. "Gini index" was performed to assign a score and rank the questions. These scores which are denoted as "Mean Decrease Gini" by the importance measure indicate how much each question contributes to the homogeneity in the privacy-related-information. After ranking the questions, a top ten questions were selected in order to minimise the number of questions.

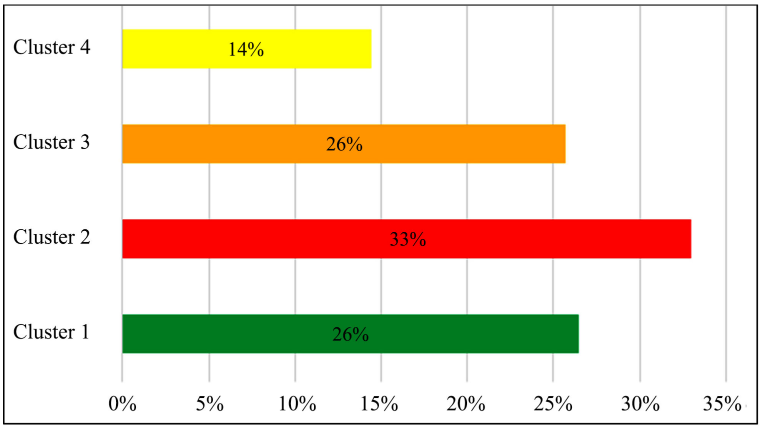

Figure 1. The percentage of each cluster.

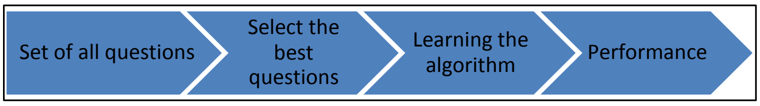

Figure 2. Feature selection methods.

Table 2. The 10 most important questions.

N	Category	Data	Weight
1	Shopping	Identity	12.41
2	Navigation	Approximate location	12.33
3	Shopping	Contacts	12.31
4	Navigation	Contacts	11.57
5	Navigation	Exact location	11.1
6	Productivity	Calendar	11.00
7	Navigation	Phone	10.76
8	Productivity	Identity	10.35
9	Lifestyle	Phone	9.70
10	Shopping	Approximate location	9.14

4.3. Accuracy of Predictions

This section aims to create some models of the data and estimate their accuracy. However, there are different algorithms therefore; the following steps were performing in order to determine the best algorithm:

1) Set-up the test harness to use 10-fold cross-validation.

2) Build five different algorithms to predict the user's mobile app privacy preferences.

3) Select the best algorithm.

Ten folds cross-validations were performed to estimate accuracy. This divides the data into 10 groups, train in 9 and test on 1 and release for all combinations of train-test splits. The process was repeated three times for each algorithm with different splits of the data into 10 groups; in an effort to get a more accurate estimate to predict the user's mobile app privacy preferences. Five different machine-learning algorithms were evaluated in order to determine the best algorithm would be good on this problem. The following list shows the five algorithms: Feature Selection Methods

- Linear Discriminant Analysis (LDA).
- Classification and Regression Trees (CART).
- K-Nearest Neighbors (kNN).
- Support Vector Machines (SVM) with a linear kernel.
- Random Forest (RF).

Figure 3 shows the evaluation results of each algorithm and compare the spread and the mean accuracy of each model. It is clear from Figure 3 that the most accurate model, in this case, was SVM. SVM classifiers achieve the highest accuracy 86% with 10 questions.

Overall accuracy for 46 questions and 4 clusters is 99%. When asking users to answer ten of questions related to privacy decisions, the accuracy decreases to 86%. This result reflects the exploration of tradeoffs between accuracy and the number of questions—in other words, tradeoffs between accuracy and user burden. The result also decreases by 80% (46 to 10 questions) of the user's effort.

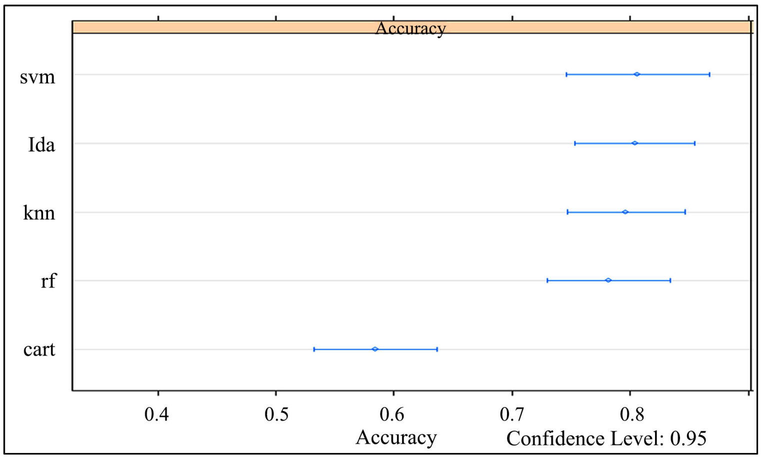

Figure 3. Feature selection methods.

5. Conclusion

The outcomes of this research show that it could significantly reduce the burden and frustration of the user while allowing users to better control information. In particular, the classifiers could be built to assign users to the privacy profiles that most closely capture their privacy preferences. In order to reduce user burden in terms of decisions, machine learning assigns a weight to each question about privacy-related information. Hence, it could be easy to minimise the number of questions, which in turn could actually help reduce user burden. Five different machine-learning algorithms were evaluated in order to determine the best algorithm. SVM algorithm achieved the highest accuracy, which is 86% for just 10 questions. The research was particularly encouraging as they offer the prospect of remarkably alleviating user burden.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Anton, A.I., Earp, J.B. and Young, J.D. (2010) How Internet Users' Privacy Concerns Have Evolved. *IEEE Privacy & Security*, **1936**, 21-27. https://doi.org/10.1109/MSP.2010.38

[2] TRUSTe (2016) TRUSTe/NCSA Consumer Privacy Infographic—US Edition. https://www.truste.com/resources/privacy-research/ncsa-consumer-privacy-index-us

[3] Federal Trade Commission (2013) Mobile Privacy Disclosures—Building Trust through Transparency. 29.

[4] Kelley, P.G., Consolvo, S., Cranor, L.F., Jung, J., Sadeh, N. and Wetherall, D. (2012) A Conundrum of Permissions: Installing Applications on an Android Smartphone. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 7398, 68-79. https://doi.org/10.1007/978-3-642-34638-5_6

[5] Pew Research Center (2015) App Permissions: An Analysis of Android Phones. https://www.pewinternet.org/2015/11/10/an-analysis-of-android-app-permissions

[6] Agarwal, Y. and Hall, M. (2012) Protect My Privacy: Detecting and Mitigating Privacy Leaks on iOS Devices Using Crowdsourcing Categories and Subject Descriptors. *Proceeding of the* 11*th Annual International Conference on Mobile Systems*, *Applications*, *and Services*, Vol. 6, 97-110. https://doi.org/10.1145/2462456.2464460

[7] Song, Y. and Hengartner, U. (2015) PrivacyGuard: A VPN-Based Platform to Detect Information Leakage on Android Devices. *Proceedings of the* 5*th Annual ACM CCS Workshop on Security and Privacy in Smartphones and Mobile Devices*, Denver, 12 October 2015, 15-26. https://doi.org/10.1145/2808117.2808120

[8] Song, Y. (2015) PrivacyGuard: A VPN-Based Approach to Detect Privacy Leakages on Android Devices. 15-26.

[9] Liu, B., Lin, J. and Sadeh, N. (2013) Reconciling Mobile App Privacy and Usability on Smartphones: Could User Privacy Profiles Help. *Proceedings of the* 23*rd International Conference on World Wide Web*, Seoul, 7-11 April 2014, 201-212.

https://doi.org/10.1145/2566486.2568035

[10] Lin, J., Liu, B., Sadeh, N. and Hong, J.I. (2014) Modeling Users' Mobile App Privacy Preferences: Restoring Usability in a Sea of Permission Settings. 12*th USENIX Security Symposium*, Menlo Park, CA, 9-11 July 2014, 199-212.

[11] Alshehri, A. and Alotaibi, F. (2019) Profiling Mobile Users Privacy Preferences. *International Journal of Digital Society*, **10**, 1436-1441.

[12] Enck, W., *et al*. (2014) TaintDroid: An Information-Flow Tracking System for Real-Time Privacy Monitoring on Smartphones. *ACM Transactions on Computer Systems*, **32**, 5. https://doi.org/10.1145/2619091

[13] Balebako, R., Jung, J., Lu, W., Cranor, L.F. and Nguyen, C. (2013) Little Brothers Watching You: Raising Awareness of Data Leaks on Smartphones. *Proceedings of the Ninth Symposium on Usable Privacy and Security*, Newcastle, 24-26 July 2013, Vol. 12, 1-11. https://doi.org/10.1145/2501604.2501616

[14] Le, A., Varmarken, J., Langhoff, S., Shuba, A., Gjoka, M. and Markopoulou, A. (2015) AntMonitor: A System for Monitoring from Mobile Devices. *Proceedings of the* 2015 *ACM SIGCOMM Workshop on Crowdsourcing and Crowdsharing of Big* (*Internet*) *Data*, Vol. 15, London, 17 August 2015, 15-20. https://doi.org/10.1145/2787394.2787396

[15] Zhou, Y., Zhang, X., Jiang, X. and Freeh, V.W. (2011) Taming Information-Stealing Smartphone Applications (on Android). 4*th International Conference on Trust and Trustworthy Computing*, Pittsburgh, 22-24 June 2011, 93-107. https://doi.org/10.1007/978-3-642-21599-5_7

[16] Hornyack, P., Han, S., Jung, J., Schechter, S. and Wetherall, D. (2011) These Aren't the Droids You're Looking for: Retrofitting Android to Protect Data from Imperious Applications. *Proceedings of the* 18*th ACM Conference on Computer and Communications Security*, Chicago, 17-21 October 2011, 639-652. https://doi.org/10.1145/2046707.2046780

[17] Egele, M., Kruegel, C., Kirda, E. and Vigna, G. (2011) PiOS Detecting Privacy Leaks in iOS Applications. *Proceedings of the* 18*th Annual Network & Distributed System Security Symposium*, San Diego, 6-9 February 2011, 11.

[18] Bal, G., Rannenberg, K. and Hong, J. (2014) Styx: Design and Evaluation of a New Privacy Risk Communication Method for Smartphones. 29*th IFIP TC* 11 *International Conference*, Marrakech, 2-4 June 2014, 113-126. https://doi.org/10.1007/978-3-642-55415-5_10

[19] Yang, Z., Yang, M., Zhang, Y., Gu, G., Ning, P. and Wang, X.S. (2013) AppIntent: Analyzing Sensitive Data Transmission in Android for Privacy Leakage Detection. *Proceedings of the* 2013 *ACM SIGSAC Conference on Computer & Communications Security*, Berlin, 4-8 November 2013, 1043-1054. https://doi.org/10.1145/2508859.2516676

[20] Tsai, L., *et al*. (2017) Turtle Guard: Helping Android Users Apply Contextual Privacy Preferences. 13*th Symposium on Usable Privacy and Security*, Menlo Park, 9-11 July 2014, 145-162.

[21] Olejnik, K., Dacosta, I., Machado, J.S., Huguenin, K., Khan, M.E. and Hubaux, J.P. (2017) SmarPer: Context-Aware and Automatic Runtime-Permissions for Mobile Devices. *IEEE Symposium on Security and Privacy*, San Jose, 22-26 May 2017, 1058-1076. https://doi.org/10.1109/SP.2017.25

[22] Wijesekera, P., *et al*. (2018) Contextualizing Privacy Decisions for Better Prediction (and Protection). *Proceedings of the* 2018 *CHI Conference on Human Factors in Computing Systems*, Montreal, 21-26 April 2018, Paper No. 268.

https://doi.org/10.1145/3173574.3173842

[23] Chua, W.Y. and Chang, K.T.T. (2016) An Investigation of Usability of Push Notifications on Mobile Devices for Novice and Expert Users. 49*th Hawaii International Conference on System Sciences*, Koloa, 5-8 January 2016, 5683-5690. https://doi.org/10.1109/HICSS.2016.703

[24] Wisniewski, P.J., Knijnenburg, B.P. and Lipford, H.R. (2017) Making Privacy Personal: Profiling Social Network Users to Inform Privacy Education and Nudging. *International Journal of Human Computer Studies*, **98**, 95-108. https://doi.org/10.1016/j.ijhcs.2016.09.006