The 35th Annual Conference of the Japanese Society for Artificial Intelligence, 2021

スマートフォンアプリの起動ログデータについての分析手法の検討

Study of Analysis Methods for Startup Log Data of Smartphone Applications

和田　伸一郎[*1]
WADA, Shinichiro

川畑　泰子[*2]
KAWAHATA, Yasuko

立教大学大学院社会学研究科[*1]
Graduate School of Sociology, Rikkyo University.

立教大学大学院社会学研究科[*2]
Graduate School of Sociology, Rikkyo University.

Abstract: In this paper, using the application launch log data (10 segments of each gender age) of 5000 smartphone users (about 2.4 million lines for one week in January 2019), we conducted an analysis to identify the characteristics and differences of smartphone usage in each gender age. In this study, we conducted a marketing basket analysis using the Apriori algorithm in "R" to reveal the access transitions between applications, and compared the network structure of each segment in parameters of confidence and support. Results of fine tuning of the parameters, we were able to remove a very small number of applications with extremely high number of launches, or many applications with extremely low number of launches as results. In this paper, we will discuss the differences between applications used by each age groups and gender.

1.　はじめに

本研究では，フラー社（千葉県柏市）との共同研究において、約 240 万行のスマートフォンユーザーのアプリ起動ログデータの提供を受け、そのデータを分析することによって各性年代においてスマートフォンの利用動向のそれぞれの特徴とはどのようなものか、その差異とはどのようなものかを明らかにすることを目的とする。

2.　データセット

2.1　概要

本発表で使用したのは、フラー社（千葉県柏市）から、特別の協力を得て提供を受けた、2019 年 1 月 25 日から 1 月 31 日の一週間の 5000 ユーザーの 1 秒単位のスマートフォンアプリ起動ログデータ(2,436,141 万行)である。なお、本データは、フラー社が Android ユーザーより許諾を得たうえで、ユーザー個人が特定できないようにランダムの ID 割り振った上で取得されたものである。また、ユーザーの性年代の構成が均等になるよう、10 代男女から年代ごとに 50 代以上の男女まで、計 10 セグメント、500 名に揃えられた上での 5000 ユーザーのデータとなっている。起動ログが含まれる本データでのアプリの対象となるのは、Google 製の AndroidOS にインストールされているもの、つまり Google Play から、2019 年 1 月時点でインストール可能だった 7736 個のアプリである（ただし後述するように、一部のアプリは除去した）。

2.2　前処理

アプリは、Google 社によって 27 種類のカテゴリーに分けられている。しかし以下の理由から、分析対象から除外したアプリがある。本論での研究成果では、各性年代の利用動向の特徴を明確にするという目的を優先することから、以下のアプリデータを除去した。2つのカテゴリーである PERSONALIZATION(ランチャー系)、 TOOLS(システム系)のアプリログデータを除外した。また、上記の目的から同じく、17 カテゴリーあるゲームアプリのログデータも除外し、最終的に 30 カテゴリー、4688 のアプリを分析対象とした。

分析手法

今回、分析手法として、マーケティング・バスケット分析（アソシエーション分析）を用いた。アプリの利用動態・遷移過程に関して、ネットワーク分析が可能であり、有向グラフを考察することにより、各性年代それぞれの特徴およびパラメータにおける考察ができると考えた。なお、マーケティング・バスケット分析を行うにあたって、Exploratory(R 言語)における arules パッケージ（Apriori アルゴリズム）を使って分析し、その後 igraph を使って、有向グラフを生成した。

3.1 バスケット分析

本分析で用いたバスケット分析では、各性年代のアカウントの 1 週間のアプリ起動ログデータを用いた。分析対象とするアプリ起動ログIを(1)(2)の通り設定した。

$$I = \{i_1 i_2, \cdots\cdots\cdots, i_k\} \tag{1}$$

(1)はつまり、以下のように記述できる。

$$I = \{AppID_1 AppID_2, \cdots\cdots\cdots, AppID_k\} \tag{2}$$

対応するトランザクションに関しても

$$T = \{t_1 t_2, \cdots\cdots\cdots, t_k\} \tag{3}$$

(3)のように記述できる。t_kは(2)同様に分析対象ユーザー

連絡先:和田伸一郎 立教大学大学院社会学研究科, 東京都豊島区西池袋３－３４－１

単位に次に遷移する$UserID_1 UserID_2, \cdots\cdots, UserID_k$を指す。

本分析で求められる支持度(Support)は、仮にA→Bへのアプリへのアクセス遷移があるとした場合、全体のアクセスログ件数(N)とした中で算出する仕組みとなっており、

$$Support(A \Rightarrow B) = \frac{|A \cup B|}{N} \tag{4}$$

(4)のように記述できる。そして、あるアプリに対するアクセスの信頼度(Confidence)に関しては

$$\text{Confidence}(A \Rightarrow B) = \frac{Support(A \cup B)}{Support(A)} \tag{5}$$

と記述することができる。また本分析では、リフト値(6)を基準に上位の(4)(5)のスコアをルール内に入れるようにしている。

$$\text{Lift}(A \Rightarrow B) = \frac{Support(A \cup B)}{Support(A) \cdot Support(B)} \tag{6}$$

3.2 パラメータ・チューニング

本分析では、データが大きいため、適切な特徴量を抽出するには、パラメータ・チューニングを行う必要があり、その結果、一定程度の適正な値を特定した。

まず、データが大きすぎると、しばしばメモリーエラーなどが起きうる。これについては、Apriori アルゴリズムの強みが発揮された。例えば、10 アイテムあるとする。この場合、1 アイテム同士の組み合わせが 90 種類となり、すべての組み合わせを計算しようとすると、約 5 万 7000 回の大きな共起計算が必要となる。このデータの場合、先述したアプリを除外した上でも、5000 人(バスケット)のユーザー(購買者)がいるので、膨大な計算になってしまい、事実上計算は非常に困難になる。これに対し、このアルゴリズムでは「支持度が一定以下のアイテムを含む組み合わせは、最初から確信度を計算しない」といった省略を行うことで、計算を高速に処理することを実現可能にすることができる。次に、適切な特徴量を抽出するためには、データは大きければ大きいほどいいということがあるとはいえ、研究対象としたデータはべき乗分布となっているため、いかに極端に高頻出するアプリ(例えば LINE など)、あるいは極端に低頻出のアプリを除外した上で分析を行うかが、問題となってくる。これを行うために、以下のようなパラメータ・チューニングを行った。

例えば、40 代女性データで、ルールの数を多めにとって 700 にすると、この性年代の起動総回数の 75 % の支持度(アプリ A からアプリ X へと遷移したユーザーの割合)をカバーできる。しかし、トランザクションの数が多くなりすぎて、特徴量が見えなくなってしまう。これを解決するために、カバー率を減らして、最適値を探した。その結果、ルール数＝90 ～120、確信度を 0.5 にすると、支持度＝ 約 6 ～27 % となり、特徴量がよりはっきり見られ、これらの値が適度な範囲であることが分かった。

3.3 グラフの説明

以下の図 1 は 40 代女性の場合の計算結果である。アソシエーションのルール数を 120 とした。これはノード(バブル)の数を 120 にすることを意味する。ノードの大きさは、支持度(support)を表し、ノードの色の濃さは確信度(confidence)を表す。また、アプリ名からノードに向かう矢印は、「これらのアプリが使われると･･･」という条件を表し、ノードからアプリ名に向かう矢印は「アプリが一緒に(あるユーザーのクラスタ内で)利用される傾向にある」という結果を表す。

また、このグラフの描画に使われている Fruchterman-Reingold アルゴリズムは、その数理的性質から「関連性の強いもの同士が近くなるように配置する」傾向がある。逆に言えば、関連性の弱いものが遠くなる結果をもたらすと考えられる。このグラフが全体的にばらけて見やすく描画されているのは、これゆえだと思われる。

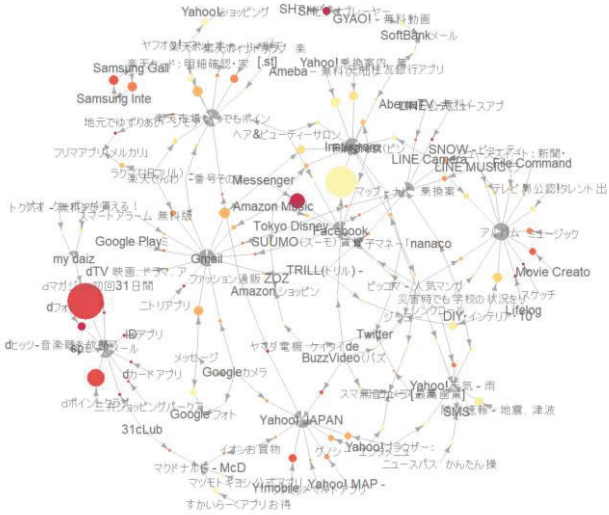

図1 40 代女性における計算結果

3.4 分析内容

40 代女性(50 代_60 代女性にも同様の傾向が見られた)の場合、他の性年代と比較すると、ショッピング／ポイントアプリが多いことが分かる。マクドナルド、すかいらーく、ニトリ、ファミリーマート、ヤマダ電機、トクバイ(無料チラシ)などが見られ、それらの多くは、メールアプリ、あるいはブラウザへと遷移していることも分かった。

他の性年代間でも、グラフから、それぞれに固有の大きなものから小さなものまで様々な特徴があることが分かった。このことから、各性年代で、アプリの利用動向が異なり、一定程度において、それぞれの生活パターンが推察できることも分かった。

今後の課題

通知オンにしているアプリとそうではないアプリとでは、起動回数に大きな差が出てくると考えられる。ただし、これらについ

ては、ログデータからは分からないので、別途アンケート調査などを実施する必要があるだろう。

次に、今回の分析は時系列で行っていないため、あるアプリを使った前後にどのアプリを使っているかを明らかにできなかった。これについては、分析手法から検討を加える必要があるだろう。

今後の予定

本共同研究では、フラー社より 2020 年 1 月から 9 月のデータもすでに提供を受けている。したがって、今後の予定としては、スマートフォンアプリログデータからコロナ禍での人々の生活スタイルの変化を分析する予定である。

【謝辞】:本発表は、フラー社(千葉県柏市)と立教大学社会学研究科木村忠正研究室との共同研究からの多大な協力を得た。また、元データからの加工データ作成にはフラー社の大野康明氏からたいへんなご尽力をいただいた。記して感謝したい。そのほか、共同研究の関係者の方々に対してもあわせて、記して感謝したい。

参考文献

[Agrawal 94] Agrawal, Rakesh, Srikant, Ramakrishnan.: Fast algorithms for mining association rules. VLDB, Vol.1215, pp.487-499(1994)

[Agrawal 96] Agrawal, Rakesh, Mannila, Heikki, Srikant, Ramakrishnan, Toivonen, Hannu, Verkamo, A Inkeri.: Fast discovery of association rules. Advances in knowledge discovery and data mining, Vol.12(1), pp.307-328(1996)

[Deng 19] Deng, Tao, Kanthawala, Shaheen, Meng, Jingbo, Peng, Wei, Kononova, Anastasia, Hao, Qi, Zhang, Qinhao, David, Prabu.: Measuring smartphone usage and task switching with log tracking and self-reports, Mobile Media & Communication, Vol. 7(1), pp.3-23 (2019)

[Do 11] Do, Trinh Minh Tri, Blom, Gatica-Perez, Jan, Daniel.: Smartphone usage in the wild: a large-scale analysis of applications and context ICMI '11: Proceedings of the 13th international conference on multimodal interfaces, pp.353-360 (2011)

[Ertek 06] Ertek, Gurdal, Demiriz, Ayhan.: A Framework for Visualizing Association Mining Results. Lecture Notes in Computer Science, Vol.4263, pp.593-602 (2006)

[Fruchterman 91] Fruchterman, Thomas MJ, Reingold, Edward M.: Graph drawing by force‐directed placement. Software: Practice and experience, Vol.21(11), pp.1129-1164(1991)

[Hahsler 11] Hahsler, Michael, Chelluboina, Sudheer, Hornik, Kurt, Buchta, Christian. The arules R-package ecosystem: Analyzing interesting patterns from large transaction datasets. Journal of Machine Learning Research, 12:1977–1981 (2011)

[Taddy 19] Taddy, Matt.: BUSINESS DATA SCIENCE: Combining Machine Learning and Economics to Optimize, Automate, and Accelerate Business Decisions. McGraw-Hill (2019)