

スマートフォンのアプリケーション利用履歴を用いた インフルエンサ推定とアプリケーション利用予測

スマートフォンのアプリケーション利用履歴から、個人が周囲の人へ与える影響を分析し、将来の利用を予測する。本検討では、利用者がアプリケーションをダウンロードし実行する順番は個人間の影響度によって表現できるというモデルと、影響関係には潜在グループ構造があるという仮定を利用している。約160人の大学生から収集した利用履歴により仮説を検証し、高い予測精度が得られることを示す。なお、本検討は大阪大学サイバーメディアセンターに設置した共同研究部門における共同研究により実施した。

サービス&ソリューション開発部 かたぎり まさじ 片桐 雅二 えとう みのる† 栄藤 稔

1. まえがき

スマートフォンが急速に普及しつつあり、Android™*1アプリケーションは2011年中に40万種類を超えたと言われ、今後一層多様化し、種類は増え続けるであろう。各利用者に対してアプリケーションを適確に選び出し推薦することが、利用者層を広げていくうえでも極めて重要である。

一方、周囲の人への影響力が強い人（インフルエンサ）を見つけ出して重点的に働きかけることで、効率的に製品／サービスなどの普及を図るマーケティング手法が注目されている。インフルエンサは図1に示すように周囲に影響を与えるため、有

効なインフルエンサの把握ができれば、増収、ビジネス拡大、トレンド予測などに効果があるといわれている。

本稿では、このインフルエンサの考え方を応用して、スマートフォン・アプリケーションのより良い推薦の実現を試みる。具体的には、利用者がアプリケーションをダウンロードし実行する順番は個人間の影響度によって表現できるというモデルを仮定して、スマートフォンのアプリケーション利用履歴から利用者間の影響関係を見つけ出す。その影響関係を用いてアプリケーションの利用予測をすることで、より良い予測性能の実現を試みる。

本検討では、大阪大学との共同研

究により約160人・6カ月間の学生実験を行い、収集した利用履歴を用いて提案モデルの能力を評価し仮説を検証した。結果から、利用者間の影響関係に潜在的なグループ構造が存在すること、その活用がアプリケ

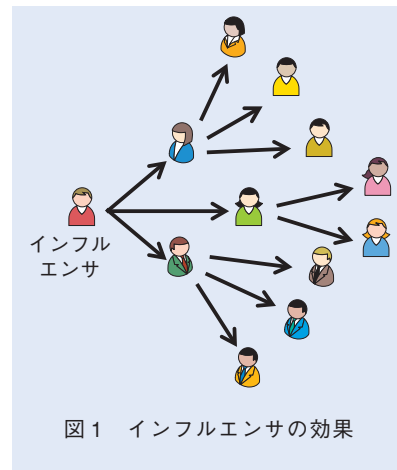


図1 インフルエンサの効果

© 2012 NTT DOCOMO, INC.
本誌掲載記事の無断転載を禁じます。

† 現在、研究開発推進部

*1 Android™：米国Google, Inc.が提唱する携帯端末を主なターゲットとしたオープンソースプラットフォーム。Android™は、米国Google, Inc.の商標または登録商標。

ーション・ダウンロード予測能力の向上に有効であることが示され、既存の代表的予測手法と比較して優れた予測が実現できることを確認した。なお、さらなる詳細については文献[1]を参照されたい。

2. 個人間の影響関係モデル

2.1 個人間の影響関係

アプリケーションを使い始める場面を想定すると、友人知人から紹介されてということもあるが、マスメディアからの情報や、その他の多様なきっかけによることが多い。このため、影響を与える先を友人知人などの情報交換を直接行う相手のみと考えて、影響関係をモデル化することは不自然である。このためここでは、人は直接関係のない人に対しても影響を与えていると考え、直接情報交換を行う相手だけでなく、直接関係ない人も含めた影響関係をモデル化することにする。これは2人の利用者を考えて、一方の利用者がいるアプリケーションを先に使い、他方の利用者が追従することが多いとき、この2人の間には強い影響関係があるとみなすことになる。このような個人間の順序関係に着目して購買などの予測をすることの有効性は川前らが文献[2]で指摘している。

アプリケーションを推薦する場合、その利用者が近い将来ダウンロードするアプリケーションを予測し推薦することが一般的である。このために現在用いられる代表的な手法

は協調フィルタリング (CF: Collaborative Filtering)^{*2}であり、さまざまな拡張が研究されている[3]。ここで説明のため、利用者がいくつかのアプリケーションをダウンロードし、初めて実行した時刻の関係の一例を図2に示し、利用者1が次にダウンロードするアプリケーションを予測することを考える。図2では3人の利用者および5つのアプリケーションを扱っており、水平軸は時刻を表している。円で囲まれた数字はその時刻に数字の示すアプリケーションが、対応する利用者によりダウンロードされたことを意味する。CFでは、履歴中の順番を考慮しないので、利用者2と3は利用者1との類似度に差がなく、利用者1が次にダウンロードするアプリケーションとして、アプリケーション4と5は等しく推薦される。しかし、ダウンロードの順序に着目すれば、利用者2は明らかに利用者1の先行者すなわちインフルエンサである。また、利用者3は1の追従者である。

このため、アプリケーション5のほうが4よりも利用者1にダウンロードされる確率が高いと考えられる。この例が端的に示すように、利用者の振る舞いに先行者/追従者の関係・傾向が存在する場合には、個人間の影響を用いることが有効である。直感的にも情報感度の高い利用者が早期にアプリケーションを開拓し、一般的な利用者が追従する傾向は自然であるので、有効性が期待できる。

2.2 モデルの定式化

ここでは、アプリケーション・ダウンロードの連鎖の発生を、ベルヌーイ試行^{*3}を基本とした確率モデルで定式化する。またアプリケーションには、万人が使うものから、ごく一部の利用者しか使わないものなど多様なものが存在する。個人の嗜好性をより反映し影響関係を捉えやすくするため、エントロピー (情報量) による重み付けを導入して、利用者uが利用者vへ与える影響力

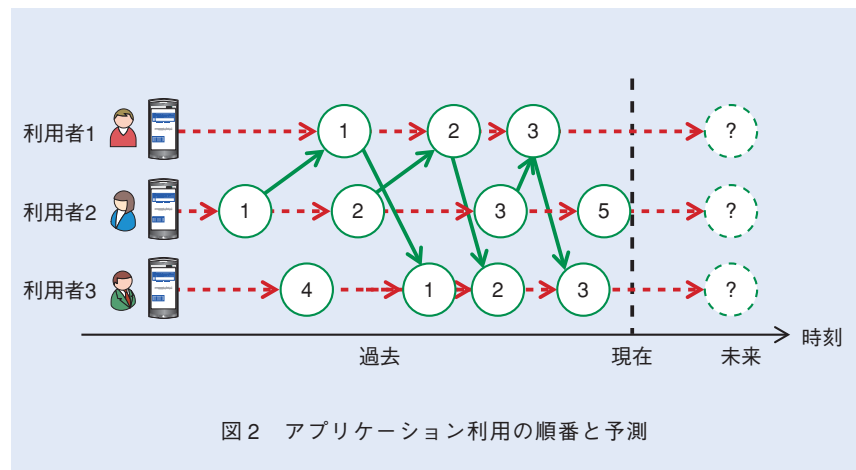


図2 アプリケーション利用の順番と予測

*2 協調フィルタリング: 多数のユーザについて、購買履歴などの嗜好情報をあらかじめ蓄積しておき、目的のユーザと嗜好の類似した他のユーザの情報を用いて、そのユーザに対する予測や推薦を行う手法。

*3 ベルヌーイ試行: 最も基本的な確率モデル。一定の確率で表が出るコインを投げて、どちらの面が出るかを観察することに相当する。

$P_i(u \rightarrow v)$ を式(1)と定義する.

$$P_i(u \rightarrow v) = \frac{\sum_{a \in A_{u \rightarrow v}} (-\log(u_a/u_{glob}))}{\sum_{a \in A_u} (-\log(u_a/u_{glob}))} \quad (1)$$

ここで、 A_u は u がダウンロードしたアプリケーションの集合、 $A_{u \rightarrow v}$ は v より先に u がダウンロードしたアプリケーションの集合、 u_a と u_{glob} はアプリケーション a の利用者数および全利用者数である。すべての利用者どうしの影響力 $P_i(u \rightarrow v)$ を u 行 v 列の要素にもつ、サイズ(利用者数×利用者数)の行列を影響力行列 R と呼ぶ。

独立カスケードモデル^{*4}[4]の考えに従い、得られた影響力行列 R を用いれば、利用者 u がアプリケーション d をダウンロードする同時影響確率は、式(2)のように定式化することができる。

$$P_{in}(d|u) = [1 - \prod_{u'} \{1 - P_i(u' \rightarrow u)\}] \quad (2)$$

ここで、 u' は、 u より先に d をダウンロードした利用者である。

この $P_{in}(d|u)$ は、利用者 u がアプリケーション d をダウンロードする予測確率なので、この値が大きいアプリケーション d を順に利用者 u へ推薦すれば良い。

3. 潜在グループ構造

2章で述べたモデルに従って、予測を履歴から計算しようとする、すべての利用者の組に対しその影響関係を安定して得るために、双方でダウンロードしたアプリケーション

が一定数以上必要となる。しかし、すべての利用者の組において共通のアプリケーションを多数ダウンロードしているとは限らず、また新規の利用者に多くの履歴は期待できないので、何らかの工夫が必要となる。このような状況に対応するため、川前らは文献[2]にて、影響関係をマルコフ過程^{*5}と見なしエルゴード性^{*6}を利用する手法を用いたが、多段連鎖による関係性を盛り込むのが目的であり、観測の不足を補うことにはならない。これに対し本検討では、影響を与えあう関係中に潜在的なグループ構造を仮定することにより、見かけ上の影響関係を安定に推定し、より良い予測・推薦ができることを示す。筆者の知る限り類似の取組みはこれまでに報告されていない。

一般にマーケティングでは、人の嗜好や行動を捉えるために、セグメント化して捉えようとすることが多い。アプリケーションのダウンロードにもこのセグメントに相当する潜在的なグループがあるという仮説をたて、検証を行った。潜在的なグループ構造を仮定すると、グループ単位で類似の行動をとるという仮説からグループ内のユーザ間で行動が補完を行え、個々のデータが少ない場合でも予測性能の安定化が期待できる。検証に用いた利用履歴データについて次節で述べる。

3.1 利用履歴データ

大阪大学と共同で、利用状況を記録する実験用ソフトウェアを搭載し

たスマートフォン(XperiaTM*7)を約160名の大学生実験協力者に6か月間貸与し、自由に使用するよう指示して利用履歴を収集した。実験協力者は学内学生に対し公募した。募集および応募受付の際に実験の趣旨を説明し、データの取扱い(個人情報保護)と研究目的の利用について許諾を得た。

利用履歴は端末上で匿名化処理を行い、3Gネットワーク経由でサーバに集められる。利用履歴は、<時刻、匿名化ユーザID、実行したAndroidアプリケーションのパッケージ名>の3項組のレコードで構成される。

収集した利用履歴から、各利用者において各アプリケーションの初回実行分のレコードを抽出した。図3に初回実行レコードの実験開始からの累積数の推移を示す。モデル評価を行うために図3の状況を勘案し、学習セット^{*8}を89日間分(2011年2~4月)、テストセット^{*9}を31日間分(2011年5月)抽出した。さらに利用者数が3未満のアプリケーションは有効な連鎖の推定が困難なため対象から削除した。これらの処理により学習セットは3,383レコード(155利用者、291アプリケーション)、テストセットは249レコード(98利用者、116アプリケーション)となった。

3.2 潜在グループ構造の検証

潜在グループを抽出するために、

*4 独立カスケードモデル：情報などが伝搬・拡散していく現象の代表的な確率モデルの1つ。確率的伝搬の連鎖が独立に確率的に生じる形を表現するモデル。

*5 マルコフ過程：複数の取り得る状態が決まっていて、ある時刻の状態から次の時刻の状態への遷移が、現在の値によって確率的に決まる過程。

*6 エルゴード性：マルコフ過程において、状態遷移を何度も繰り返すと、最初の状態に関係なく一定の状態確率分布に収束する性質。

*7 XperiaTM：Sony Mobile Communications ABの商標または登録商標。

*8 学習セット：正解が分かっている訓練用のデータから特性を獲得し、それに基づいて新たなデータに対する予測を行う手法を評価する場合、予測能力を評価するデータとは別に用意する、特性を獲得するための正解付きの訓練用データのこと。

影響力行列Rに対して、行列分解^{*10}を用いた低ランク近似^{*11}を行った。影響力行列Rの要素はアプリケーション・ダウンロードの連鎖の生起確率を意味するため、各要素の値は非負である必要がある。このため、行列分解にはNMF (Non-negative Matrix Factorization)^{*12}[5]を用いて要素の値が非負となるように制限した。NMFの正則化項には、提案されている代表的なものを試し、最も結果の良いものを用いた。

仮説の検証は、グループ数を変化させアプリケーション・ダウンロードの予測精度を観察することにより行った。結果を図4に示す。予測精度にはperplexity^{*13}という尺度を用いた。perplexityは値が小さいほど高い予測精度を意味する。潜在グループ構造がない場合は、一般的にグループ数が多いほど表現能力が高まり予測精度も高まる。これに対し、図4からグループ数が6程度において予測精度のピークがあることが確認できる。これは、影響力関係において潜在グループの存在を意味している。

次に、影響力行列Rで表される関係の可視化を試み、低ランク近似なしの結果を図5(a)に、低ランク近似ありの結果を図5(b)に示す。図5(a)(b)は影響力が一定値以上の関係のみを表示対象として、一般的なグラフ可視化手法であるForce-Directedレイアウト^{*14}により構造を可視化したものである。各ノードは利用者を表し、エッジが一定値以上

の影響力の存在を表す。図5(b)のノードの色は各利用者の所属する潜在グループを示している。グラフ上多数のノードへ向けたエッジをもつノードがインフルエンサである。低ランク近似しないと容易に見いだせないグループ構造が、低ランク近似により抽出されたことが確認でき

る。あわせて図5(b)からは、同一潜在グループ内での影響関係と異なる潜在グループ間での影響関係が混在していることが観察できる。

4. ダウンロード予測精度の比較

影響力行列を用いた提案手法の予

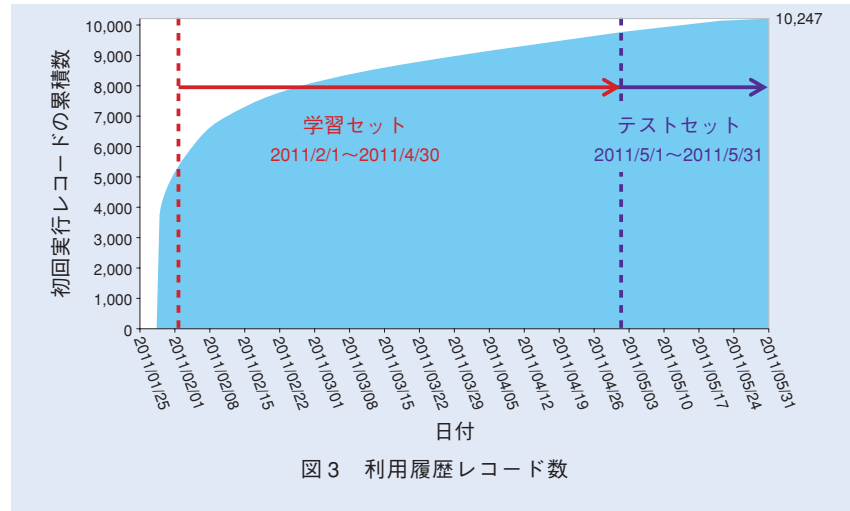


図3 利用履歴レコード数

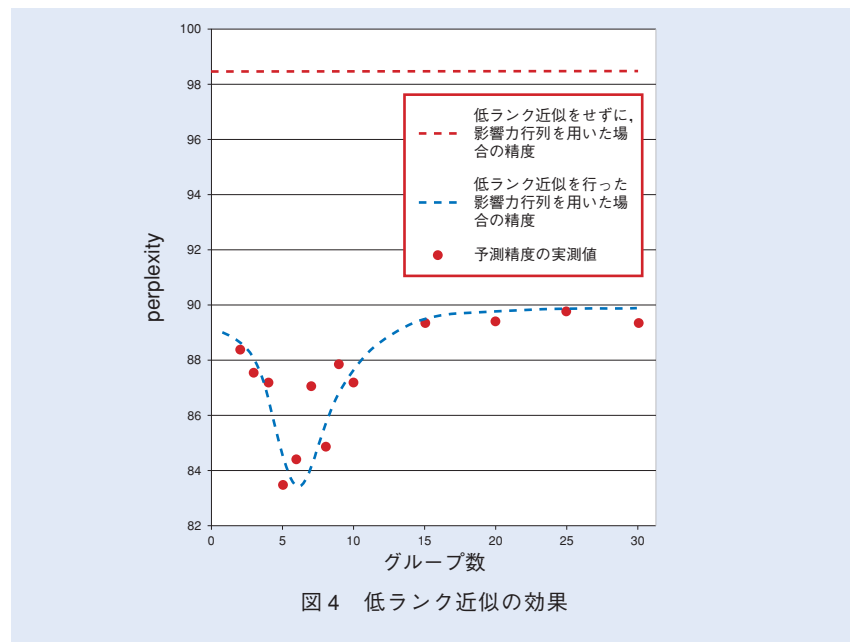


図4 低ランク近似の効果

*9 テストセット：正解が分かっている訓練用のデータから特性を獲得し、それに基づいて新たなデータに対する予測を行う手法を評価する場合、特性を獲得するためのデータとは別に用意する、予測能力を評価するための試験用データのこと。
*10 行列分解：(m×n)次元の行列を、(m×k)と(k×n)次元の2つの行列の積で近似

するように、行列の因数分解をすること(通常はk≪m, n)。
*11 低ランク近似：任意の行列をより低いランクの行列で近似すること。行列分解で得られる2行列の積は、ランクがkの近似行列となる。
*12 NMF：行列分解の一手法。与える行列も、分解して得られる行列も、要素がす

べて非負の値をもつという制約のうえで行列分解を行う。
*13 perplexity：主に言語モデルの評価に用いられる定量的評価尺度で、厳密にはtest-set perplexityと呼ばれる尺度。直感的には、モデルが与える出現確率と実際に観察された頻度との乖離している程度を表す。

測精度と、代表的な既存手法（人気ランキングによる手法、CFの代表的手法であるユーザベースCF）の予測精度を比較した。予測精度は前章と同様に perplexity により評価した。結果を図6に示す。低ランク近似無しの影響関係を用いた予測精度は十分でないが、低ランク近似した影響関係を利用すると、既存手法を上回る性能が得られ、提案手法の有効性が確認できた。

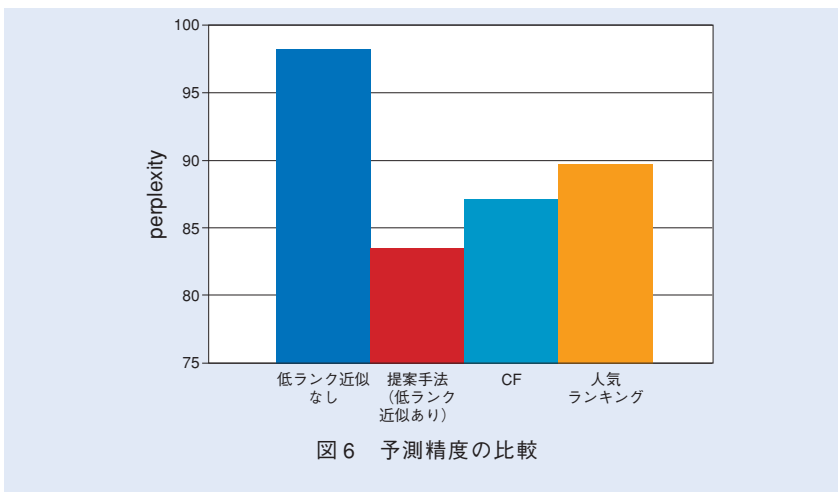
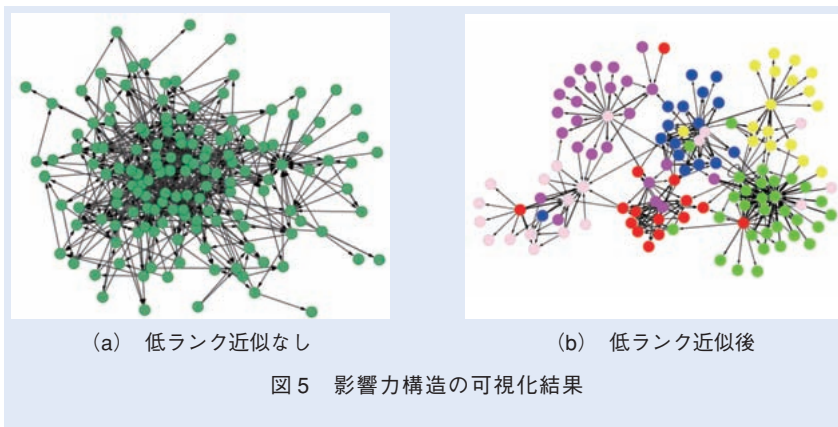
5. あとがき

本稿では、利用者がアプリケーションをダウンロードし実行する順番は個人間の影響度によって表現できるというモデルを仮定して、個人間の影響関係によるモデルを定式化するとともに、約160人の大学生による6カ月間の実験で得た利用履歴を用いて、潜在グループ構造が存在することを確認した。さらに潜在グループ構造を活用すると、アプリケーション利用の予測能力が向上し、既存手法を上回ることを示した。

今後は、より大規模データに対応できる効率的な処理手法の開発を行い、実フィールドへの適用可能性および効果の検証を進めていきたい。また、抽出された影響関係から得られるインフルエンサをマーケティング施策に適用した場合の有効性検証は今後の課題である。

参考文献

[1] 片桐 雅二, 榮藤 稔: “Social Influence



Modeling on Smartphone Usage,” Proc. of the 7th International Conference on Advanced Data Mining and Applications - Part II, pp.292-303, Dec. 2011.

[2] 川前 徳章, 坂野 鋭, 山田 武士: “Personalized recommendation based on the personal innovator degree,” Proc. of the third ACM conference on Recommender systems, RecSys '09, ACM, pp.329-332, Oct. 2009.

[3] G. Adomavicius and A. Tuzhilin : “Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions,” IEEE Trans. on Knowledge and Data Engineering 17, pp.734-749, Jun. 2005.

[4] J. Goldenberg, B. Libai and E. Muller : “Talk of the Network: A Complex Systems Look at the Underlying Process of Word-of-Mouth,” Marketing Letters, pp.211-223, Aug. 2001.

[5] D.D. Lee and H.S. Seung : “Learning the parts of objects by non-negative matrix factorization,” Nature, Vol.401, No.6755, pp.788-791, Oct. 1999.

* 14 Force-Directed レイアウト: グラフ構造データを美しく描画する手法の一つ。グラフの頂点と辺に仮想的な力を割り当て、力学的に安定となる配置を探索する。グラフの辺の長さがほぼ等しく、辺ができるだけ交差しないような配置が得られる。