

Analysis of Application Used on Smartphones

Fareha Noor¹, Praveen Babu G²

¹M.Tech. Student, Department of Information Technology (DIT), JNTUH University College of Engineering, Science, and Technology, Kukatpally, Hyderabad, 500085, Telangana, India.

²Associate Professor of CSE, Department of Information Technology (DIT), JNTUH University College of Engineering, Science, and Technology, Kukatpally, Hyderabad, 500085, Telangana, India.

ABSTRACT:

As smartphones have become indispensable personal devices, the number of smartphone users has increased dramatically over the last decade. These personal devices, which are supported by a variety of smartphone apps, allow people to access Internet services in a convenient and ubiquitous manner. App developers and service providers can collect fine-grained app usage traces, revealing connections between users, apps, and smartphones. We present a comprehensive review of the most recent research on smartphone app usage analysis in this survey. Our survey summarizes advanced technologies and key patterns in smartphone app usage behaviors, all of which have significant implications for all relevant stakeholders, including academia and industry. We begin by describing four data collection methods: surveys, monitoring apps, network operators, and app stores, as well as nine publicly available app usage datasets. We then systematically summarize the related studies of app usage analysis in three domains: app domain, user domain, and smartphone domain. We make a detailed taxonomy of the problem studied, the datasets used, the methods used, and the significant results obtained in each domain. Finally, we discuss future directions in this exciting field by highlighting research challenges.

Keywords: Smartphone, Analysis, Internet Services, Apps.

1. INTRODUCTION

People can now use their smartphone apps to access a variety of Internet services, including instant messaging (e.g., WhatsApp, WeChat), online socializing (e.g., Twitter, Weibo), electronic commerce (e.g., Amazon, Taobao), and online payment (e.g., PayPal, Alipay). These services have become an important part of the infrastructure of the modern information society, making smartphone apps a necessity in daily life. According to a report from Statista, the number of apps available in Google Play, the official app store of Android, has increased exponentially from 16,000 in December 2009 to 2,893,806 in July 2021. The app market is expected to generate 935.2 billion US dollars in business value by 2023. Such a vast and vital app market has attracted developers and service providers to investigate app usage behavior to better develop and deliver mobile apps. Understanding app usage behaviors has significant implications for all relevant stakeholders, including smartphone manufacturers, network operators, market intermediaries, app developers, and end consumers. To improve device performance and extend usage time, smartphone manufacturers can optimize the scheduling of various smartphone resources, such as CPU, memory, and battery power, based on the usage patterns of specific apps. Based on app traffic patterns, network operators can dynamically optimize traffic offloading schemes and improve network services. Furthermore, network operators and market intermediaries can

provide personalized services, such as accurate recommendations and targeted advertisements, by profiling mobile users' preferences and interests from their app usage behaviors. By doing so, operators and intermediaries can improve the quality of experience (QoE) while increasing profits. App developers can better understand customer satisfaction and market trends by analyzing app usage and profiling app popularity, which may provide excellent guidance for upgrading existing apps and designing new apps.

2. LITERATURE REVIEW

Mobile phone data for informing public health actions across the COVID-19 pandemic life cycle

The coronavirus 2019–2020 pandemic (COVID-19) poses unprecedented challenges for governments and societies around the world (1). Nonpharmaceutical interventions have proven to be critical for delaying and containing the COVID-19 pandemic (2–6). These include testing and tracing, bans on large gatherings, nonessential business and school and university closures, international and domestic mobility restrictions and physical isolation, and total lockdowns of regions and countries. Decision-making and evaluation or such interventions during all stages of the pandemic life cycle require specific, reliable, and timely data not only about infections but also about human behavior, especially mobility and physical copresence. We argue that mobile phone data, when used properly and carefully, represents a critical arsenal of tools for supporting public health actions across early-, middle-, and late-stage phases of the COVID-19 pandemic.

Understanding the challenges of mobile phone usage data

Driven by curiosity and our own three diverse smartphone application usage datasets, we sought to unpack the nuances of mobile device use by revisiting two recent Mobile HCI studies [1, 17]. Our goal was to add to our broader understanding of smartphone usage by investigating if differences in mobile device usage occurred not only across our three datasets, but also in relation to prior work. We found differences in the top-10 apps in each dataset, in the durations and types of interactions as well as in micro-usage patterns. However, it proved very challenging to attribute such differences to a specific factor or set of factors: was it the time frame in which the studies were executed? The recruitment procedure? The experimental method? Using our somewhat troubled analysis, we discuss the challenges and issues of conducting mobile research of this nature and reflect on caveats related to the replicability and generalizability of such work.

Finding spatiotemporal patterns of mobile application usage

Understanding mobile application usage patterns is significant for producing better services and enriching user experience. The understanding of spatiotemporal patterns of application usage is still limited. In this paper, we aim at finding spatiotemporal mobile app usage patterns and propose a framework to capture who, when, where, and what applications are used. We first collect a large-scale and real-world application usage dataset covering over 400 thousand active users and 600 million records. In order to introduce spatial features, we partition the collection area into small regions. By grouping regions of similar point-of-interest attributes, we then map 796 regions onto 13 region clusters with semantic meanings. As a result, the original data is reformed as a tensor of four dimensions, i.e., users, application categories, region clusters, and time-slots. We then leverage a multi-way clustering algorithm on the tensor to extract coupling relations between different dimensions. Finally, we discover 508 distinct spatiotemporal application usage patterns with meaningful labels, including E- readers,

Digital payers, Uber/DiDi drivers, Young parents, Travellers and Travel planners. The results produced by our framework can serve a series of applications, e.g., identifying user habits and inferring user demographics, which are helpful in customized services like personalized recommendations.

Effective and real-time in-app activity analysis in encrypted Internet traffic streams

The mobile in-App service analysis, aiming at classifying mobile internet traffic into different types of service usages, has become a challenging and emergent task for mobile service providers due to the increasing adoption of secure protocols for in-App services. While some efforts have been made for the classification of mobile internet traffic, existing methods rely on complex feature construction and large storage cache, which lead to low processing speed, and thus not practical for online real-time scenarios. To this end, we develop an iterative analyzer for classifying encrypted mobile traffic in a real-time way. Specifically, we first select an optimal set of most discriminative features from raw features extracted from traffic packet sequences by a novel Maximizing Inner activity similarity and Minimizing Different activity similarity (MIMD) measurement. To develop the online analyzer, we first represent a traffic flow with a series of time windows, which are described by the optimal feature vector and are updated iteratively at the packet level. Instead of extracting feature elements from a series of raw traffic packets, our feature elements are updated when a new traffic packet is observed and the storage of raw traffic packets is not required. The time windows generated from the same service usage activity are grouped by our proposed method, namely, recursive time continuity constrained KMeans clustering (rCKC). The feature vectors of cluster centers are then fed into a random forest classifier to identify corresponding service usages. Finally, we provide extensive experiments on real-world traffic data from Wechat, Whatsapp, and Facebook to demonstrate the effectiveness and efficiency of our approach. The results show that the proposed analyzer provides high accuracy in real-world scenarios, and has low storage cache requirement as well as fast processing speed.

POWERFUL: Mobile app fingerprinting via power analysis

Which apps a mobile user has and how they are used can disclose significant private information about the user. In this paper, we present the design and evaluation of POWERFUL, a new attack which can fingerprint sensitive mobile apps (or infer sensitive app usage) by analyzing the power consumption profiles on Android devices. POWERFUL works on the observation that distinct apps and their different usage patterns all lead to distinguishable power consumption profiles. Since the power profiles on Android devices require no permission to access, POWERFUL is very difficult to detect and can pose a serious threat against user privacy. Extensive experiments involving popular and sensitive apps in Google Play Store show that POWERFUL can identify the app used at any particular time with accuracy up to 92.9%, demonstrating the feasibility of POWERFUL.

Carat: Collaborative energy diagnosis for mobile devices

We aim to detect and diagnose energy anomalies, abnormally heavy battery use. This paper describes a collaborative black-box method, and an implementation called Carat, for diagnosing anomalies on mobile devices. A client app sends intermittent, coarse-grained measurements to a server, which correlates higher expected energy use with client properties like the running apps, device model, and operating system. The analysis quantifies the error and confidence associated with a diagnosis, suggests actions the user could take to improve battery life, and projects the amount of improvement. During a

deployment to a community of more than 500,000 devices, Carat diagnosed thousands of energy anomalies in the wild. Carat detected all synthetically injected anomalies, produced no known instances of false positives, projected the battery impact of anomalies with 95% accuracy, and, on average, increased a user’s battery life by 11% after 10 days (compared with 1.9% for the control group).

3. METHODOLOGY

Recent years have witnessed a surge in research on data-driven mobile app usage analysis. Starting around 2010, this field grew steadily until 2012, but experienced a remarkable increase thereafter, indicating its burgeoning significance. We plan to conduct a thorough literature survey to capture these developments.

Disadvantages:

- Network service improvement is very less.
- Quality of experience also not good

Our survey provides an in-depth review of recent smartphone app usage analysis research. We cover data collection methods, datasets, and studies in app, user, and smartphone domains, benefiting academia and industry stakeholders.

Advantages:

- The smartphone domain research focuses on smartphone characteristics, with two main areas of investigation: app energy drain and app traffic patterns. These two areas aim to improve smartphone performance by analyzing app energy and traffic consumption patterns.
- User identification, focusing on the individual level, aims to identify a user based on his or her app usage behaviors.

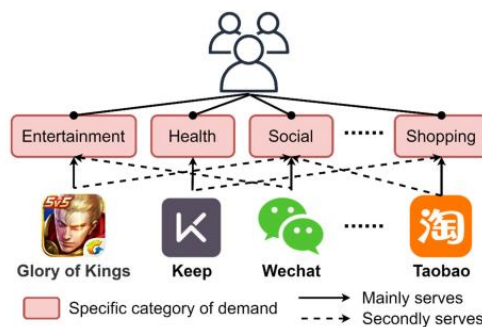


Fig.1: System architecture

MODULES:

To carry out the aforementioned project, we created the modules listed below.

- importing the packages: using this module we will import all packages
- exploring the dataset –English text for analysis: Using this module we will upload dataset
- data processing: Using this module we will read data for processing
- Splitting the data to train and test: Using this module will divide dataset into train & test for processing

- building the model for Smartphone app usage time - Naive Bayes, KNN, Bagging Classifier, Random Forest, Decision Tree, SVM, Voting Classifier, K-Means, DBSCAN, Linear Regression
- building the model for Smartphone app Review analysis based on TfIDF embedding - Logistic Regression, Random Forest, Decision Tree, SVM, KNN, XGBoost, MLP, AdaBoost, Naive Bayes, Voting Classifier
- training the model: Using this module algorithms trained for processing & prediction building the model with Voting Classifier and XGBoost since it gives better accuracy comparing with Other Models around 88%
- Flask Framework with Sqlite for signup and signin: Using this module user will get register & login
- User gives input as Feature Values : Using this module user gives input for prediction the given input is preprocessed for prediction
- a - feature is processed using pandas and numpy arrays
- b - review is tokenized and transformed with tfidf
- trained model is used for prediction: Using this module predicted result displayed
- final outcome is displayed through frontend.

4. IMPLEMENTATION

Random Forest

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

Support Vector Machine

Support Vector Machine(SVM) is a supervised machine learning algorithm used for both classification and regression. Though we say regression problems as well its best suited for classification. The objective of SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points.

Voting Classifier

A voting classifier is a machine learning estimator that trains various base models or estimators and predicts on the basis of aggregating the findings of each base estimator. The aggregating criteria can be combined decision of voting for each estimator output.

Logistic regression

Logistic regression is a statistical analysis method to predict a binary outcome, such as yes or no, based on prior observations of a data set. A logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables.

MLP

A multilayer perceptron (MLP) is a fully connected class of feedforward artificial neural network (ANN). The term MLP is used ambiguously, sometimes loosely to mean any feedforward ANN, sometimes strictly to refer to networks composed of multiple layers of perceptrons (with threshold activation); see § Terminology. Multilayer perceptrons are sometimes colloquially referred to as "vanilla" neural networks, especially when they have a single hidden layer.

MNB

The Multinomial Naive Bayes algorithm is a Bayesian learning approach popular in Natural Language Processing (NLP). The program guesses the tag of a text, such as an email or a newspaper story, using the Bayes theorem. It calculates each tag's likelihood for a given sample and outputs the tag with the greatest chance.

KNN

KNN is one of the simplest forms of machine learning algorithms mostly used for classification. It classifies the data point on how its neighbor is classified. KNN classifies the new data points based on the similarity measure of the earlier stored data points. For example, if we have a dataset of tomatoes and bananas.

XGBOOST

XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements Machine Learning algorithms under the Gradient Boosting framework. It provides a parallel tree boosting to solve many data science problems in a fast and accurate way.

ADABOOST

AdaBoost, short for Adaptive Boosting, is a statistical classification meta-algorithm formulated by Yoav Freund and Robert Schapire in 1995, who won the 2003 Gödel Prize for their work. It can be used in conjunction with many other types of learning algorithms to improve performance.

BAGGING CLASSIFIER

A Bagging classifier is an ensemble meta-estimator that fits base classifiers each on random subsets of the original dataset and then aggregate their individual predictions (either by voting or by averaging) to form a final prediction.

K-MEANS

K-Means clustering is an unsupervised learning algorithm. There is no labeled data for this clustering, unlike in supervised learning. K-Means performs the division of objects into clusters that share similarities and are dissimilar to the objects belonging to another cluster. The term 'K' is a number.

DBSCAN

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a base algorithm for density-based clustering. It can discover clusters of different shapes and sizes from a large amount of data, which is containing noise and outliers.

LINEAR REGRESSION

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.

5. EXPERIMENTAL RESULTS

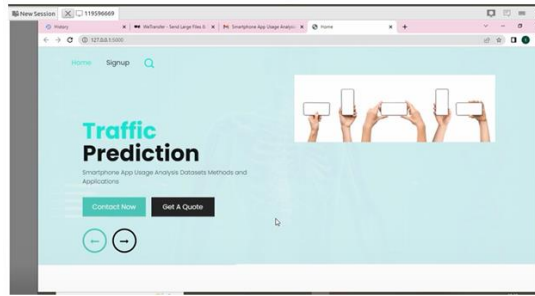


Fig.2: home page

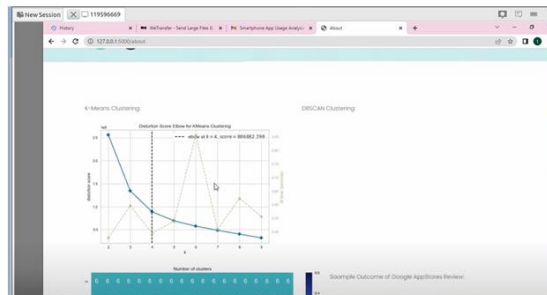


Fig.3: Graph

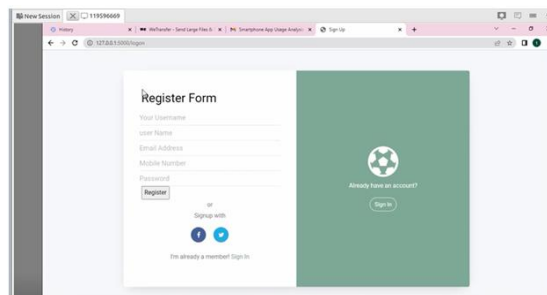


Fig.4: Register page

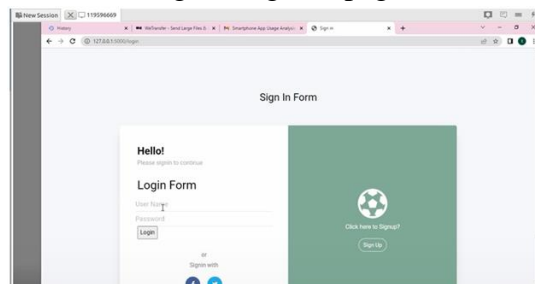


Fig.5: Login Page

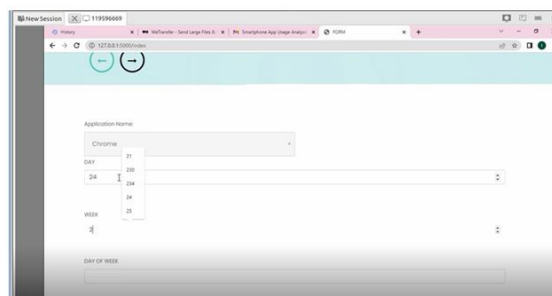


Fig.6: upload input

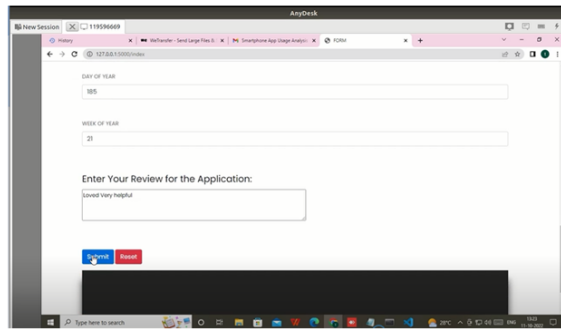


Fig.7: upload input

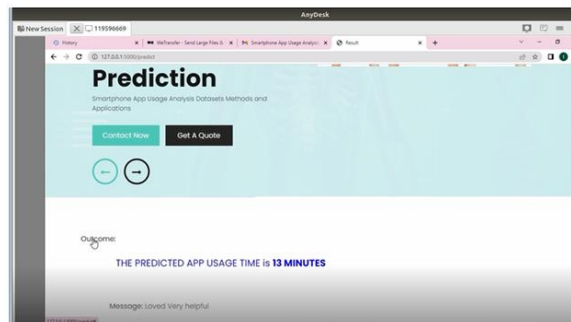


Fig.8: prediction

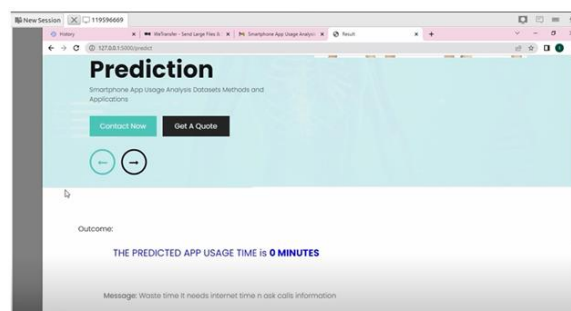


Fig.9: prediction result

6. CONCLUSION

The research efforts on smartphone app usage analysis are surveyed and summarized in this paper. We first introduced and compared various data sources, such as surveys, monitoring apps, network operators, and app stores. For the research community, we presented a set of public datasets and discussed privacy and ethical issues. The related studies in the app, user, and smartphone domains were surveyed, respectively. We made a detailed taxonomy for each research domain based on the problem investigated, the characteristics of the datasets used, the methods used, and the key results obtained. Finally, we discussed two current research challenges and identified five future research directions for this hot topic.

7. FUTURE WORK

The future scope of research on smartphone app usage analysis is promising and multifaceted. Firstly, exploring advanced machine learning and AI techniques to extract deeper insights from app usage data will be essential. Additionally, addressing evolving privacy concerns and ethical considerations is

crucial. Future studies should also focus on the impact of emerging technologies, such as augmented reality and 5G, on app usage patterns. Furthermore, investigating the influence of socio-cultural factors and user behaviors on app adoption and abandonment will be valuable. Finally, collaborations between academia, industry, and policymakers can help shape responsible app usage and enhance user experiences.

REFERENCES

1. N. Oliver et al., “Mobile phone data for informing public health actions across the COVID-19 pandemic life cycle,” *Sci. Adv.*, vol. 6, no. 23, 2020, Art. no. eabc0764.
2. K. Church, D. Ferreira, N. Banovic, and K. Lyons, “Understanding the challenges of mobile phone usage data,” in *Proc. 17th Int. Conf. Human–Comput. Interact. Mobile Devices Services*, 2015, pp. 504–514.
3. T. Li, Y. Li, T. Xia, and P. Hui, “Finding spatiotemporal patterns of mobile application usage,” *IEEE Trans. Netw. Sci. Eng.*, early access, Nov. 30, 2021, doi: 10.1109/TNSE.2021.3131194.
4. Statista. “Number of available applications in the google play store from December 2009 to July 2021.” 2021. [Online]. Available: <https://www.statista.com/statistics/266210/number-of-availableapplications-in-the-google-play-store/>
5. Statista. “Worldwide mobile app revenues in 2014 to 2023.” 2021. [Online]. Available: <https://www.statista.com/statistics/269025/worldwide-mobile-app-revenue-forecasts>
6. J. Liu, Y. Fu, J. Ming, Y. Ren, L. Sun, and H. Xiong, “Effective and real-time in-app activity analysis in encrypted Internet traffic streams,” in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Disc. Data Min.*, 2017, pp. 335–344.
7. Y. Chen, X. Jin, J. Sun, R. Zhang, and Y. Zhang, “POWERFUL: Mobile app fingerprinting via power analysis,” in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, 2017, pp. 1–9.
8. A. J. Oliner, A. P. Iyer, I. Stoica, E. Lagerspetz, and S. Tarkoma, “Carat: Collaborative energy diagnosis for mobile devices,” in *Proc. 11th ACM Conf. Embedded Netw. Sensor Syst.*, 2013, p. 10.
9. F. Xu, Y. Li, H. Wang, P. Zhang, and D. Jin, “Understanding mobile traffic patterns of large scale cellular towers in urban environment,” *IEEE/ACM Trans. Netw.*, vol. 25, no. 2, pp. 1147–1161, Apr. 2017.
10. M. Zeng et al., “Temporal–spatial mobile application usage understanding and popularity prediction for edge caching,” *IEEE Wireless Commun.*, vol. 25, no. 3, pp. 36–42, Jun. 2018.