スマートフォンの利用履歴を用いた
コミュニケーション構造推定に関する研究

提出先　　大阪大学大学院情報科学研究科

提出年月　2013年1月

片桐　雅二

研究業績目録

I. 学術論文

[1] 片桐 雅二, 栄藤 稔, 竹村 治雄 : スマートフォンアプリケーション実行ログからのインフルエンサ推定, 情報処理学会論文誌 データベース, Vol. 5, No. 3, pp. 75–85, 2012 年 9 月.

[2] Masaji KATAGIRI, and Minoru ETOH: Implicit Influencing Group Discovery from Mobile Applications Usage, IEICE Transactions on Information and Systems, Vol. E95-D, No. 12, pp. 3026–3036, December 2012.

II. 国際会議（査読あり）

[1] Masaji KATAGIRI, and Minoru ETOH: Social Influence Modeling on Smartphone Usage, Proceedings of the 7th International Conference on Advanced Data Mining and Applications (ADMA 2011), Part II, pp. 292–303, Beijing, China, December 2011.

[2] Yuka IKEBE, Masaji KATAGIRI[*1], and Haruo TAKEMURA: Friendship Prediction using Semi-Supervised Learning of Latent Features in Smartphone Usage Data, Proceedings of the 4th International Conference on Knowledge Discovery and Information Retrieval (KDIR 2012), pp. 199–205, Barcelona, Spain, October 2012.

III. 国内シンポジウム等（査読あり）

[1] 片桐 雅二, 栄藤 稔: スマートフォンアプリ実行ログからのインフルエンスグループの発見によるインフルエンサとイノベータの推定, 第 4 回 Web とデータベースに関するフォーラム (WebDB Forum 2011), 2011 年 11 月.

[*1] Corresponding author

IV. 機関論文誌

[1] 片桐 雅二, 栄藤 稔: 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, NTT DOCOMO テクニカル・ジャーナル, Vol. 20, No. 2, pp. 54–58, 2012 年 7 月.

V. その他の発表

[1] Masaji KATAGIRI: Social Modeling through Mobile Application Usage, Proceedings of the 3rd International Workshop on Mobile Information Retrieval for Future (MIRF 2011), pp. 89–95, November 2011.

内容梗概

ここ数年において急激にスマートフォンの普及が進んだ．高速な移動通信技術の普及とも合わさり，モバイルにおける様々な処理がクラウド形式にて実現可能となってきた．これらが社会の様々な部分に浸透することにより、多種多様な情報を処理可能な電子データとして扱い価値化しようとするビッグデータの流れが顕著となってきている．さらには，これらの情報の価値化によって最終的には現実世界/実社会をより良くするための貢献が期待されていることをより強く意識し，CPS（Cyber Physical System）の重要性が認識されてきている．しかしながら，CPS の実現には，技術の進歩と合わせて，社会およびその主役である人間についての理解を深めることが必須である．

本論文では，スマートフォンの利用履歴を分析・活用することによって，利用者および社会を理解することを目指して行った研究について述べる．特に，社会は人と人との間にコミュニケーションがなされて構成されていることに着目し，コミュニケーション構造の推定を試みる．スマートフォンは従来の携帯電話端末と比較して多機能であり，より多くの日常的活動に活用されることが期待されるため，その利用履歴には個人や社会を表す有用な情報が含まれることが想定される．そこでこの利用履歴を分析することにより，誰と誰との間にどのようなコミュニケーションがあるのかを推定する．具体的には，クチコミ情報を発信しているのは誰で，その影響を受けやすいのは誰か．あるいは直接交流（コミュニケーション）のある友人は誰と誰かを推定する．

スマートフォンの利用傾向とコミュニケーション構造の間の関係は明らかでないため，本研究においては，スマートフォンを用いた大規模なモニター実験（約 160 人，6 か月間）を行い実履歴データを収集するとともに，アンケート調査もあわせて行い，コミュニケーション状況を把握し，これらを用いて仮説検証型の解析研究を行う．

まずは，スマートフォンにおけるアプリケーションの利用順序には，社会構造（人間関係）による影響（インフルエンス）が反映されているとする仮説に基づき，潜在特徴モデルを構築すると，利用順序の予測が高精度に可能となることを示す．また，潜在特徴モデルと予測精度の関係を考察し，影響関係においては潜在グループ構造が存在することを示す．潜在特徴の獲得には非負行列分解（Nonnegative Matrix Factorization; NMF）による低ランク近似を用いる．予測精度の向上には，NMF により得られる 2 つの因子行列について，その値がほぼ 0 となる

要素の頻度を高めるスパースネス制御が重要であることを示し，結果的にモニター実験のデータでは 5~6 の潜在グループの存在が確認できるとともに，予測精度としては既存の各種の協調フィルタリング手法を上回る精度が得られることを示す．

次に，スマートフォンのアプリケーション利用履歴と友人関係情報が既知であるとして，実際にクチコミ等によって周囲に直接影響を及ぼすインフルエンサを推定・抽出する手法を提案する．上述のアプリケーション利用予測において得た影響関係は，利用順序を説明するモデルであるため，新しいものを好んで試用する傾向の人が広く影響を及ぼしているように表現される．しかしながら，このような人が必ずしも周囲に直接影響を及ぼしているとは限らないので，ここでは周囲に直接影響を及ぼしている利用者の推定・抽出を試みる．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を，個人間の影響度をパラメータとする確率モデルにより表現する．利用者同士に友人関係が有る場合にのみ情報伝達が起こり直接影響を与え得ることに着目し，個人間の影響度を直接影響とそれ以外の要素の混合として表す．ダウンロードの連鎖の生起確率密度分布をベータ分布によりモデル化し，MCMC 法を用いて母数推定を行う．そして獲得された直接影響の大きさからインフルエンサを推定する．モニター実験により得たデータを用いて，提案手法の正当性および効果を確認する．行動が早いという先行性指標および既存の手法よりも提案手法が推定能力において優れていることを確認する．

さらには，スマートフォンの利用履歴を用いて，友人関係ネットワークを半教師付学習により求める手法を提案する．上述のインフルエンサ推定では友人関係情報を入力情報として用いるが，実際には完全な友人関係情報を得ることは容易でない．通話履歴やメール履歴等を観測することにより，友人同士であることを確認・推定できるが，観測できるのは全体の一部であり，すべての友人関係を把握することは網羅性の観点から困難である．このことから，半教師付学習により一部の判明している友人関係を利用して全体を推定することを試みる．友人同士においては興味の一部もしくは全部を共有していることを仮説としてモデルを構築する．教師付学習によるリンク推定の手法に対して，利用履歴（アプリケーションの利用履歴とインターネットアクセス履歴を結合したもの）を行列分解することにより得られる潜在特徴を組み合わせ，リンク推定と潜在特徴を同時に最適化することで，半教師付学習を実現する．さらには利用者同士が友人関係である可能性を示す指標としてシグモイド関数を用いた極化潜在利用者特徴を提案する．モニター実験により得たデータを用いて，モデルの妥当性を確認する．既存・最新のリンク推定手法と比較して，より良い友人推定の性能が得られることを実験的に確認する．

最後に，結言として上記の研究を総括し、課題と今後の展望について述べる．なお，本研究の基礎データとして用いるスマートフォン利用履歴を収集するために著者らが実施したモニター実験の概要について，付録に付記する．

目次

図目次

表目次

第 1 章

序 論

1.1 背 景

Blog, Twitter®[67], Facebook®[14], LinkedIn®[41], mixi®[50] 等のソーシャルメディアの急速な普及から，膨大な量の情報がネットワーク上に集積され始めた．さらには個人がメディアを発信することができるようになり，社会を動かすほどの力を持ち始めた．そして，このソーシャルメディアの大きな特性として，大規模処理が可能な電子データであることがあげられる．

また，近年における急速なスマートフォンの普及により，日常生活においてモバイル・デバイスを活用することが進展しつつある．移動通信の技術の進歩と普及および低廉化により，スマートフォンを含む様々な電子デバイスがネットワークに接続され，着実に様々な情報が電子データとして取り扱われるようになってきている．これらを統合することにより，人の活動を総合的に理解しサポートすることへの期待が高まりつつある．旧来から提唱されているコンテキスト・アウェア [4] なサービスやインタフェースの実現である．しかしながら、これらが実際に実用に耐える能力を発揮し普及するためには，これまでの取組以上に深く人を理解することが求められよう．

さらには，上述した高機能スマートフォンの普及と移動通信の普及・低廉化に加え，クラウドサービスが容易に利用できる形で提供されるようになり，多様な情報が電子化され記録・保存されるようになってきた．ストレージコストの劇的な低減も伴い，容易に個人が大量な電子データを所持できるようになりつつある．このため，日常の多様な出来事を様々なセンサにより計測・把握し記録・保存を行い，分析・可視化等により有効に活用しようというライフログも実現・実用可能となってきた．データを集積することが十分実現可能となりつつある現在，そのデータをどのように分析・加工し，利用者に対する価値を創り出すかが極めて重要である．2011 年の東日本大震災においても，災害時対応や復旧・復興において，ソーシャルメディア等の有用性が示されたものの，情報システム全般としての貢献が十分であったか，あるいは

具体的にどのような貢献ができればさらに良かったか，そのためには何が必要であるかを，考えさせられる大きなきっかけとなった．

この様に，技術の発展に伴うサイバー空間の拡がりとともに，それを現実の世界に適切に結び付け，より良い社会・世界を構築しようという CPS（Cyber Physical System）[15] の考え方が注目を集めつつある．しかし，社会・世界をより良くするためには，社会・世界を，そしてそれらを構成する人について，より良く知る必要がある。これまでは，社会学・心理学等の分野で扱われてきた課題に対して，ビッグデータの分析・処理という新しいパラダイムでのアプローチが，まさに始まろうとしている．

1.2 本研究の課題

前節で述べた通り，ビッグデータを活用して，よりよい社会を作りだすことへの貢献が求められている．これまでの情報システム関連の研究では、より良い社会を作り出すことへの視点が必ずしも十分ではなかった．これは，速度や性能などの向上が必ずしもより良い社会の実現に直接結びつかず，よりよい社会を実現するために技術がなすべきことが必ずしも明らかでないことに起因している．

社会は人により構成されており，人と人をつなぐのは多様なコミュニケーションである．ここでいうコミュニケーションとは，直接・間接・暗黙等の多様なもの含んだ広義なコミュニケーションを意図している．すなわち，人は様々な形で影響を及ぼしあいながら社会を構成している．そこで本研究では，ビッグデータを活用して，このコミュニケーションの構造を明らかにする技術を確立することを課題として取り組むことにする．特に本研究においては，人と人の間での影響（インフルエンス）の与え方を分析して，各種の予測やマーケティング等に応用できる価値のある情報を推定・抽出することを目指す．

通常このような分析を行うためには，心理学実験の手法である行動観察を行い分析する手法 [80] が用いられてきた．しかし，人手による行動観察を大規模に行うことは，その負荷の大きさから明らかに現実的でない．最近のネット利用の普及を活用して，大規模なサンプリング調査を行う研究（例えば [78]）が報告されているが，基本は回答を依頼する調査であるので，継続的に行うことは現実問題として考えられない．そこで，通常の活動に伴い受動的に収集可能である履歴等のデータを大規模に集積し，これを活用して分析・推定する手法の確立が望まれる．このとき注意するべきことは，現実的に（少なくとも近い将来には）入手可能と考えられる情報を用いて，目的を達成するために必要な技術を確立することである．

本研究においては，スマートフォンの利用履歴に着目し，コミュニケーションの構造を推定する技術の確立を課題とする．直接の通話履歴等を観察するだけでは，影響関係を得ることができないことや，直接的コミュニケーションの一部しか把握できないことから，コミュニケーション構造を推定するためには著しく不十分である．そこで，アプリケーションの利用履歴や

ネット閲覧履歴を活用することで，コミュニケーション構造を推定することを目指す．これらの利用履歴には，利用者の興味・嗜好が十分に反映されていると考えられることと，通常のスマートフォンの基本ソフトには履歴を取得する機能が備えられており実現性の観点からも現実的であることから，活用するデータとして好適である．

新製品などの新しい事物が人々によって採用され普及していく過程については，Katz らの研究 [30] に端を発し，Rogers のイノベータ理論 [60]，Bass の普及モデル [5] 等を基礎として，経済学や社会学において多くの研究がなされてきた．多くの研究は，マクロな現象としてとらえるものであったが，近年になってより詳細に個々の伝播をモデル化して捕捉しようとする研究がなされてきた [56]．マーケティング施策への応用や，個々の利用者に対する各種の利用予測・推薦へ適用するためには，個々の伝播のモデル化が必要である．個々の伝播のモデルとしては，独立カスケードモデル（IC モデル），線形閾値モデル（LT モデル）が基本モデルとして用いられることが多い [33, 21].

しかしながら従来の研究を適用してスマートフォンの利用履歴からコミュニケーションの構造を推定しようとすると，次のような課題がありそのまま適用するだけでは十分な結果が得られない．(1) 従来の研究は伝播が起こりうる経路としてソーシャルネットワークが与えられることを前提としているが，実際にはこれを得ることができない．(2) 伝播の発生頻度は多くないため観測データがスパースとなり精度を得るために工夫が必要．(3) 自然発生的に伝播したものと，クチコミ等の直接的コミュニケーションにより生じた伝播とが混在して観察されるが，これらを区別する手法が確立されていない．

本研究では上記の課題に対して，次の 3 種類のコミュニケーション構造の推定に取り組む．これらの取組みの全体像を図 1.1 に示す．

1. 暗黙の影響関係構造
2. 直接的影響関係構造（クチコミの構造）
3. 友人関係構造

暗黙の影響関係構造とは，間接的な影響や外部要因等を含めて，利用履歴全体を影響構造により説明するモデルである．利用予測を行う場合は，この影響構造を用いることができる．直接的影響関係構造とは，暗黙の影響関係構造から特にクチコミ等による直接的コミュニケーションによって及ぼされる影響の関係を抽出したものである．マーケティング施策として影響力の強い顧客を抽出したい場合には，この構造が必要となる．ここで直接的影響の抽出のために友人関係情報を利用する．しかし友人関係についても現実的には容易に全体を得ることができないため，観測から知ることができる一部の友人関係情報を用いて利用履歴から友人関係構造を推定する手法も合わせて検討する．

上記の各構造は，いずれも観察できるデータから直接得ることができない．そこで，観察できるデータの背後には，これらの構造があるとの立場から，潜在構造モデリングの枠組みによ

り構造を推定する．すなわち，仮説として構造の確率モデルを設定し，観察されたデータを用いて統計的機械学習によりモデルのパラメータを推定し，交差検定によりそのモデル仮説の妥当性を評価検証する，という手順を踏むことで構造を推定する．

1.3　本論文の概要と構成

本論文は，全 5 章で構成される．

以下，第 2 章では周囲への暗黙の影響をモデル化することによりアプリケーション利用予測を行う手法について述べる．提案手法の予測能力を既存の代表的予測手法と比較することにより評価し，高い予測能力が得られることを示すとともに，暗黙の影響関係には潜在グループ構造が存在することが実験的に見出されることを示す．

第 3 章では，さらに詳細に周囲へ及ぼす影響について踏み込み，直接的にどれ位周囲に影響を与えているのかを，友人関係の情報を合わせて用いることにより明らかにする手法について述べる．これにより，直接的な影響を周囲に与える人を，利用履歴を分析することから見つけ出すことができるようになる．直接的に影響を周囲に与える人はインフルエンサと呼ばれ，プロモーション，マーケティング，トレンド予測等において価値が高いので，その抽出は実用的に大変有用であるが，実際に有効なインフルエンサを把握することが困難であった．

第 4 章では，利用履歴を分析することから友人関係の有無を推定する手法について述べる．これにより第 3 章で入力情報として利用する友人関係情報を，部分的な観測から推定により得ることができる．リンク推定の手法と利用履歴から得られる利用者潜在特徴を統合することにより，既存のリンク推定手法よりも高い推定能力が得られることを示す．

第 5 章では，結論を述べるとともに、課題と今後の展望を示す．

なお，第 2 章では文献 [28, 29, 74, 75] により公表した結果に基づいて論述し，第 3 章では文献 [76] にて公表した結果に基づき論述する．第 4 章は文献 [26] により公表した結果に基づき論述する．

また，第 2 章 ~ 第 4 章の研究を行う上で基礎データとして用いたスマートフォン利用履歴を収集するために著者らが実施したモニター実験の概要を付録に示す．

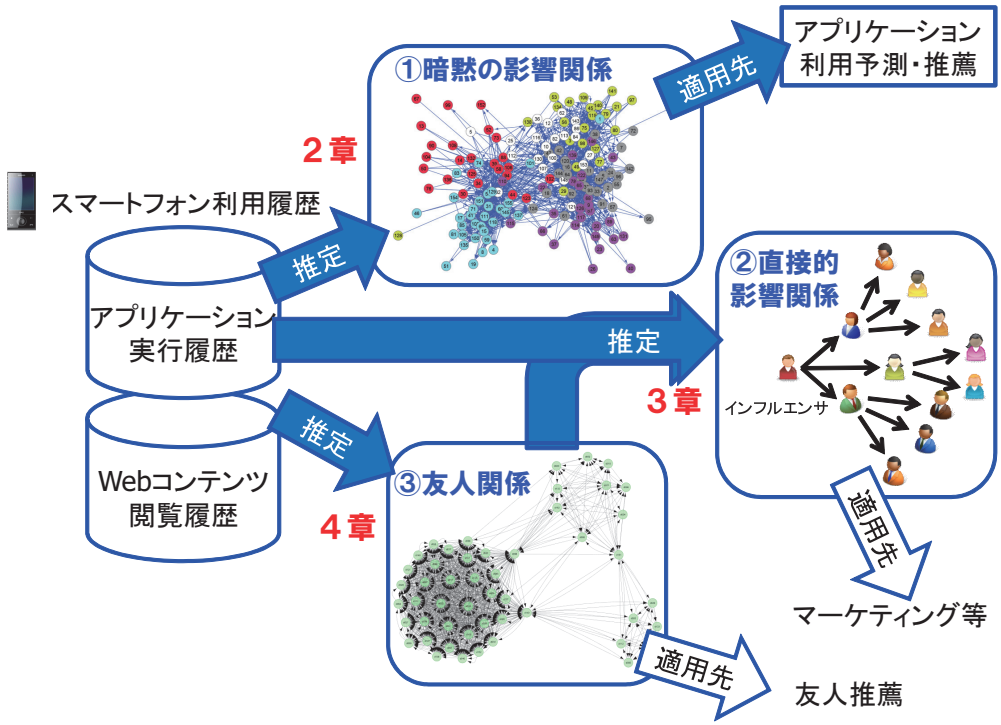

①暗黙の影響関係
適用先
アプリケーション
利用予測・推薦
2章
スマートフォン利用履歴
推定
アプリケーション
実行履歴
推定
②直接的
影響関係
インフルエンサ
3章
推定
Webコンテンツ
閲覧履歴
③友人関係
4章
適用先
マーケティング等
適用先
友人推薦

図 1.1　取組み全体像

第 2 章

暗黙の影響関係の構造推定とアプリケーション利用予測

2.1 はじめに

最近の研究結果から，個人間の影響関係を考えることが，各人の行動パターンを理解することや商品等の推薦をするうえで，重要かつ効果的であることが示されている [31]．今，ユーザ B が購入するアイテムには，ユーザ A がすでに購入しているものが多い傾向があるとしよう．この場合，暗黙の影響関係 $A \rightarrow B$ があると考える．暗黙の影響関係には，周囲の友人等に対してクチコミ等により直接的な影響を与える場合の他に，Liu らが文献 [42] にて議論している間接的な影響関係（直接的に関係のない人に対して介在者を経由して与える影響）や，ユーザ A が単に新しいものが好きで早い段階で新しい物を購入する傾向があるような場合これによる効果も含まれる．本章では，購買の履歴を観察することによって，個人間における暗黙の影響関係を抽出する手法について述べる．[28, 29, 74, 75]

個人間の影響関係を扱う関連研究として，ネットワーク上での情報の拡散を模擬するモデルが提案されている．広く用いられているモデルとして，独立カスケードモデル（IC モデル）と線形閾値モデル（LT モデル）があり，双方とも情報拡散を扱う基本的な確率モデルである [33, 21]．IC モデルは，ソーシャルネットワークが既知であることと，各リンクには拡散確率が割り当てられていることを仮定する．LT モデルも同様に，ソーシャルネットワークの構造が既知であることと，各リンクには重み値が割り当てられていることを仮定している．LT モデルは，当該アイテムを購買済みの周囲のノードからの重み値の総和が確率的閾値を超えた場合，購買の伝搬（拡散）が起こることをモデル化している．これらのモデルに基づいて，拡散確率を最適化問題として求める手法がいくつか提案されている [25, 35, 62, 24]．一方，IC/LT モデル以外にも，Song らは購買履歴から情報フローネットワークを構築し，これを用いた将来の購買予測を提案している [64]．また川前らは，購買に関する周囲への影響は経過時間に応

じて指数的に減少するという仮説に基づき personal innovator degree を提案した [31]．上記の関連研究はすべて，個人間の影響度合いを個々に（独立に）推定する必要がある．拡散（伝搬）の起こる経路が与えられたソーシャルネットワークに限定される場合は，推定すべき影響度合いの数はそのリンク数となる．しかしながら暗黙の影響関係を考える場合は拡散（伝搬）の起こる経路を事前に限定することができないため，すべての個人間の組み合わせについて影響度合いを推定する必要が生じる．このため，規模が大きくなると計算量が大きくなり推定が困難となる．さらには，個人間の影響度合いは双方がともに購買した共通アイテムの数から推定することになるため，すべての個人間において一定量以上の共通アイテムが存在することが望まれるが，実際にはこの共通アイテムの数は少ないことが多く，安定して拡散確率をを推定することが困難である．

このような状況に対応するためには，適切なモデル仮説を導入し利用可能な購買データからの推定精度と効率の向上を図る手法が必要である．例えば購買の予測に用いられる協調フィルタリング（Collaborative Filtering，以下 CF）では，個人の購買パターンと個人の嗜好性の相関を仮定し，購買履歴が類似している利用者の購買履歴もしくは過去に購買した商品と購入者が類似している商品を手がかりに，将来の購買を予測する [66]．CF の実現手法として行列分解を利用した手法が提案されており良好な性能を示すことが報告されている [37]．この場合，利用者–アイテム行列を対象に行列分解を行い，結果として潜在利用者特徴と潜在アイテム特徴を得ていることになる．著者は，影響を与える関係のモデルとしても同様なアプローチが有望であると考え，行列分解手法を用いた潜在特徴モデルの適用を提案する．すなわち本検討においては利用者間の影響度合いを表す行列を行列分解し，潜在影響特徴と潜在被影響パターンを得ることとなる．また，推定する影響度合いは情報拡散（伝搬）の発生確率を表すことから，影響度合いは非負の値を持つように拘束すべきである．このため，行列分解の手法には，NMF（non-negative matrix factorization）[38] を用いる．

NMF は，解釈性の良い結果を得られるデータ解析ツールとして，近年広く用いられるようになってきた．当初は，全体を意味のある部分に分解することができる行列分解手法として，主に映像解析向けとして提案された [38]．その後，NMF は音楽解析，テキストマイニング，遺伝子解析などの様々な分野に応用されその有用性が示されており，それらの詳細は Wixiang のサーベイ [71] に示されている．さらに最近では，NMF を用いたコミュニティ発見の手法も提案された [69]．また，NMF の求解手法には，より良く特徴抽出を行うための改良提案が複数なされており，代表的な手法を表 2.1 に示す．

上述の影響を与える関係のモデル化について，学生約 160 人の実データを用いてその有効性を実験的に確認する．代表的な NMF を用い潜在特徴の次元数および必要な制御パラメータを変化させ，クロス・バリデーションの要領にてそのモデルの妥当性（推定性能）を観察し結果を評価分析する．この結果から次のような重要な発見が得られた;

- 提案モデルによると低い潜在特徴次元数において推定性能が極大となり，かつ CF より優れた推定性能が確認されたことから，影響を及ぼす関係には潜在グループ構造が存在することが示唆される
- 行列分解においてはスパースネスを制御することが重要である — 適度にスパースな特徴ベクトルがより良い推定性能を示す

ここで行列分解においてスパースネスを制御するとは，行列分解により得られる 2 つの因子行列を構成する特徴ベクトルについて，その要素の値がほぼ 0 である頻度を大きくし，少数の要素のみが大きい値を持つように，制約をかけて求解することを指す．

図 2.1 に，実際のスマートフォンの利用データから獲得した影響構造の例を示す．各ノードは利用者を表し、ノードの色は所属する潜在影響グループ（G1~G6）を表している．各エッジはその力が閾値を超える主要な影響関係の存在を表している．ここで、潜在影響グループは影響を及ぼすパターンにおける潜在特徴から得たものであり，コミュニティ発見のようにエッジの密度や接続性からグループを得たものではない．グループ内の影響関係がグループ間の影響関係より強い傾向にあるため，図がコミュニティ発見の結果と類似したものとなっているが，その意味付けは異なっていることに注意する．図 2.1 についてのさらなる詳細は、2.3.3 節にて述べる．

ここではスマートフォンの利用履歴を，暗黙の影響関係を分析するターゲットとして選択した．これは，スマートフォンの急速な普及から，スマートフォンにおける利用者の振る舞いの理解が今後重要となると考えたためである．しかしながら、ここで記述する手法はスマートフォンの利用履歴に限定されず，他の利用履歴や購買履歴等に適用することができる．

著者の知る限りにおいて，潜在特徴モデルを利用者の影響関係に適用し，実データによりその妥当性を確認した先行研究はない．本研究の主要な貢献は下記の通りである．

- 個人間の影響関係に対し潜在特徴モデルを適用し，より良い予測性能が得られるモデルを提案
- 約 160 名の学生による 4 か月間の実際のスマートフォン利用データにより実験的に効果を検証；潜在グループ構造の存在が明らかになった，スパースネス制御が重要であることが明らかとなった，さらには最新の手法と比較しても提案手法の予測性能が上回ることを示した．

本章の以降の構成は次の通りである．2.2 節にて問題設定とモデルの定式化を説明する．2.3 節では，収集したデータとモデルに基づいた分析について述べる．この分析には潜在グループの発見と予測性能の評価が含まれる．2.4 節に考察を示す．2.5 節にて本章をまとめる．

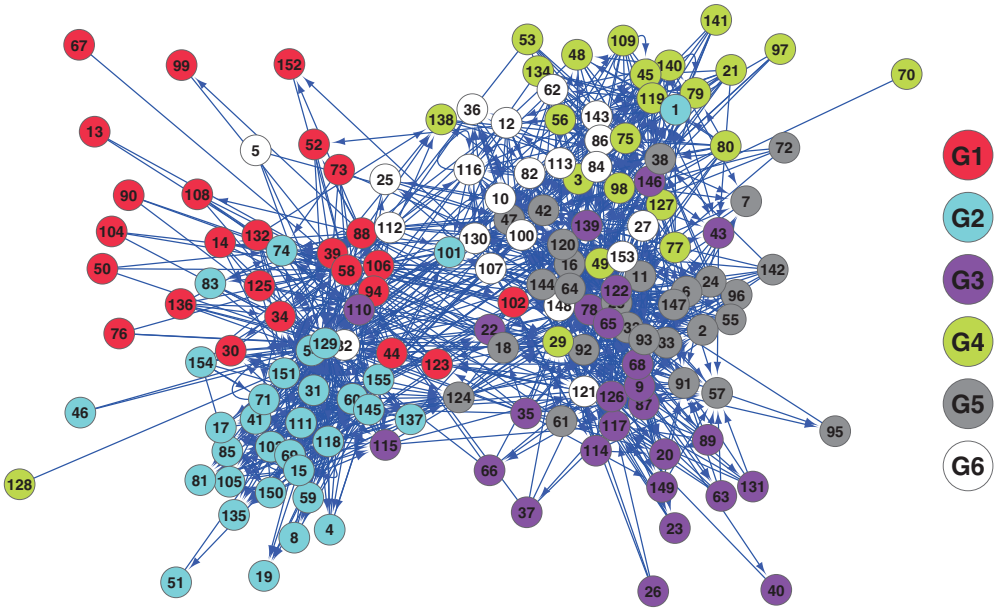


図 2.1　影響関係と潜在影響グループ

2.2 個人間の影響

2.2.1 非対称な関係

図 2.2 は簡単な，アプリケーションの利用履歴の例を示したものである．例では，3 人の利用者（user 1, 2, 3）が 5 種類のアプリケーションを利用している．ここで，水平軸は時間の推移を示し，円で囲まれた数字は，利用者が該当する時刻に当該アプリケーションを実行したことを示す．この図から明らかなように，利用者–アプリケーション関係から、アプリケーション実行の連鎖（同じアプリケーションを実行した利用者の列；青い実線矢印）が得られる．ここで，未来において各利用者がどのアプリケーションを実行するかを予測し，推薦に応用することを考える．典型的な協調フィルタリング（CF）手法 [66] を用いると，利用者 1 が次に実行するアプリケーションを予測する際には，利用者 2 と利用者 3 の履歴が等しく用いられることになる．これは利用者 1 に対する履歴の類似度が，CF では順序関係を用いないため，利用者 2 と利用者 3 とで等しくなるためである．したがってアプリケーション 4 と 5 は CF では等しく推薦されることになる．しかしながら，実行の順序関係に着目すれば，利用者 2 は利用者 1 の先行者であり，利用者 3 は利用者 1 の追随者である傾向が明らかである．このため利用者 1 はアプリケーション 4 よりもアプリケーション 5 を実行する確率が高いと考えられる．この例によって，将来予測をする上で，個人間の影響関係を加味することの重要性は明らかである．

2.2.2 提案手法の全体像

本検討で述べる，影響関係における潜在グループ発見プロセスの全体を図 2.3 に示す．学習セットに基づき影響行列 $\mathbf{R}$ を算出する．その後，$\mathbf{R}$ に対して行列分解（NMF）を行い近似影響行列 $\widehat{\mathbf{R}}$ を得る．この時行列分解に用いる潜在特徴の次元数が得られる近似影響行列 $\widehat{\mathbf{R}}$ のランク数となる．得られた近似影響行列 $\widehat{\mathbf{R}}$（または影響行列 $\mathbf{R}$）と学習セットを用いて，その後のアプリケーション実行を予測する．予測性能は，予測結果とテストセットを用いて交差検定により算出する．各部の処理については，以降の節にてより詳細に説明する．

2.2.3 影響係数と影響行列 R

本節では，影響行列 $\mathbf{R}$ およびこれを構成する影響係数を定義する．ここで用いる影響係数は，Goyal らが文献 [24] にて提案している静的モデルに類似している．ソーシャルネットワークがあらかじめ与えられることを仮定しないところが主要な差異である．このため，観測されるアプリケーション利用履歴を用いて，すべての利用者に対する影響係数を推定算出する必要

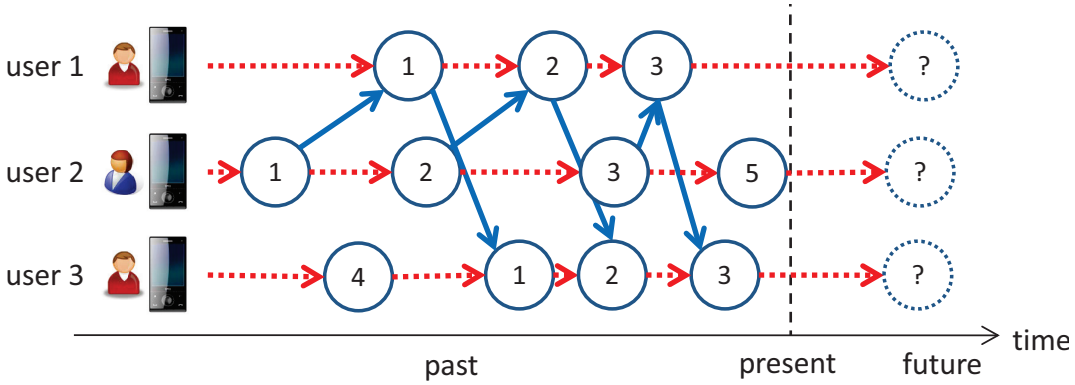


図 2.2　アプリケーション利用の連鎖

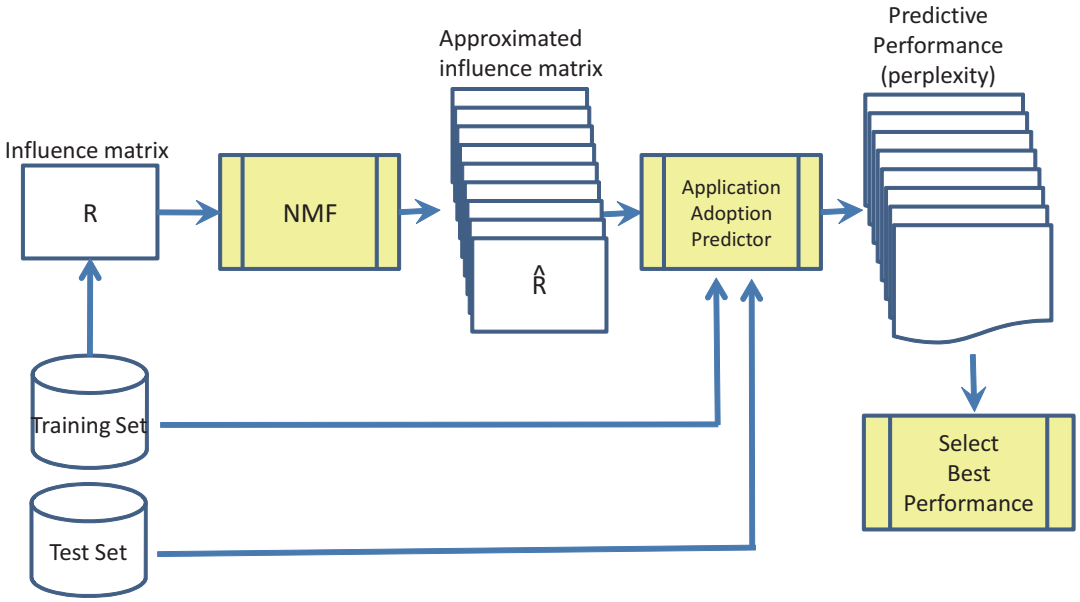


図 2.3　潜在グループモデル獲得の処理構成

がある．

頻度に基づく影響モデリング

利用者 u は利用者 v に対して，静的な影響確率 $P_r(u \to v)$ を持つこととして、これを影響係数と呼ぶ．利用者 u がアプリケーション d を実行した後，利用者 u は周囲のアプリケーション d を未だ実行していない利用者 v に対して，一定期間 τ の間はいつでも影響を与えようとすると考える．ここで，係数 τ は定数であると仮定する．影響を与える試行はベルヌーイ試行と見なす．これにより試行の成功確率の最尤推定結果は，総試行回数に対する成功回数の比で表される．すなわち，u が v に対して影響を及ぼす確率は次式により推定される．

$$P_r(u \to v) = \frac{\|A_{u \to v}\|}{\|A_u\|} \tag{2.1}$$

ここで $\|A_u\|$ は, 利用者 u が学習セットにおいて実行した総アプリケーション数である．$\|A_{u \to v}\|$ は学習セットにおいて，利用者 u が実行しその後利用者 v が実行したアプリケーションの数である．

$P_{inf}(d|u)$ が，アプリケーション d が利用者 u によって周囲からの影響により実行される連結確率を表すことにする．本検討においては，利用者 u に影響を与える周囲の様々な人々はお互いに独立であることを仮定する．これにより連結確率 $P_{inf}(d|u)$ は次式により定義することができる．

$$P_{inf}(d|u) = \left[1 - \prod_{u'} \{1 - P_r(u' \to u) Y(u', d)\} \right] \tag{2.2}$$

$$Y(u', d) = \begin{cases} 1 & \text{if 利用者 } u' \text{ がアプリケーション } d \text{ を過去 } \tau \text{ 以内に実行した} \\ 0 & \text{それ以外} \end{cases} \tag{2.3}$$

エントロピーの導入による影響モデリングの改善

スマートフォンのアプリケーションには，非常に普及しているものから，ごく一部の利用者にしか使われないニッチなものまで、多様なものが存在する．非常に普及し誰でも使うようなアプリケーションよりも，後者の一部の利用者にしか使われないものの方が，個人間の伝搬をより特徴的に捉えやすいと考えられる．このためこの考え方を取り込むため，エントロピーを用いた影響係数を提案する．エントロピーは杞憂さを表す尺度と考えられる．提案するモデリング手法では，アプリケーションのユニークユーザ数を用いてアプリケーションのエントロピーを求める．すなわち，エントロピーに基づく影響確率 $\widehat{P_r}(u \to v)$ を次式により求める．

$$\widehat{P_r}(u \to v) = \frac{\sum_{a \in A_{u \to v}} (-log\left(\frac{u_a}{U_{glob}}\right))}{\sum_{a \in A_u} (-log\left(\frac{u_a}{U_{glob}}\right))} \tag{2.4}$$

ここで，U_{glob} は実験中に使われた Android アプリケーションの世界中での総ユーザ数であり，u_a はアプリケーション a の（世界中での）ユニークユーザ数である．u_a と U_{glob} の真の値を得ることは困難であるため，u_a には “Android Market” にて記録・公表されている概略ダウンロード数を用いる．また U_{glob} には，仮の値として 2,000,000 を実験では用いることにする．$P_{inf}(d|u)$ を $\widehat{P_r}(u' \to u)$ に基づいて計算することは，(2.2) 式に対して $P_r(u' \to u)$ を $\widehat{P_r}(u' \to u)$ に入れ替えることにより実現できる．

頻度に基づくモデリングとエントロピーを導入したモデリングについて事前評価実験を行った．この結果によれば，エントロピーを導入したモデリングがわずかに良好な結果となった．このため，以降においてはエントロピーを導入したモデリングを用いることとする．係数 τ の影響についても事前評価実験を行った．結果は τ が長ければ長いほどより良いモデリング性能を表した．このため，以降のすべての実験は，$\tau = \infty$ の設定にて行うこととする．

ここで, 影響行列 $\mathbf{R}$ を, その u 行 v 列の要素に影響係数 $P_r(u \to v)$ を値として持つ $U_{exp} \times U_{exp}$ の行列，と定義する．ただし U_{exp} は，実験協力者の総数を表している．

2.2.4　影響関係における潜在特徴

NMF は，非負行列 $\mathbf{X}$（$n \times m$ 行列）を 2 つの非負行列 $\mathbf{W}$（$n \times k$ 行列）と $\mathbf{H}$（$k \times m$ 行列）に，$\mathbf{X} \simeq \mathbf{WH}$ となるように分解する．この時通常 k は n，m と比較して小さい値（$k \ll n, m$）とする．ここで，$\mathbf{WH}$（$n \times m$ 行列）を $\mathbf{X}$ のランク k による低ランク近似行列と呼ぶことにし，$\widehat{\mathbf{X}}^{(k)}$ と表記する．NMF には行列分解において行列要素の値を非負に保つ特徴がある．

利用者間影響行列 $\mathbf{R}$ を NMF により行列分解して，ランク k の近似行列 $\widehat{\mathbf{R}}^{(k)}$ を得る．$\mathbf{R}$ と $\widehat{\mathbf{R}}^{(k)}$ は同じサイズの行列であるので，利用者のアプリケーション利用を影響行列 $\mathbf{R}$ から求める手法をそのまま $\widehat{\mathbf{R}}^{(k)}$ に適用することができる．$\mathbf{R}$ もしくは $\widehat{\mathbf{R}}^{(k)}$ を行列分解した結果，得られる行列 $\mathbf{W}$ と $\mathbf{H}$ は因子行列となる．この因子行列の内容について直観的な解釈を与えると，図 2.4 に示す通り，$\mathbf{W}$ はサイズが 利用者数 $\times k$ の影響の与え方に関する潜在特徴行列となり，$\mathbf{H}$ はサイズが $k \times$ 利用者数 の各潜在特徴の内容（被影響パターン）を表す基底行列となる．

様々な NMF のバリエーションおよび動作パラメータに対して得られる最適動作性能（operational optimal boundary）を得るために，標準的に提案されている複数の NMF アルゴリズムに対し各々パラメータ探索を行い予測性能を計測した．NMF は非線形の最適化処理を行うため，正則化項の種類および動作パラメータにより異なる結果となる．表 2.1 に実験に用いた NMF アルゴリズムのリストを示す．すべての NMF 計算には R[58] 環境におけるパッケージ NMF[18] を用いた．各 NMF アルゴリズムについて対応する動作パラメータを変化させ得

られた近似影響行列 $\widehat{\mathbf{R}}^{(k)}$ について各々予測性能を得た．ランク k ごとに最も良い予測性能値を採用することとして，ランク k に対する性能の変化を観察した．この性能変化から，交差検定の要領により，小さい k に予測性能の極大値が確認できれば，個人間の影響関係における潜在特徴次元と推定することができる．

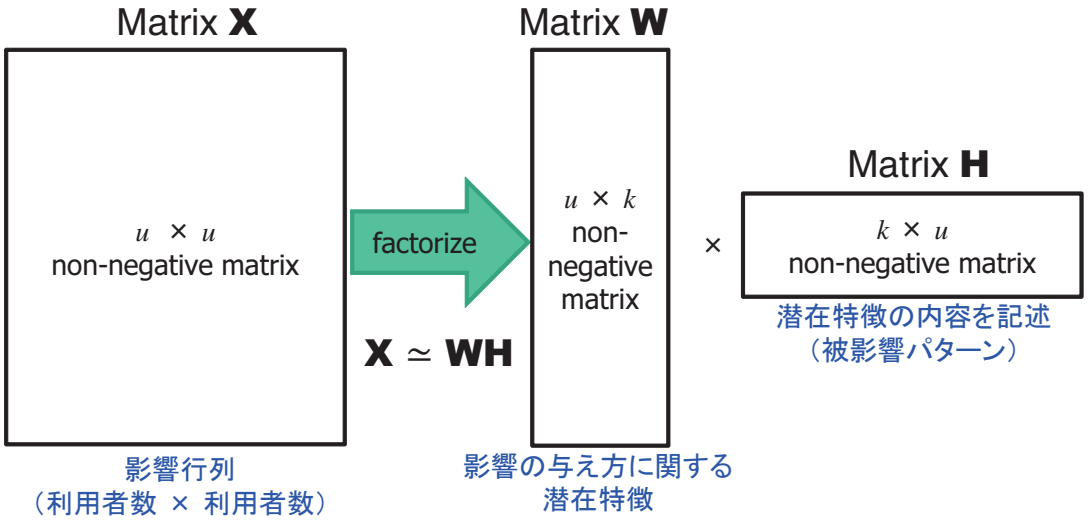


図 2.4　影響行列の行列分解

表 2.1　NMF アルゴリズム

#	文献	説明
1	[10]	Standard NMF, based on Kullbach-Leibler divergence
2	[39]	Standard NMF, based on Euclidean distance
3	[55]	Non-smooth NMF based on Kullbach-Leibler divergence
4	[3]	Modified version of [39] based on Euclidean distance
5	[73]	Pattern-Expression NMF based on Euclidean distance
6	[34]	Alternating Least Square (ALS) approach

2.3 実験と結果の分析

2.3.1 利用するデータ

大阪大学においてスマートフォンの利用モニター実験を行いデータを収集した．行ったモニター実験の詳細について付録 A に示す．157 人の学生が実験に協力者として参加した．実験協力者を募集する際に，3 人以上 50 人以下の友人グループで応募することを条件とした．結果的に 6 グループから構成される 157 人が参加した．実験用の動作監視ソフトウェアをスマートフォン Xperia® に搭載し，協力者に貸与した．協力者には実験の目的および収集データの研究目的の利用について説明し許諾を得た．

動作監視ソフトウェアは，アプリケーションの実行を監視し履歴を記録する．履歴情報は匿名化処理を行った後に 3G ネットワーク経由でサーバに収集される．各履歴情報は、時刻情報，匿名化ユーザ ID，実行されたアプリケーションのパッケージ名から構成される．

収集された履歴情報から，利用者毎に各アプリケーションの初回実行時の履歴を抽出する．図 2.5 に抽出した初回実行の履歴の累積件数を時間軸に対して示す．ここでは，時間軸により履歴情報を，学習セットとテストセットに分割した．実験開始時の急峻な立ち上がり期を避け，学習セットには 2011 年 2 月 ~ 4 月の 89 日分を用いることとした．テストセットには学習セットの直後の 5 月の 31 日分を用いる．さらに，学習セットにおいて利用者数が 3 未満のアプリケーションは除外した．以上の処理により，学習セットは 3,383 履歴（利用者数：155，アプリケーション数：291）となり，テストセットは 249 履歴（利用者数：99，アプリケーション数：166）となった．なお実行されたアプリケーションはすべて無料アプリケーションであった．

図 2.6 は，各利用者が学習セットにおいて実行したアプリケーション数について，降順に並べたものである．最大の実行アプリケーション数は 110 であった．表 2.2 に，協力者のデモグラフィック属性分布とともに，学習セットにおける実行アプリケーション数の平均値を示す．図 2.7 は，任意の利用者の組において双方がともに実行したアプリケーション数を算出し，降順に並べたものである．

2.3.2 評価方法

まず影響行列 $\mathbf{R}$ を，式 (2.4) を用いて学習セットから算出する．次にランク k を変化させ近似影響行列 $\widehat{\mathbf{R}}^{(k)}$ を，$\mathbf{R}$ に基づいて算出する．得られた $\mathbf{R}$ または $\widehat{\mathbf{R}}^{(k)}$ を用いて，式 (2.2) と利用履歴（学習セット）により，テストセット期間におけるアプリケーション実行を予測する．予測性能は，予測結果とテストセットのデータを比較することにより評価する．評価指標としては，(1) テストセット perplexity [17]（以降，perplexity），(2) mean average precision (MAP)

[53] を用いる．perplexity は，言語モデルの評価に標準的に用いられる尺度であり，確率領域における MAE/RMSE に相当する．perplexity は確率モデルが実際の観測に対して出現確率の観点でどれだけ符合しているかを表すため，本検討においては潜在グループ構造の確認のために用いる．一方，MAP は推薦や情報獲得の分野において共通に用いられる性能指標であるので，予測性能の尺度として用いる．予測性能は，現時点において最善の能力を有するとされる協調フィルタリング手法（biased-SVD[36]，SVD++[36]，wALS[53]，NMF-based CF）および伝統的なユーザベース協調フィルタリング [66]，人気度による推薦の各手法と比較評価する．評価結果は 2.3.5 節にて述べる．

2.3.3　潜在グループの発見

NMF による低ランク近似影響行列 $\widehat{\mathbf{R}}^{(k)}$ により得られた最適動作性能（operational optimal boundary）を図 2.8 に示す．この図から明らかにランク 5~6 において perplexity の減少（すなわち予測性能の向上）が見られる．このことから，個人間の影響関係には 5~6 次元程度の潜在特徴構造が存在することが想定され，各利用者の特性はこれらの特徴次元に対する重み係数により表現できる可能性が示唆された．

さらに図 2.9 は，図 2.8 と同じデータから固有値分解（SVD）[22] により低ランク近似影響行列を得た場合の予測性能を，NMF による性能に合わせて表示したものである．いずれのランク数においても，NMF を用いた場合が SVD による場合よりも良好な性能を示している．SVD の場合には，影響行列 $\mathbf{R}$ をそのまま用いて予測した場合の性能（98.2635）よりも，多くのランク数において予測能力が低くなることが観察される．NMF を用いた $\widehat{\mathbf{R}}^{(k)}$ は，SVD の結果からわかるように低ランク近似が予測能力を向上させる保証が無いにも関わらず，すべてのランク数において上回ることは注目に値する．非負制約の効果と推定される．加えて，図 2.9 からは SVD による低ランク近似影響行列 $\widehat{\mathbf{R}}^{(k)}$ の場合においても，NMF の場合と同様にランク数が低い場合に perplexity 値が低くなる現象が観察されている．これは個人間の影響関係に潜在特徴構造が存在することをあらためて確認する証拠と言えよう．

図 2.10 は，perplexity が最善を示したランク $k = 6$ における近似影響行列を構成する分解された因子行列を可視化したものである．ここで行列の各要素の値は利用者毎に正規化（利用者毎に総和が 1 となるように）してある．また値の大きさを色で表している．赤色は高い値を示し明るい黄色は小さい値を表している．図 2.10(a) が影響の与え方を表す係数行列 $\mathbf{W}$ であり，図 2.10(b) は影響パターンの基底行列 $\mathbf{H}$ である．2.2.4 節で述べた通り，係数行列 $\mathbf{W}$ の垂直軸および基底行列 $\mathbf{H}$ の水平軸は個々の利用者を表している．係数行列 $\mathbf{W}$ の各利用者に対するベクトルは，その利用者の影響の与え方をパターン毎の重みで表したもので，その利用者が誰に影響を及ぼすかを示す．一方，基底行列 $\mathbf{H}$ における基底ベクトル（長さが利用者数のベクトル）は，各パターンが具体的に誰にどれ位の割合で影響を与えるものかを示している．例え

ば，パターン 1 は基底行列 **H** における左端の 39 列が示す利用者に対して強い影響を与えることを表している．ここで，これらの行列には双方向性があることに注意する．すなわち，影響を与える立場で考えることと，影響を受ける立場で考えることができる．このため，図 2.10(b) を影響を受ける場合の係数行列，2.10(a) を基底行列と考えることもできる．

図 2.10(a) を見ると，各利用者には一つの支配的な影響パターンがあることがわかる．すなわち，支配的な影響パターンが共通である利用者同士は，影響の与え方が類似していることになる．ここで，支配的な影響パターンが共通である利用者集合を，影響グループと呼ぶことにする．同様に図 2.10(b) において，支配的な影響を受けるパターンが共通である利用者集合を，被影響グループと呼ぶことにする．影響グループ G1 に所属する利用者は，被影響グループ g1 の利用者に対して主として影響を及ぼす．また，すべての利用者は各々いずれかの影響グループおよび被影響グループに所属するため，被影響グループ g1 の利用者は，いずれかの影響グループに所属しており，これを把握することにより，影響グループ G1 が各影響グループに対してどの程度影響を与えるのかを知ることができる．上記の考えにより，影響グループが各影響グループに対して及ぼす影響の強さを人数比で集計した結果を，表 2.3 に示す．表 2.3 から得られる影響グループ間の影響の強さ関係を可視化すると，図 2.11 のような構造で表される．表 2.3 から，影響グループ内の影響が，影響グループ間の影響と比較して強い傾向があることが読み取れる．これは同一影響グループに所属する利用者間での相互影響が密であることを表す．さらに図 2.11 から，影響グループ間の関係を読み取ることができ，G3~G6 の関係が密であること，G1 や G2 は G3~G6 とは比較的に独立であること，を知ることができる．

図 2.1 は近似影響行列 $\widehat{\mathbf{R}}^{(6)}$ について，グラフ構造の可視化を行った結果を示したものである．利用者–利用者間の影響関係の強さが上位 6% の関係を $\widehat{\mathbf{R}}^{(6)}$ から抽出し，エッジが存在するものとして可視化を行った．影響関係の強さが上位 6% のものにより，$\widehat{\mathbf{R}}^{(6)}$ 全体の要素値の総和の 35% を占めていた．グラフ構造の可視化には広く用いられている force-directed layout アルゴリズムを用いた．図 2.1 には，上記にて議論した所属する影響グループを色分けして表示した．図 2.11 の影響グループ間の関係が，図 2.1 の描画においては影響グループによる操作を一切していないにも関わらず，図 2.1 における影響グループの分布状況と整合していることが観察できることからも，影響グループの構造関係が裏付けられる．

2.3.4 スパースネス制御の効果

図 2.8 の最適動作性能を示した NMF アルゴリズムを各動作点（ランク数）について表 2.4 に示す．この表から明らかであるように，“nsNMF” アルゴリズムが 20 以下のランクにおいて良好な性能を示している．この “nsNMF” は，分解した行列のスパースネスを制御しつつ，同時に積が元の行列を良く近似するように最適化する手法である．他の NMF アルゴリズムよりも安定して良い能力を示したことから，本検討の主題である影響関係における潜在グループ構

表 2.2　実験協力者の属性分布

属性		人数	平均実行 AP 数
所属	文学部	5	21.6
	人間科学部	14	26.7
	法学部	12	21.0
	経済学部	13	38.5
	理学部	8	24.9
	医学部	23	25.7
	薬学部	3	41.7
	工学部	24	37.6
	基礎工学部	33	30.3
	外国語学部	19	22.4
	情報科学研究科	1	13
入学年	2006	1	13
	2007	2	37.0
	2008	50	26.2
	2009	28	34.5
	2010	74	28.9
性別	女性	46	24.4
	男性	109	31.0
Total		155	29.1

表 2.3　グループ内とグループ間のインフルエンス

人数 (比率 (%))		影響を受ける側の所属する影響グループ					
		G1	G2	G3	G4	G5	G6
影響を与える側の所属する影響グループ	G1	20 (51.3%)	5 (12.8%)	3 (7.7%)	6 (15.4%)	2 (5.1%)	3 (7.7%)
	G2	1 (3.2%)	24 (77.4%)	1 (3.2%)	2 (6.5%)	2 (6.5%)	1 (3.2%)
	G3	2 (6.9%)	0 (0.0%)	11 (37.9%)	4 (13.8%)	10 (34.5%)	2 (6.9%)
	G4	0 (0.0%)	2 (10.0%)	3 (15.0%)	7 (35.0%)	2 (10.0%)	6 (30.0%)
	G5	1 (5.9%)	0 (0.0%)	5 (29.4%)	4 (23.5%)	6 (35.3%)	1 (5.8%)
	G6	1 (5.3%)	1 (5.3%)	3 (15.8%)	0 (0.0%)	6 (31.6%)	8 (42.1%)

造を発見するうえで，スパースネス制御は重要な要素であることが示唆された．

“nsNMF” には制御パラメータ θ（$0 \leqq \theta \leqq 1$）がありスパースネスへの制約の強さを制御する．$\theta = 1$ とすることにより最も強くスパースとなる制約が課される．$\theta = 0$ では制約は効力を持たず通常の NMF と同様の動作をする．図 2.12 に，異なる θ を用いた際の予測性能の変化を示す．$\theta = 0$ の場合，特性曲線は緩やかであり，θ の値を増していくと急峻になっていくことがわかる．また θ が 0.6 を超えると，低ランクにおける性能向上（perplexity の減少）が見られなくなった．分解した行列をスパースにするということは，影響を与える人および影響を受ける人の範囲を狭めることに相当する．このことから，共振回路と同様の動作となっていることが想定される．すなわち，θ が共振回路の Q 値に相当し，Q 値（θ）が上限に達するまでは，Q 値の上昇とともに共振回路特性は鋭敏となる．ランク数は連続値でなく離散値を取るため，共振回路特性が鋭敏となり過ぎるともはや共振が観察できなくなることが考えられ，これが Q 値（θ）の上限となると考えられる．

2.3.5 アプリケーション利用予測

図 2.13 に，実験データから得られた提案する潜在特徴モデルの予測性能を示す．図中には比較対象として，広く用いられている推薦技術を用いた場合の予測性能も示した．図 2.13 において，“Latent Str. (NMF)” は前節にて議論した潜在特徴モデルを示す．“Influence” は低ランク近似をしていない元の **R** を用いて予測した場合を示す．“CF (User_COS)” はコサイン類似度を類似尺度として用いた伝統的なユーザーベース協調フィルタリングを示す．“Popularity” は，学習セットにおけるユニークユーザ数を用いた単純な人気度による予測を示す．図 2.8 と同様に，潜在特徴モデルについて $rank = 6$ 周辺で性能が向上していることが確認できる．

図 2.14 に，様々な CF 手法との性能比較結果を示す．“CF (NMF)” は行列分解に NMF を用いた潜在特徴モデル協調フィルタリングである．“CF (SVD_bias)” および “CF (SVD++)” は SVD を用いた潜在特徴モデル協調フィルタリング [36] であり，映画レーティングのタスクにおいて高い予測性能を示すことで知られている．“CF (wALS_item)”，“CF (wALS_user)”，“CF (wALS_universal)” は one-class 設定のタスクへの適用を特に意識して提案された協調フィルタリング [53] である．ここで注意が必要であるのは，本検討にて取り扱っているタスクは one-class 設定のタスクであるということである．one-class 設定とは，学習の時点において負例が与えられず，正例のみしか与えられない状況で，未知の例に対して正もしくは負を予測するようなタスクである．学習セットにおいて，あるアプリケーションを実行しなかったからといって，その利用者がそのアプリケーションが嫌いなので実行しなかった（今後も実行しない）のか，あるいは気付かなかったので実行しなかったのか（今後実行する可能性あり）のか，区別をつけることができない．このため学習においては正例のみしか与えられないことになる．“CF (NMF)”，“CF (SVD_bias)”，“CF (SVD++)” については，AMAN (all missing as

negative) 方式 [53] により性能を評価した．なお，各種 CF 手法の性能評価には GraphLab[43] を用いた．

図より提案する潜在特徴モデルが，最も良好な性能を示した．また one-class 設定向けの “wALS_item” が既存手法の中では良好な性能を示したが，ランク数 5 以上においては僅かに提案手法が上回った．個人向け推薦への実応用を想定する場合には，既存手法との性能差が僅かであっても，それに応じた効果（増収・利便性向上等）が継続的に期待できる．この結果より，提案モデルの妥当性が確認されるとともに，個人向け推薦として有用であることが示された．

2.4　考察

2.4.1　潜在グループの特徴

潜在グループ（影響グループと被影響グループ）は，影響を与える／受けるパターンの類似性により構成される．このことから，同一グループに所属する利用者は類似したアプリケーション嗜好性を持つことや，類似したデモグラフィック属性を持っている可能性が考えられる．これらを確認するために，影響グループに対する，デモグラフィック属性や利用アプリケーションの相関の有無を分析した．

表 2.5 に，影響グループに対するデモグラフィック属性の分布を示す．一定の分布の片寄が学部，性別ともに見られるが，顕著でなくその解釈が困難である．各影響グループにおいてより多くの利用者が実行した上位 20 アプリケーションを表 2.6 に示す．表からは，各影響グループにおける顕著な利用アプリケーションの嗜好性を読み取ることは困難である．しかしながら，いくつかの特徴的なアプリケーションを見つけることができる．例えば，G4 には薬学部の学生が多く含まれており，薬品の検索アプリケーションが G4 では特徴的によく使われていることが読み取れる．

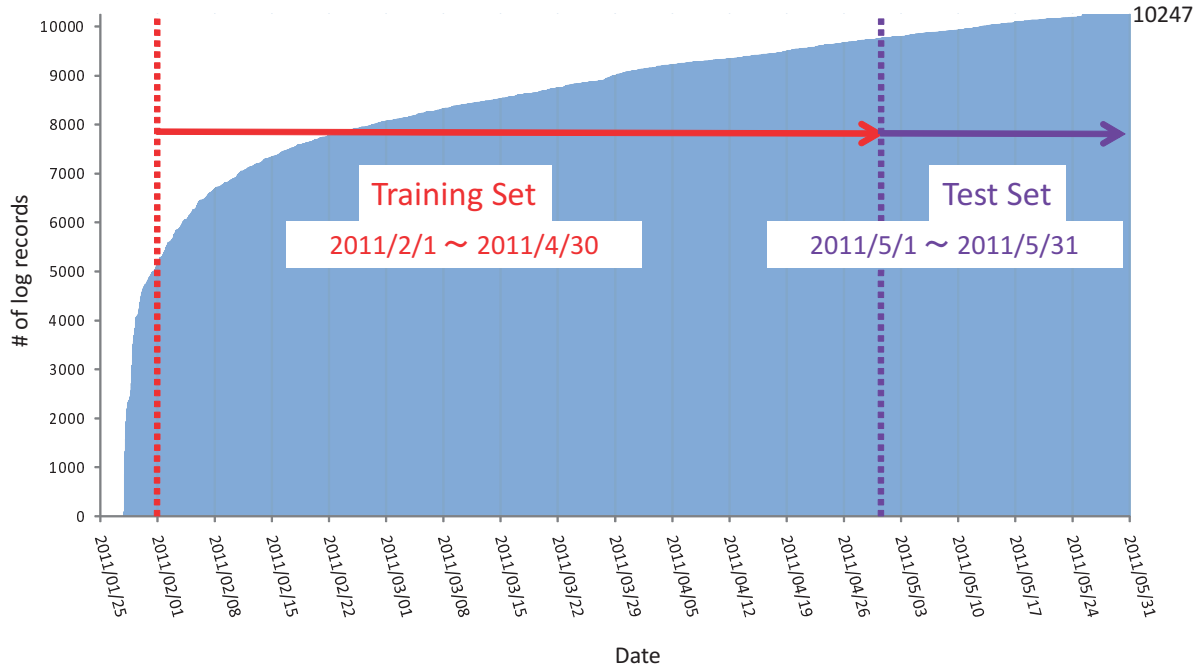


図 2.5 履歴レコードの累積数

表 2.4 図 2.8 において最善の結果を示した NMF アルゴリズム

ランク数	最善の結果を示した NMF アルゴリズム
2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20	Non-smooth NMF based on Kullbach-Leibler divergence (nsNMF)
25, 30, 35, 40, 45, 50, 60	Pattern-Expression NMF based on Euclidean distance (PE-NMF)

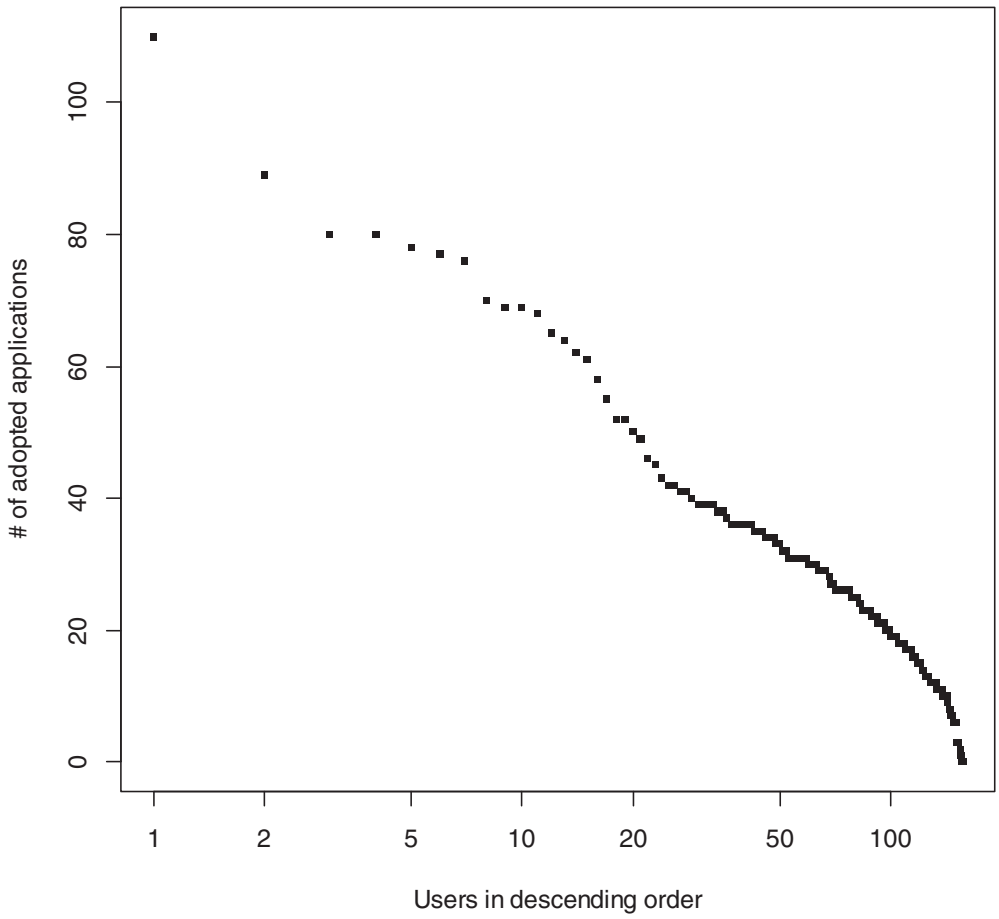


図 2.6　実行アプリケーション数（降順ソート）

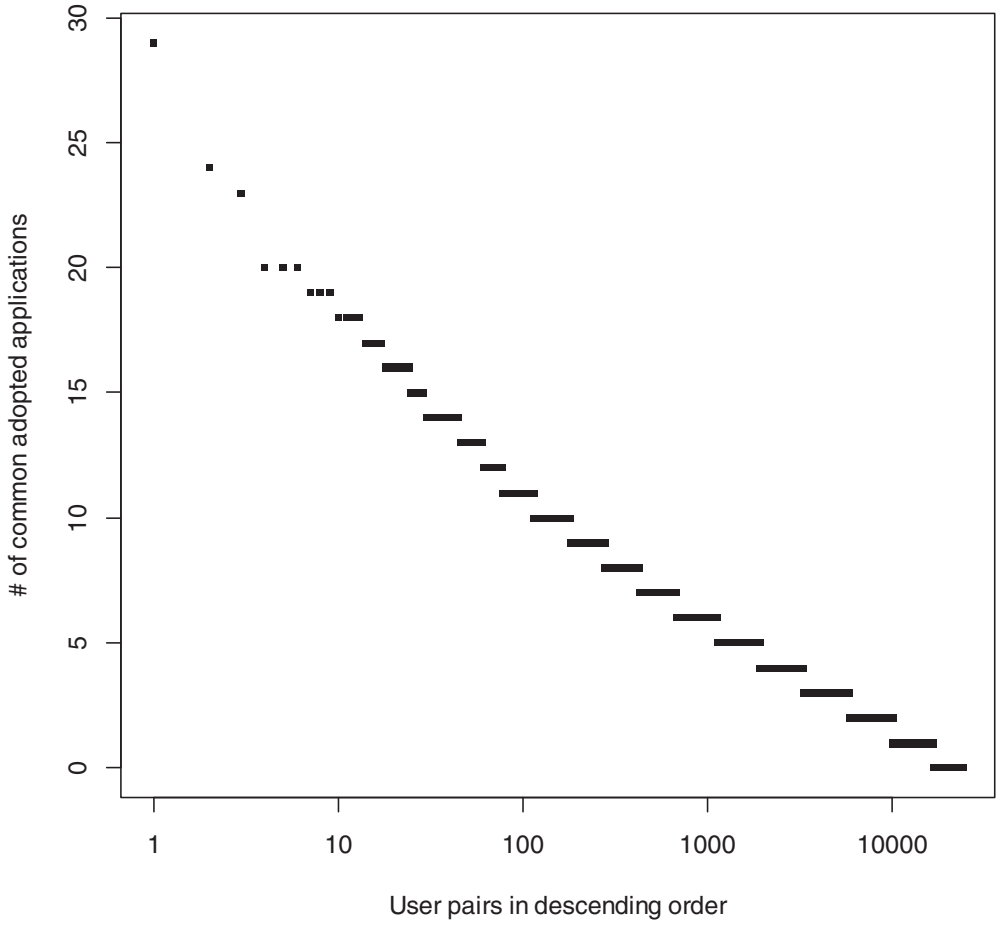


図 2.7 利用者間の共通実行アプリケーション数（降順ソート）

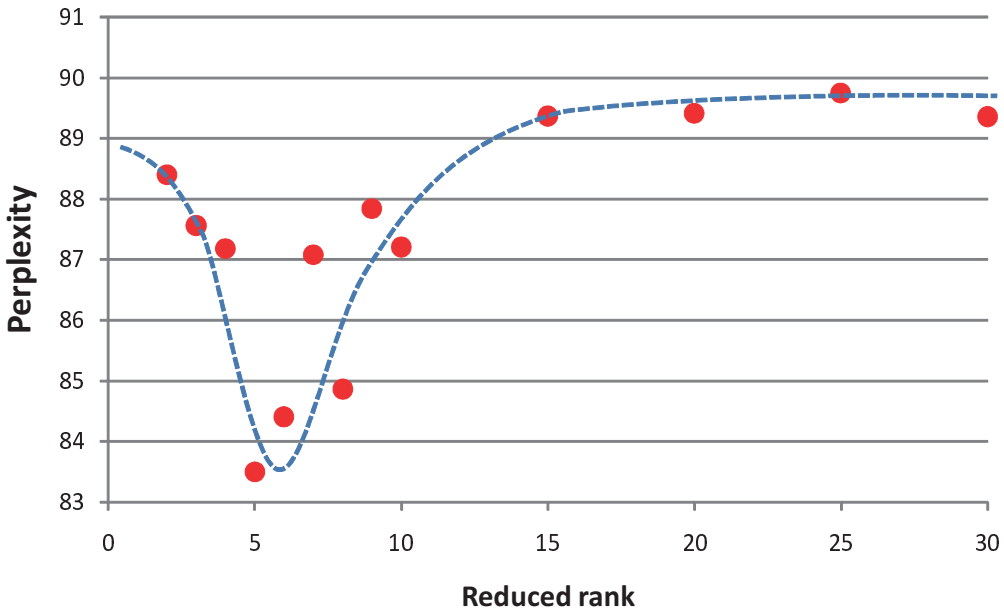


図 2.8　NMF により低ランク近似した影響行列の予測性能（perplexity）

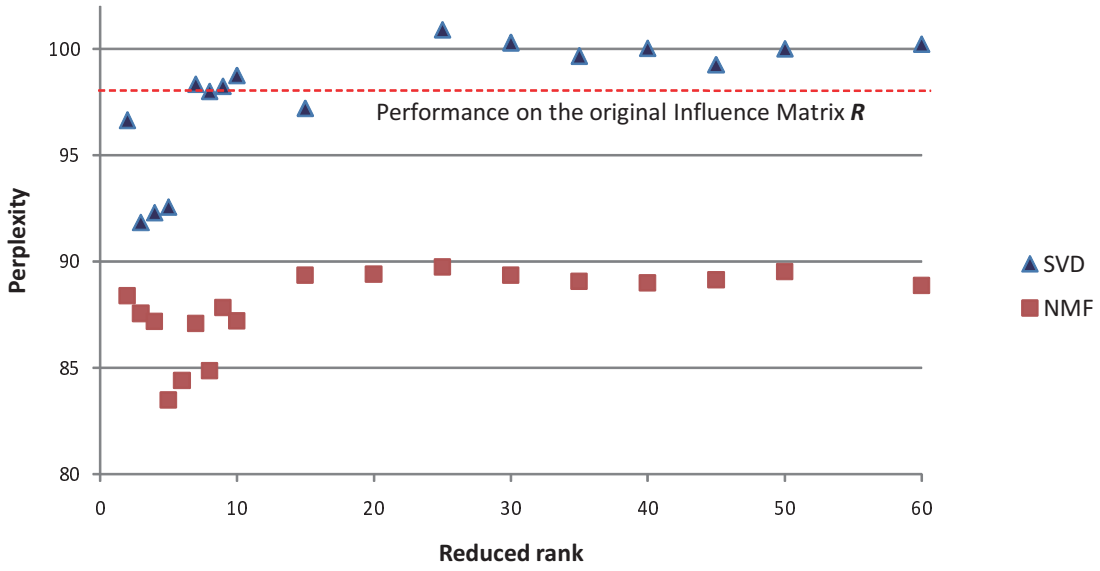
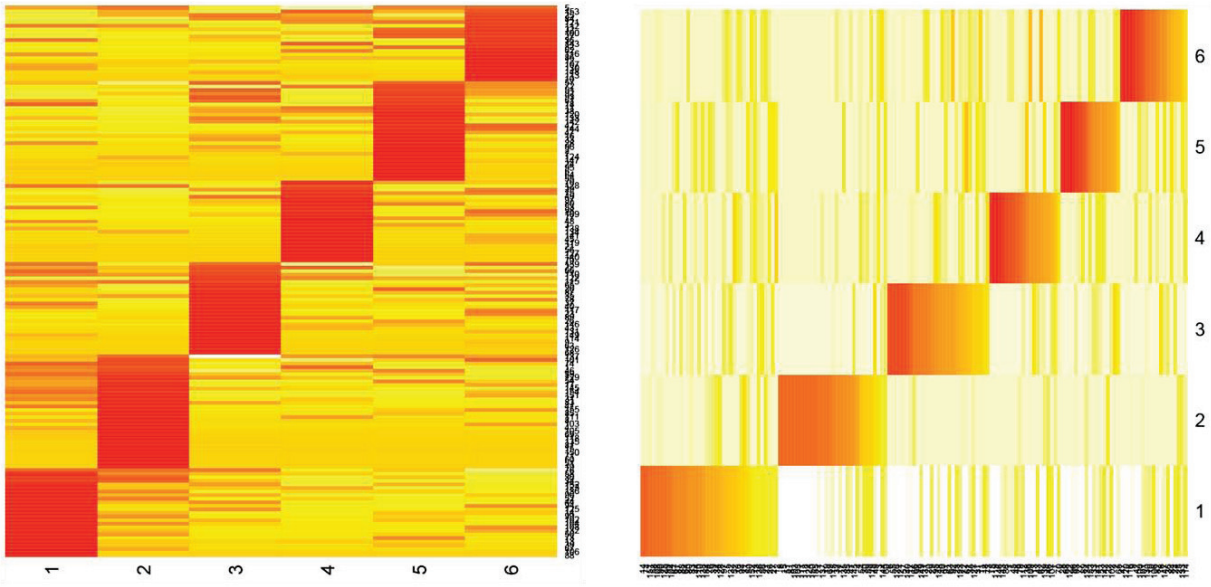


図 2.9 SVD により低ランク近似した影響行列の予測性能（perplexity）



(a) Coefficient matrix　　(b) Basis matrix

図 2.10　影響行列の行列分解結果 ($k = 6$)

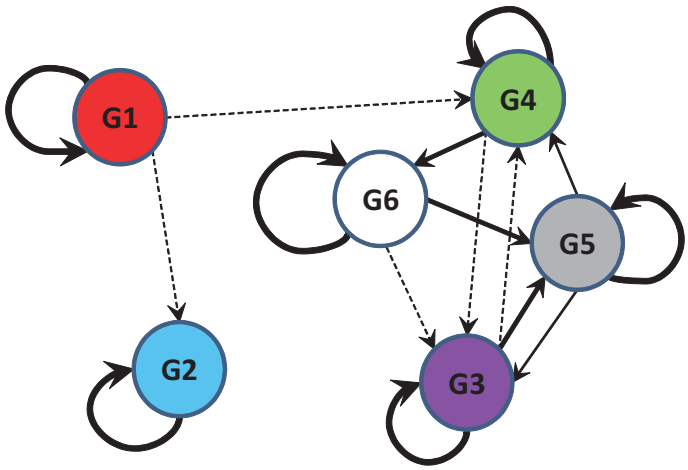


図 2.11　影響グループ間の関係

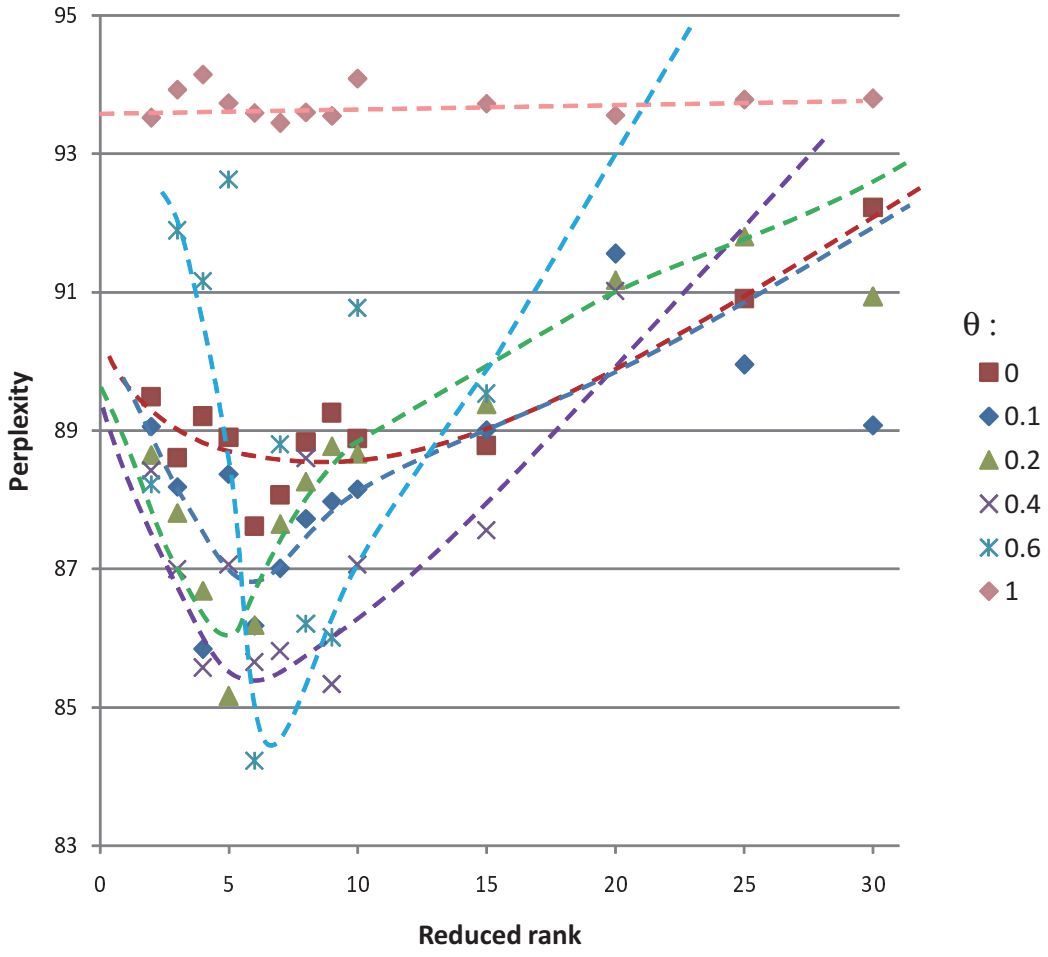


図 2.12　nsNMF におけるパラメータ θ と予測性能（perplexity）の関係

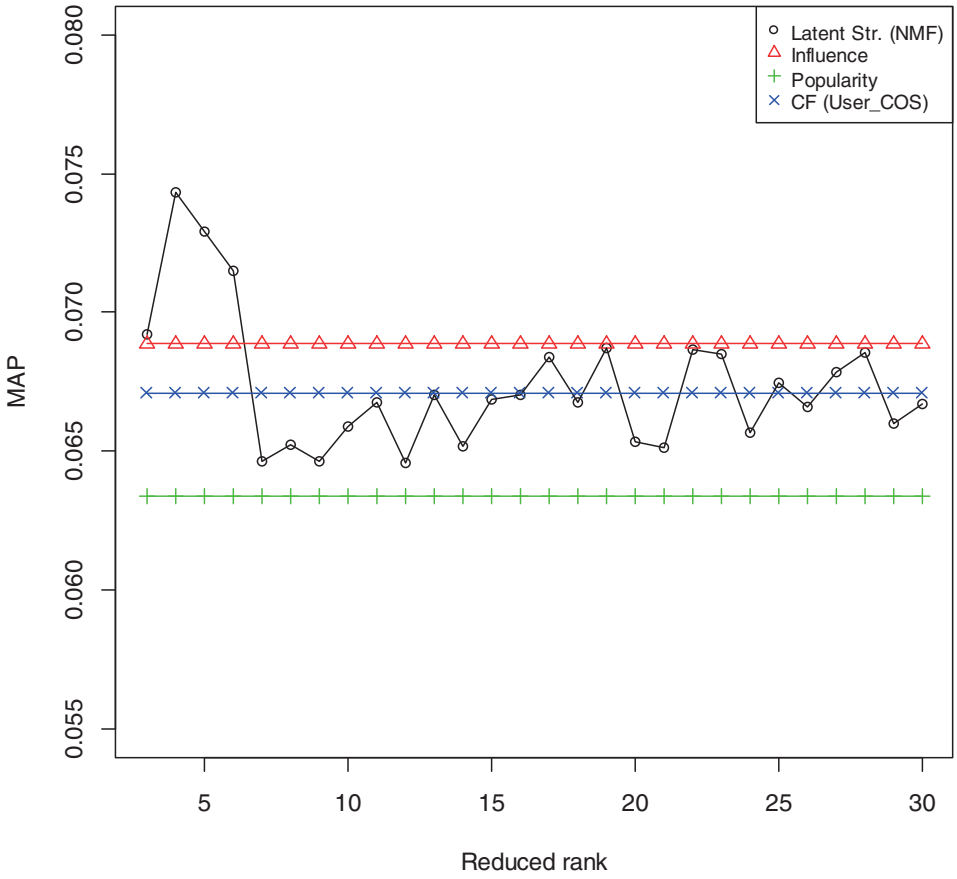


図 2.13　ランク数と予測性能 (MAP)

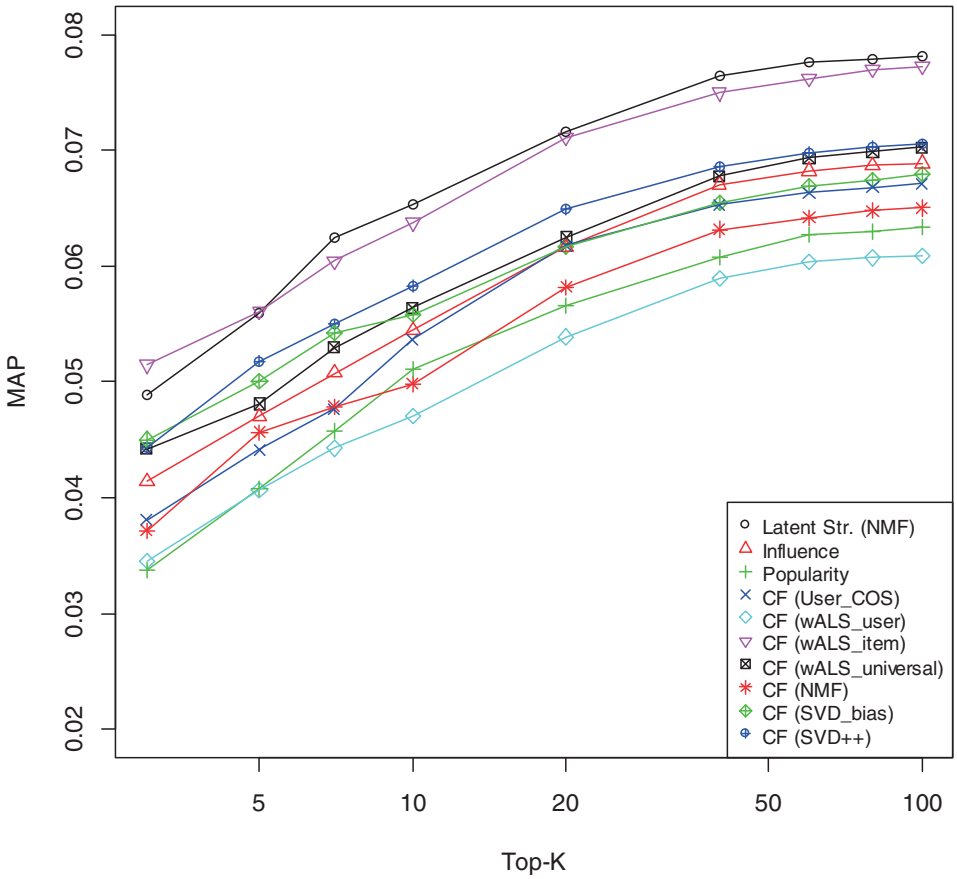


図 2.14 予測性能 (MAP) の比較

2.4.2　その他の因子との相関度合い

利用者の属性等と影響の強さとの間に相関関係があるのであれば，これを活用して予測性能の向上が図れる可能性がある．このため，利用者間の属性の関係と影響の強さとの間の相関分析を試みた．すべての利用者組について，ある属性の値が等しい組と異なる組の 2 グループに分類し，影響の強さに関してこの 2 グループの間の相関比を算出した．分析した属性は，学部，入学年，応募グループ（友人グループ；2.3.1 節参照），実行したアプリケーション数が平均以上／以下，所属する影響グループ，である．

表 2.7 に結果を示す．所属する影響グループ以外に関しては相関関係は認められなかった．この結果から，影響の強さとこれらの属性が一致していることとの間には強い関係を見出せず，すなわちこれらの属性を用いて影響の強さを推定することは困難であることがわかる．

2.4.3　影響の授受度合いの可視化

バイラル・マーケティング（例えば [20]）への活用を想定すると，周囲の人々へクチコミ等により直接影響を及ぼすインフルエンサ（オピニオン・リーダー）を見つけることができれば，働きかける先を絞ることが可能となり全体の効率を高めることができ有用である．これまで論じてきた影響の強さを利用すると，次のように各利用者に対して周囲影響を与える強さ $influentiality^i$ および周囲からの影響を受けやすさ $influenceability^i$ を定義することができる．ここで添え字の ^{i} は暗黙の影響に基づいていることを示している．

$$influentiality_u^i = \frac{\Sigma_v P_r(u \to v)}{U_{exp} - 1} \tag{2.5}$$

$$influenceability_u^i = \frac{\Sigma_v (P_r(v \to u))}{U_{exp} - 1} \tag{2.6}$$

図 2.15 に，影響行列 $\mathbf{R}$ と最も予測性能が高くなった近似影響行列 $\widehat{\mathbf{R}}^{(6)}$ を用いてそれぞれ算出した各利用者の $influentiality^i$ と $influenceability^i$ の散布図を示す．図中の各プロットは各利用者を表す．また矢印は，各利用者のプロットの影響行列 $\mathbf{R}$ による位置から近似影響行列 $\widehat{\mathbf{R}}^{(6)}$ による位置への変動分を示しており，大きな変動が生じていないことを確認できる．また，$influentiality^i$ と $influenceability^i$ の間には顕著な相関関係が認められない．さらには外れ値の存在も確認できる．履歴の中に少数のアプリケーションの実行しかない場合には，$P_r(u \to v)$ の算出値が不安定となることが考えられ，直感的には外れ値の原因と推定される．

上記の $influentiality^i$ および $influenceability^i$ は，暗黙の影響に基づいて算出したものであるので，2.1 節で述べたとおり直接的に及ぼす影響以外の効果も含まれていることに注意が必要である．したがって，$influentiality^i$ の大きい利用者が必ずしも真のインフルエンサである

とは限らない．この真のインフルエンサを見つけ出すためには，さらなる工夫が必要となる．この課題については第 3 章にて述べる．

2.5 むすび

本章では，スマートフォンのアプリケーション利用履歴を取り合げて，暗黙の影響構造について分析し，潜在グループ構造を発見したことを述べた．そのためには非負行列分解（NMF）をスパースネス制御とともに用いることが重要であった．低ランク近似を影響行列に施し予測能力を観察することから，潜在的な影響グループの存在を実験的に確認した．また，影響グループ内の相互影響と，影響グループ間の影響に分類し，影響関係を解釈した．この潜在グループモデルに基づきアプリケーションの利用予測を行う手法を提案し，最先端の既存手法と比較しても高い精度で予測が可能であることを示した．

本研究においては，特定の実データを用いた実験結果によってしか提案モデルの有効性を確認していない．モデルの有効性はデータ（すなわち利用者の振る舞い）に依存するため，様々な異なる種類の振る舞いにおいて，提案モデルが有効であるかどうかは明らかでない．しかしながら，本検討により少なくともスマートフォンの利用履歴において本モデルが有効であることが確認されたことは，他の類似タスクにおいても本モデルの有効性を検討するに値するという意味において有用な知見を提供するものと考える．

今後の発展としては，より洗練化した拡散モデルを利用することによりさらなる性能改善を図ることがあげられる．また，大規模データに対しても対応できるようスケーラブルな手法に発展させることが現実問題に適用する上では必須である．さらには，影響関係はその内容トピックに依存することが考えられるため，トピック毎に影響関係を推定するような拡張が必要・有望と思われる．また本章では，暗黙の影響関係を扱ったため，直接的な影響関係は必ずしも正しく獲得できない．これについては第 3 章にて述べる．

表 2.5　影響グループ構成者の属性分布

		影響グループ						
		G1	G2	G3	G4	G5	G6	合計
所属	文学部	1	1	2	0	1	0	5
	人間科学部	3	6	1	0	2	2	14
	法学部	2	3	2	2	1	2	12
	経済学部	2	2	0	3	4	2	13
	理学部	1	1	3	0	2	0	7
	薬学部	5	3	1	8	3	3	23
	医学部	1	1	1	0	0	0	3
	工学部	4	3	4	1	8	5	25
	基礎工学部	3	6	10	6	4	4	33
	外国語学部	3	6	1	3	3	3	19
	情報科学研究科	0	0	1	0	0	0	1
性別	女性	11	10	2	10	5	8	46
	男性	14	22	24	13	23	13	109
合計		25	32	26	23	28	21	155

表 2.6 各影響グループにおける人気アプリケーション

順位	影響グループ					
	G1	G2	G3	G4	G5	G6
1	com.android.calculator2	Android	com.android.inputmethod.latin	com.cookpad.android.activities	com.cooliris.media	com.google.android.voicesearch
2	com.android.inputmethod.latin	com.google.android.providers.enhancedgoogle search	net.binzume.android.nicoplayer	com.moxier.mail	com.android.packageinstaller	com.sonyericsson.android.basicwords
3	com.facebook.katana	com.sonyericsson.android.timescape	com.android.alarmclock	cn.bluesky.neatreversi	com.android.wallpaper.livepicker	com.mobisystems.office
4	com.mobisystems.office	com.sonyericsson.conversations	com.android.packageinstaller	jp.co.gnavi.activity	com.facebook.katana	com.sonyericsson.pccompanion
5	com.sonyericsson.quadrapop	com.android.browser	com.google.android.youtube	com.android.calculator2	com.google.android.voicesearch	com.sonyericsson.quadrapop
6	com.android.vending	com.google.android.youtube	com.pdanet	com.google.android.apps.maps	com.google.zxing.client.android	com.spritemobile.backup.semc
7	com.cooliris.media	com.android.phone	com.skype.raider	com.sonyericsson.android.servicemenu	com.joelapenna.foursquared	jp.picolyl.led_light
8	com.sonyericsson.android.basicwords	com.google.android.apps.maps	com.adobe.reader	com.sonyericsson.quadrapop	com.nttdocomo.android.compass	com.android.calculator2
9	com.sonyericsson.android.iwnnime	com.mobisystems.office	com.android.calculator2	de.joergjahnke.mario.android.free	com.sonyericsson.android.friendpivot	com.cooliris.media
10	com.sonyericsson.android.servicemenu	com.sonyericsson.android.camera	com.cooliris.media	jp.co.c_lis.ccl.medicinesearch.android	com.sonyericsson.android.wallpaperchooser	com.facebook.katana
11	com.sonyericsson.setupwizard	com.sonyericsson.android.mediascape	com.evernote	jp.fuukiemonster.webmemo	com.sonyericsson.pccompanion	com.google.android.street
12	com.android.globalsearch	com.sonyericsson.android.socialphonebook	com.justsystems.atokmobile.trial.service	nikeno.Tenki	jqsoft.apps.mysettings	com.google.android.talk
13	com.android.wallpaper.livepicker	com.android.launcher	com.nttdocomo.android.compass	wni.WeathernewsTouch.jp	com.adobe.reader	com.nttdocomo.android.compass
14	com.nttdocomo.android.docomo_market	com.google.android.gm	com.sonyericsson.textinput.uxp	com.adobe.reader	com.android.alarmclock	com.sakura.News2010
15	com.sonyericsson.android.socialservicesetting	com.sonyericsson.search	jp.bustercurry.virtualtenho_g	com.android.alarmclock	com.android.htmlviewer	com.sonyericsson.android.servicemenu
16	com.sonyericsson.pccompanion	com.android.settings	jp.radiko.gui.main	com.android.globalsearch	com.android.music	com.sonyericsson.trackid3.client
17	com.sonyericsson.textinput.uxp	com.android.vending	org.adw.launcher	com.android.providers.applications	com.bumptech.bumpga	com.taptu.wapedia.android
18	com.android.alarmclock	com.google.android.voicesearch	asia.sonix.sekaimeigenwidget	com.bumptech.bumpga	com.clapfootgames.tankhero	de.shandschuh.sparserss
19	com.android.calendar	com.sonyericsson.android.friendpivot	com.android.phone	com.ekitan.android	com.donapon.pisces.cashbook	jp.co.c2inc.medicallite.first
20	com.cootek.touchpal	jp.co.sonyericsson.android.playnow	com.android.providers.applications	com.google.android.location	com.fgol.sharkfree	jp.co.labelgate.moratouch

表 2.7　相関比による分析

因子	相関比	
	$\mathbf{R}$	$\widehat{\mathbf{R}}^{(6)}$
所属学部が同じである組—それ以外	0.020142	0.017418
入学年が同じである組—それ以外	0.006050	0.001535
同じ申込団体である組—それ以外	0.019159	0.018326
双方とも実行アプリ数が平均以上の組—それ以外	0.068715	0.098507
双方とも実行アプリ数が平均以下の組—それ以外	0.063818	0.078679
同一影響グループに所属する組—それ以外	0.092977	0.208375

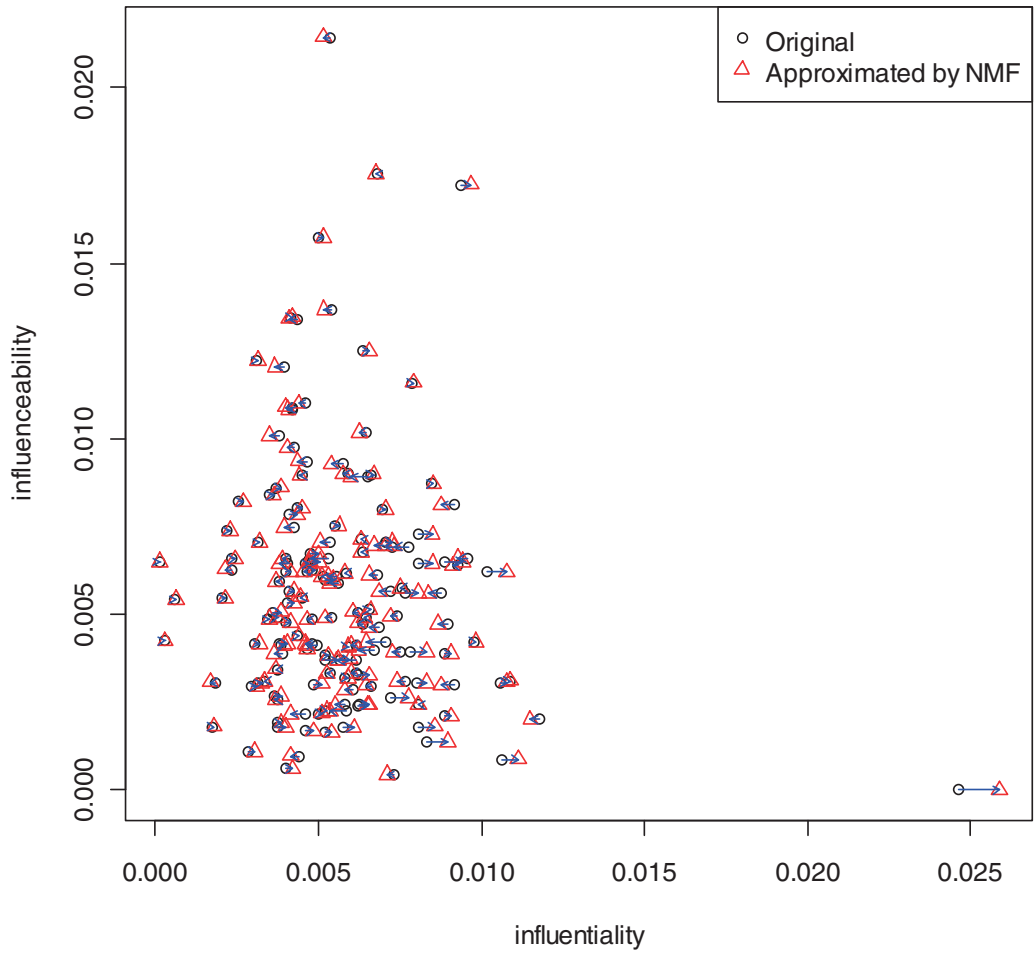


図 2.15　影響の授受度合いの分布

第 3 章

友人関係情報を用いたインフルエンサの推定

3.1 はじめに

マーケティングの研究から，人々の様々な行動（購買行動や視聴行動など）は周囲から影響を受けていること，また周囲へ与える影響度は人によって異なり，この影響度の強い人（インフルエンサ）が存在すると言われている．このインフルエンサを活用することによって，販売促進等の効率を向上することが可能であるため，インフルエンサを推定することは重要であり注目を集めている [9, 20]．ここで，第 2 章で議論した暗黙の影響関係には，インフルエンサからの直接的影響に加え，間接的な影響や行動特性としての先行性による効果等が含まれており，暗黙の影響力が強い人とインフルエンサとは等しくないことに注意する．

影響の与え方には，直接の対人コミュニケーションによるものからソーシャルメディアやマスメディアを通じたものまで，多様な形態が想定されるが，本章では，特に実世界における周囲（友人知人家族等）に対する影響度が強い人を，利用履歴の観察により推定する手法を提案する [76]．ここでは，スマートフォンのアプリケーション利用における周囲へ与える影響を題材として，インフルエンサ推定を行うが，提案手法はスマートフォンの利用履歴以外であっても，ユーザ間の連鎖に意味が考えられる対象に対して，広く適用可能である．以降本論文では特に断らない限り，実世界における周囲（immediate environment）に居る他の人（友人知人家族等）に対して影響を与える個人をインフルエンサと呼ぶ．この定義は Katz らによりオピニオン・リーダの定義として示されたもの [30] で，現在でもインフルエンサの定義として変わらずに広く使われている [70]．さらに，個人の周囲に対して及ぼす具体的な影響行為をインフルエンスと呼ぶ．また，2 人が友人知人家族等の関係であるか否かを示す情報のことを友人関係情報と呼ぶことにする．ここでの友人関係とは，直接的な対人コミュニケーションを行う相手との関係を指すので，いわゆる友人同士の関係に加えて，家族等の相手も含まれることに注意

する．

一方，人々が新しい事物を採用する行動には，個人によりその時期（先行性）に違いがあることが Rogers により示されている [60]．Rogers は，先行性によって 5 つのカテゴリ（イノベータ，アーリーアダプタ，アーリーマジョリティ，レイトマジョリティ，ラガード）に分類した．イノベータが新しい事物を最も早く採用する層であり，ラガードは最も遅く採用する層である．購入履歴，利用履歴を観察することにより，イノベータやアーリーアダプタ等の先行者を把握することは容易である．しかしながら，先行しているからと言って，他者の行動に影響を及ぼしているとは限らないことに注意が必要である．Rogers はこれについても指摘しており，インフルエンサは偏りはあるものの，イノベータからレートマジョリティ／ラガードにかけて分布すると述べている．このため，先行度からのインフルエンサ推定は困難である．

ここで，新しい事物が人々によって採用されていくことを普及（diffusion）と呼ぶことにする．インフルエンサが与えた影響により起こる普及を直接的インフルエンス，直接的インフルエンス以外により生じる普及を特に拡散的普及と呼ぶことにする．文献によっては拡散的普及を含んだ普及のことをインフルエンスと呼ぶ場合があるが，本検討では明確に区別して用いる．

普及に関する研究は，Katz らの研究 [30] に端を発し，Rogers のイノベータ理論 [60]，Bass の普及モデル [5] 等を基礎として，特に経済学・社会学において多くの研究がなされてきた．多くの研究は普及をマクロに扱うものであったが，近年になって個人単位の普及を扱う研究がなされるようになってきた [56]．インフルエンサの推定を行うためには，普及を個人単位に扱う必要がある．個人単位の普及を扱う研究としては，次のようなものがある．

(1) 個人間の普及確率を求める [24, 62, 52]
(2) 全体の普及を最大化する個人の集合を求める [33, 35]
(3) 普及関係から行動を予測する [54, 64, 28, 61, 81, 31]
(4) 属性が類似している者同士で普及が増大する傾向（ホモフィリー）による効果と，直接的インフルエンスによるものとを区別する [2]

上記のうち，(1)~(3) は個人間の普及を扱っていて，普及のなかから直接的インフルエンスを区別し抽出することは意図していない．直接的インフルエンスを扱っているのは (4) のみであるが，直接的インフルエンスとホモフィリーの効果の比率を明らかにすることが狙いであるため，個人間の直接的インフルエンスの発生確率を個別に明らかにすることができない．Goyal らは文献 [24] において，個人間の普及のうち友人間の部分のみを取り出し，その時間差が基準より小さいものを直接的インフルエンスと見なす手法を述べているが，その妥当性について検証が十分なされていない．このため 3.4 節にて提案手法と比較しその妥当性を議論する．

また，応用の観点から類似・関連する研究として，Pan らによるスマートフォンアプリケー

ションの利用予測 [54]，Richter らによる電話の解約予測 [59] がある．Pan らの研究は種々のソーシャルグラフ（友人関係情報）が得られる時に利用予測能力が最適となるような友人関係情報を得る取組である．また Richter らの研究では，過去の通話履歴から密なつながりを持つと思われるグループを抽出するとともにその中でのインフルエンサを通話履歴のパターンから推定し，それらの特徴により解約予測ができると主張している．双方とも，個人間の影響を扱う研究であるが，Pan らの研究は直接的インフルエンスの抽出を意図していない．Richter らの研究は，通話という直接観測される友人関係からインフルエンサを抽出しているが，本論文ではスマートフォンのアプリケーション利用履歴という間接観測から推定するという問題を扱っており，彼らの手法は適用することができない．

本検討では，利用履歴からインフルエンサを推定することを目的とする．前述したように，利用履歴から利用者の先行性は把握できるが，インフルエンサは同定することができない．そこで，友人関係が既知であるとして，友人に対する普及には直接的インフルエンスが拡散的普及に追加されることをモデルとして表現し，これにより各人の直接的インフルエンスの強さを統計的に推定しインフルエンサを抽出することを試みる．また，実際には完全な友人関係は容易に得ることができないため，一部の観測から得られる友人関係を手がかりに全体を推定する手法について第 4 章にて議論する．著者の知る限りにおいて，類似のインフルエンサを推定する手法は公知文献として発表されていない．より具体的には，下記の枠組みによるインフルエンサの推定手法を本検討では提案する．

- 観測（入力）
  - 利用履歴：スマートフォン・アプリケーション・ダウンロード履歴
  - アンケート結果：友人であるか否かの情報（友人関係情報）
- 推論対象（出力）
  - 各個人の直接的インフルエンスの強さ
- 推論手法
  - 普及の発生を，ベータ分布としてモデル化
  - 拡散的普及と直接的インフルエンスの組み合わせにより普及発生の分布パラメータをモデル化
  - マルコフ連鎖モンテカルロ（MCMC）法を用いたベイズ統計による未知母数の推定

本検討の主張点は以下の 2 点である．

- 友人関係情報を利用し，利用履歴の分析からインフルエンサを推定する手法の提案．具体的には個人間の影響度を直接的インフルエンスとそれ以外の要因の組み合わせで表し，ダウンロードの連鎖はこれをパラメータとするベータ分布に従うとする確率モデルによる，直接的インフルエンスの強さの統計的推定方法として実現．

- 大阪大学学生約 160 人・3 か月分のスマートフォン利用履歴からのインフルエンサ推定結果とアンケート結果との照合による提案手法の有効性の確認.

以下，3.2 節で提案するモデルおよび未知母数の推定手法について説明する．3.3 節では実験で利用するデータについて説明する．3.4 節では評価実験の内容および結果と考察を述べ，3.5 節で本章のまとめについて述べる.

3.2　アプリケーション・ダウンロードのモデル

3.2.1　ダウンロードの連鎖

第 2 章にて，アプリケーション・ダウンロードの連鎖の発生する確率を，すべての利用者間において求めることによって，各利用者が将来ダウンロードするアプリケーションを予測できることを示した．本章においては，連鎖の発生する確率を同様に基本として，インフルエンサが推定できるように連鎖発生の確率を直接的インフルエンスによるものと拡散的普及によるものの組み合わせとして表現することによって，モデルを拡張する．以降の説明にて用いる表記をまとめて表 3.1 に示す.

アプリケーションダウンロードの連鎖は，利用者 i が各利用者 j に対して持っている連鎖発生確率 $p_{i\to j}$ に従って生起すると仮定する．ユーザ i がアプリケーション a をダウンロードした場合，その後はいつでも当該アプリケーション a を未ダウンロードの利用者 j に対して連鎖を起こそうとする．周囲に対する連鎖を起こそうとする試行はベルヌーイ試行としてモデル化する．このため，試行の成功回数は二項分布となり，観測から推定される成功確率の確率密度分布はベータ分布となる.

また，観測される利用履歴から，利用者 i が利用履歴の期間中にダウンロードしたアプリケーションの数 $||A_i||$ および，利用者 i がダウンロードした後に利用者 j がダウンロードしたアプリケーションの数 $||A_{i\to j}||$ が得られる．これにより利用者 i が利用者 j に対して連鎖を起こそうとする試行の成功確率 $p_{i\to j}$ の最尤推定結果（期待値）に相当する観測値 $y_{i\to j}$ が次式により得られる.

$$y_{i\to j} = \frac{||A_{i\to j}||}{||A_i||} \tag{3.1}$$

3.2.2　直接的インフルエンスと拡散的普及

図 2.2 の例において，日常的にこの 3 人が良くコミュニケーションをとっていて，このような連鎖が頻繁に観察される場合には，利用者 $2 \to 1 \to 3$ の順で影響が与えられている可能

表 3.1 表記一覧

表記	意味
$p_{i\to j}$	利用者 i から利用者 j に対する連鎖の発生確率
$q_{i\to j}$	利用者 i から利用者 j に対する拡散的普及による連鎖の発生確率
v_i	利用者 i が周囲に直接的インフルエンスを与える強さを示すパラメータ
α, β	ベータ分布におけるパラメータ $Beta(x\|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$
γ, θ	α,β から変換により得られるベータ分布のパラメータの別表現 $\gamma = \frac{\alpha}{\alpha+\beta}, \theta = \alpha+\beta$
A_i	利用者 i がダウンロードしたアプリケーションの集合
$A_{i\to j}$	利用者 i がダウンロードし，その後利用者 j がダウンロードした アプリケーションの集合
$y_{i\to j}$	A_i のうち $A_{i\to j}$ に含まれるアプリケーションの比率
t_{ij}	利用者 i と利用者 j の間の友人関係の有無
s_i	利用者 i の先行度
$r_{a,i}$	利用者 i がアプリケーション a をダウンロードした実験協力者中の順位
$t_{a,i}$	利用者 i がアプリケーション a をダウンロードした時刻
τ_{ij}	$A_{i\to j}$ に含まれるアプリケーションの，利用者 i と利用者 j における ダウンロード時刻差の平均
u_a	アプリケーション a をダウンロードした実験協力者の人数

性が示唆されるものの，利用者 2 がイノベータであり，利用者 3 がラガードである場合の様に，友人関係がなくても利用者 2,1,3 の順で先行性が高いため見かけ上このような観察が得られていることも考えられ，これらを利用履歴のみからは区別することができない．利用履歴から観測される連鎖には，このように直接的インフルエンスによる連鎖とそれ以外の要因による連鎖（拡散的普及による連鎖）が混在していると考えることができる．拡散的普及による連鎖は，先行しやすさを相対的に表現したものとも言える．また，友人関係のない相手には直接的インフルエンスは発生し得ないことから，友人以外に対する普及と友人に対する普及との差異には，直接的インフルエンスの強さが反映されていると考えられる．

本検討では，上記の連鎖を作りだすメカニズムとして確率モデル（生成モデル）を設定し，観測された標本に基づいて確率モデル中の未知母数を統計的に推定する．真の確率モデルは未知であるので，ここでは直接的インフルエンスと拡散的普及の組み合わせにより普及が発生するという仮説に基づいた確率モデルを設定する．そして，実験により得た実際の観測データを用いて未知母数の一つである直接的インフルエンスの強さを統計的に推定し，その結果をインフルエンサの正解データと照合することにより仮説（モデル）の妥当性を検証する．

3.2.3　確率モデル

本章で提案する，普及（連鎖）の発生メカニズムとしての確率モデルを得る上での，基本的な考え方は次の通りである．

(1) 利用者 i から利用者 j への普及および拡散的普及をベルヌーイ試行としてモデル化（近似）する．またその発生確率（試行の成功確率）の密度分布はベルヌーイ試行との親和性からベータ分布にて表す．
(2) 各利用者は各自の特性として，どのような確率で拡散的普及を起こす（起点側）および拡散的普及を受ける（終点側）かの密度分布を持ち，その分布はベータ分布で表されると仮定する[*1]．
(3) 拡散的普及の起点側・終点側の密度分布は，相互に全体確率分布の上で独立であることを仮定し，任意の起点，終点の組における拡散的普及の近似計算を可能とする．
(4) 各利用者には，固有の周囲に与える直接的インフルエンスの強さを示す値があり，拡散的普及の発生確率に相手が友人であれば加算され，またその強さは相手に依存せずどの友人に対しても等しく作用することを仮定しモデル化する．

上記の考え方 (1)~(4) に従い，確率モデルを定式化する．まず，利用者 i から利用者 j に対する拡散的普及による連鎖の発生確率 $q_{i\to j}$ は，考え方 (1) により密度分布をベータ分布として

[*1] 3.3 節で説明する収集データに関して，非友人に対しての $y_{i\to j}$ の分布の特性を分析し，拡散的普及の起点側・終点側の密度分布を近似的にベータ分布で表すことの妥当性は確認済みである．

モデル化するので分布パラメータ $\alpha_{q_{i\to j}}, \beta_{q_{i\to j}}$ を用いて式 3.2 として表現される．

$$q_{i\to j} \sim Beta(\alpha_{q_{i\to j}}, \beta_{q_{i\to j}}) \tag{3.2}$$

観測 $y_{i\to j}$ は，式 3.1 より各 i, j の組に対して単一の値しか得られないため，観測 $y_{i\to j}$ から直接的にベータ分布のパラメータ $\alpha_{q_{i\to j}}, \beta_{q_{i\to j}}$ を得ることができない．これに対して，考え方 (2)，(3) を用いることで，式 3.3 により $q_{i\to j}$ を近似として得ることができる．ここで，$q_{i\to *}$ は利用者 i の起点側の拡散的普及の発生確率を，$q_{*\to i}$ は終点側を，さらに $q_{*\to *}$ はすべてのユーザ間における拡散的普及の発生確率を示す．

$$q_{i\to j} = \frac{q_{i\to *} \times q_{*\to j}}{q_{*\to *}} \tag{3.3}$$

ベータ分布の確率密度分布の式 ($Beta(x|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$) を用いて式 3.3 を展開し，式 3.2 の分布パラメータ $\alpha_{q_{i\to j}}, \beta_{q_{i\to j}}$ を式 3.4，式 3.5 から得ることができる．これにより拡散的普及による連鎖の発生確率 $q_{i\to j}$ のモデル化を終える．

$$\alpha_{q_{i\to j}} = \alpha_{q_{i\to *}} + \alpha_{q_{*\to j}} - \alpha_{q_{*\to *}} \tag{3.4}$$

$$\beta_{q_{i\to j}} = \beta_{q_{i\to *}} + \beta_{q_{*\to j}} - \beta_{q_{*\to *}} \tag{3.5}$$

次に，考え方 (1)，(4) に従い拡散的普及と直接的インフルエンスを組み合わせて普及発生確率 $p_{i\to j}$ のモデルをベータ分布にて表現する．まず，利用者 i の直接的インフルエンスの強さを表す変数 v_i および友人関係の有無を示す変数 t_{ij} を導入して，普及のベータ分布のパラメータを，別表現 $(\gamma_{p_{i\to j}}, \theta_{p_{i\to j}})$ を用いて式 3.6，式 3.7 にてモデル化する．式 3.6 は，分布の期待値（平均値）である $\gamma_{p_{i\to j}}$ を，拡散的普及の期待値 $\gamma_{q_{i\to j}}$ に対して友人関係のある相手の場合は直接的インフルエンスにより v_i が加算されるものとして表現している．また式 3.7 は，分布の先鋭度 $\theta_{p_{i\to j}}$ が，拡散的普及の先鋭度 $\theta_{q_{i\to j}}$ と同じ値を持つことを表している．

$$\gamma_{p_{i\to j}} = \gamma_{q_{i\to j}} + v_i \times t_{ij} \tag{3.6}$$

$$\theta_{p_{i\to j}} = \theta_{q_{i\to j}} \tag{3.7}$$

ここで，t_{ij} は次式で与えられる．

$$t_{ij} = \begin{cases} 1 & \text{if 利用者 } i \text{ が利用者 } j \text{ と友人関係にある} \\ 0 & \text{それ以外} \end{cases} \tag{3.8}$$

式 3.6，式 3.7 のパラメータ $\gamma_{p_{i\to j}}, \theta_{p_{i\to j}}$ を $\alpha_{p_{i\to j}}, \beta_{p_{i\to j}}$ の形に変換し，普及発生確率 $p_{i\to j}$ を表すモデルとして式 3.9 を得る．

$$p_{i\to j} \sim Beta\Big((\gamma_{q_{i\to j}} + v_i \times t_{ij}) \times \theta_{q_{i\to j}}, \Big(1 - (\gamma_{q_{i\to j}} + v_i \times t_{ij})\Big) \times \theta_{q_{i\to j}}\Big) \tag{3.9}$$

ここで，$\gamma_{q_{i\to j}}$, $\theta_{q_{i\to j}}$ は次式で与えられる．

$$\gamma_{q_{i\to j}} = \frac{\alpha_{q_{i\to j}}}{\alpha_{q_{i\to j}} + \beta_{q_{i\to j}}} = \frac{\alpha_{q_{i\to *}} + \alpha_{q_{*\to j}} - \alpha_{q_{*\to *}}}{\alpha_{q_{i\to *}} + \alpha_{q_{*\to j}} - \alpha_{q_{*\to *}} + \beta_{q_{i\to *}} + \beta_{q_{*\to j}} - \beta_{q_{*\to *}}} \tag{3.10}$$

$$\theta_{q_{i\to j}} = \alpha_{q_{i\to j}} + \beta_{q_{i\to j}} = \alpha_{q_{i\to *}} + \alpha_{q_{*\to j}} - \alpha_{q_{*\to *}} + \beta_{q_{i\to *}} + \beta_{q_{*\to j}} - \beta_{q_{*\to *}} \tag{3.11}$$

以上により，提案する確率モデルに基づいて直接的インフルエンスの強さ v_i を推定することは，観測 $y_{i\to j}$ および入力情報 t_{ij} を用いて，観測 $y_{i\to j}$ を式 3.9 の $p_{i\to j}$ の観測とみなした際の最も尤度の高い v_i を統計的に求めることと定式化される．

3.2.4　未知母数の推定

3.2.3 節にて述べた確率モデルに基づき，観測を用いて未知母数のベイズ推定を行い，各人の直接インフルエンスの強さ v_i を求める．この時に，確率モデルから代数的に解を得る困難さを回避するため，マルコフ連鎖モンテカルロ（MCMC）法 [79] を用いる．

MCMC 法では，式 3.12 に示すベイズの定理を用いて各未知母数に対し，確率モデルに従い事後分布 ($P(\Omega|x)$) と一致する分布特性を持つ擬似乱数列を生成する．式中で Ω は未知母数を，x は観測を示す．生成された擬似乱数列を大数の法則に基づき解析することで，未知母数の値（分布）を推定することができる．擬似乱数列生成のためには，式 3.12 の右辺の各項のうち，$P(x|\Omega)$ と $P(\Omega)$ を与える必要がある．$P(x|\Omega)$ は尤度関数であり確率モデルそのものである．$P(\Omega)$ は未知母数の事前分布である．なお，分母の $P(x)$ は観測のみで与えられる規格化定数である．

$$P(\Omega|x) = \frac{P(x|\Omega)P(\Omega)}{P(x)} \tag{3.12}$$

3.2.3 節の確率モデルについて母数推定する場合，推定対象の未知母数は v_i, $\gamma_{q_{i\to *}}$, $\gamma_{q_{*\to j}}$, $\gamma_{q_{*\to *}}$, $\theta_{q_{i\to *}}$, $\theta_{q_{*\to j}}$, $\theta_{q_{*\to *}}$ である．また入力として与える情報は，利用履歴から得られる観測値 $y_{i\to j}$ と友人関係情報 t_{ij} および各未知母数の事前分布である．具体的に，MCMC 法向けの記述言語である BUGS 言語 [65] を用いて，3.2.3 節で述べた確率モデルおよび未知母数の事前分布を記述した例を次に示す．例では添え字を展開するための構文等一部を省略している．母数推定は次のような手順で行う．モデル定義，未知母数の事前分布に加えて観測値等の必要な入

力情報を与え，未知母数の初期値を乱数により生成させる．その後，MCMC サンプリングによりマルコフ連鎖を生成し未知母数の一つである直接的インフルエンス v_i の擬似乱数列を生成する．得られた擬似乱数列より初期値の影響を避けるため先頭から burn-in 標本数分を棄却し，それ以降の擬似乱数列の平均値（期待値）を v_i の推定値とする．

```
model{
  # 観測される確率変数
  p[i,j] ~ dbeta( alpha_p[i,j], beta_p[i,j] )
  qi[i]  ~ dbeta( alpha_qi[i], beta_qi[i] ) # t_ij=0 友人以外のみ
  qj[j]  ~ dbeta( alpha_qj[j], beta_qj[j] ) # t_ij=0 友人以外のみ
  qw     ~ dbeta( alpha_qw, beta_qw )       # t_ij=0 友人以外のみ

  # 未知母数とその事前分布
  v[i] ~ dnorm(0.0, 0.1)
  gamma_qi[i] ~ dbeta(1.0, 19.0)
  gamma_qj[j] ~ dbeta(1.0, 19.0)
  gamma_qw    ~ dbeta(1.0, 19.0)
  theta_qi[i] ~ dgamma(1.0, 0.02)
  theta_qj[j] ~ dgamma(1.0, 0.02)
  theta_qw    ~ dgamma(1.0, 0.02)

  # 確率モデル
  alpha_p[i,j] <-    gamma_p[i,j]  * theta_p[i,j]
  beta_p[i,j]  <- (1-gamma_p[i,j]) * theta_p[i,j]

  gamma_p[i,j] <- gamma_q[i,j]+v[i]*t[i,j]
  theta_p[i,j] <- theta_q[i,j]

  gamma_q[i,j] <- alpha_q[i,j]/(alpha_q[i,j]+beta_q[i,j])
  theta_q[i,j] <- alpha_q[i,j]+beta_q[i,j]
  alpha_q[i,j] <- alpha_qi[i]+alpha_qj[j]-alpha_qw
  beta_q[i,j]  <- beta_qi[i] +beta_qj[j] -beta_qw

  alpha_qi[i]  <-    gamma_qi[i]  * theta_qi[i]
  alpha_qj[j]  <-    gamma_qj[j]  * theta_qj[j]
  alpha_qw     <-    gamma_qw     * theta_qw
  beta_qi[i]   <- (1-gamma_qi[i]) * theta_qi[i]
  beta_qj[j]   <- (1-gamma_qj[j]) * theta_qj[j]
  beta_qw      <- (1-gamma_qw)    * theta_qw
}
```

3.3　利用するデータ

3.3.1　利用履歴

2.3.1 節にて述べた，第 2 章の検討において学習セットとして用いたデータセットを本検討においても用いる．すなわち，モニター実験にて収集したスマートフォン利用履歴から，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のみを抽出し，2011 年 2 月 1 日から 2011 年 4 月 30 日までに限定し，利用者数が 3 未満のアプリケーションを削除した結果の，3,383 レコード (155 利用者，291 アプリケーション) を利用した．（図 3.1）

3.3.2　友人関係情報

端末貸与終了時に実験協力者に対してアンケート調査を行い，実際の友人関係を調査した．前述のとおり実験協力者はグループで応募する形態としたため，得られた友人関係情報は，いくつかの集団から構成されるデータセットとなっている．

友人関係は具体的に次の方法により得た．友人であるかないかの曖昧性を排除するために，アンケートの中に「相手と話したことがあるか」という設問を設け，相手としては回答者が所属する応募グループのメンバリストを示した．グループ内の各人を相手とした場合の回答を各々求め，この結果を友人関係として用いた．ある実験協力者から相手として指定されているにも関わらず，その実験協力者を相手として回答しない場合が生じたが，その 2 人の間には友人関係があるものとしてデータを補正して使用した．

一人あたりの友人数のヒストグラムを図 3.2 に示す．図 3.2 には，3.3.3 節で説明する各インフルエンサの分布を $*$ にて表示している．前述の通り，実験協力者はいくつかの集団から構成されており，集団内は友人関係が密であることから，分布はべき乗法則を満たしていない．また，インフルエンサと友人数には，顕著な相関は認められない．

図 3.3 は，例として 50 人のグループで応募した集団の友人関係情報をグラフとして図示したものである．図中の各ノードは利用者を表し，各エッジは友人関係を示している．このグループの内訳は 3 系統の友達つながりと申告されており，この系統が図によく表れている．

3.3.3　インフルエンサ

実験協力者に対する追加アンケートを 2011 年 11 月に行い，実際のインフルエンサを調査した．調査は，他の誰かから紹介されたことがきっかけでダウンロードして使ってみたアプリ

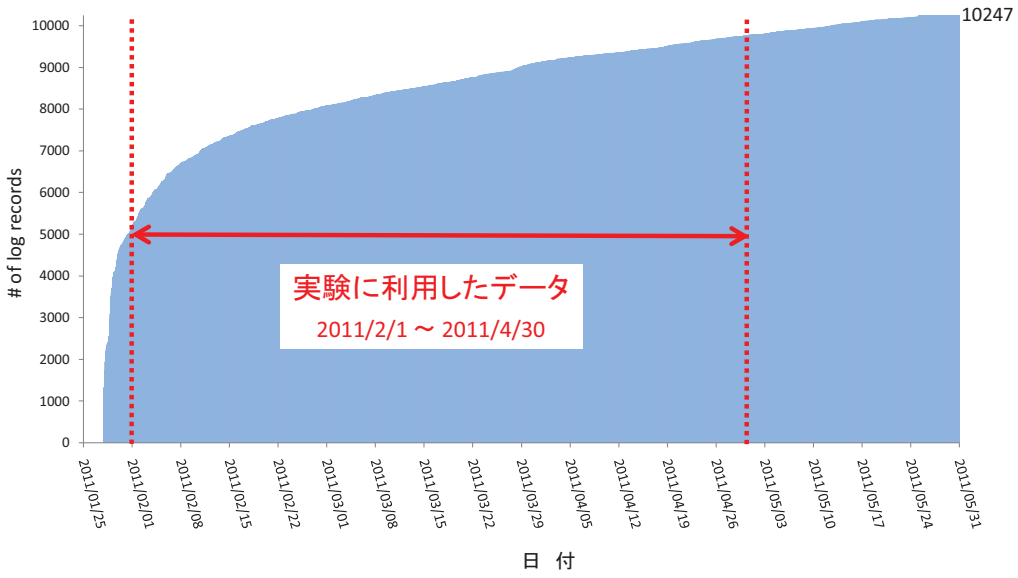


図 3.1　履歴レコードの累積数

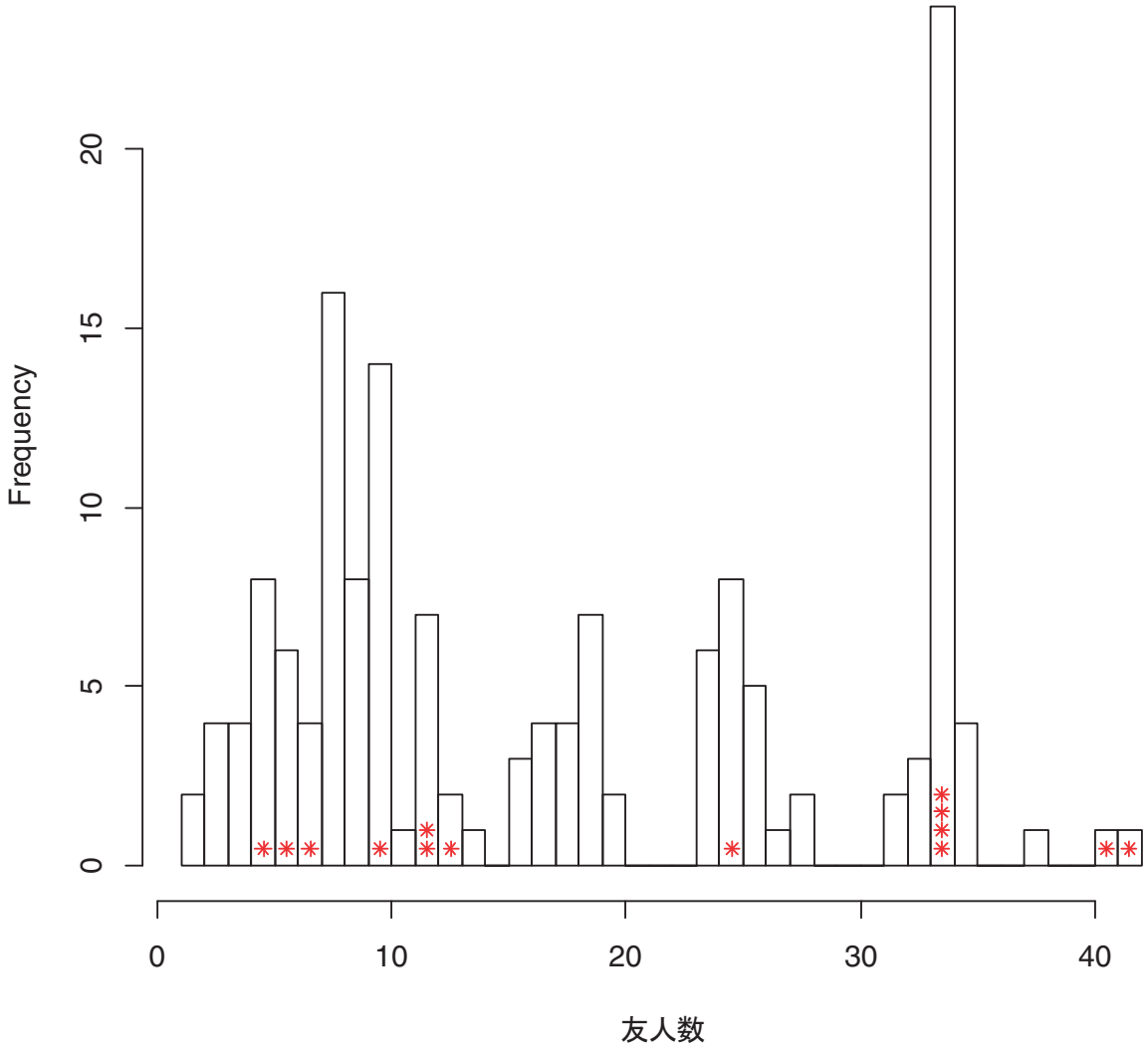


図 3.2　友人数ヒストグラム

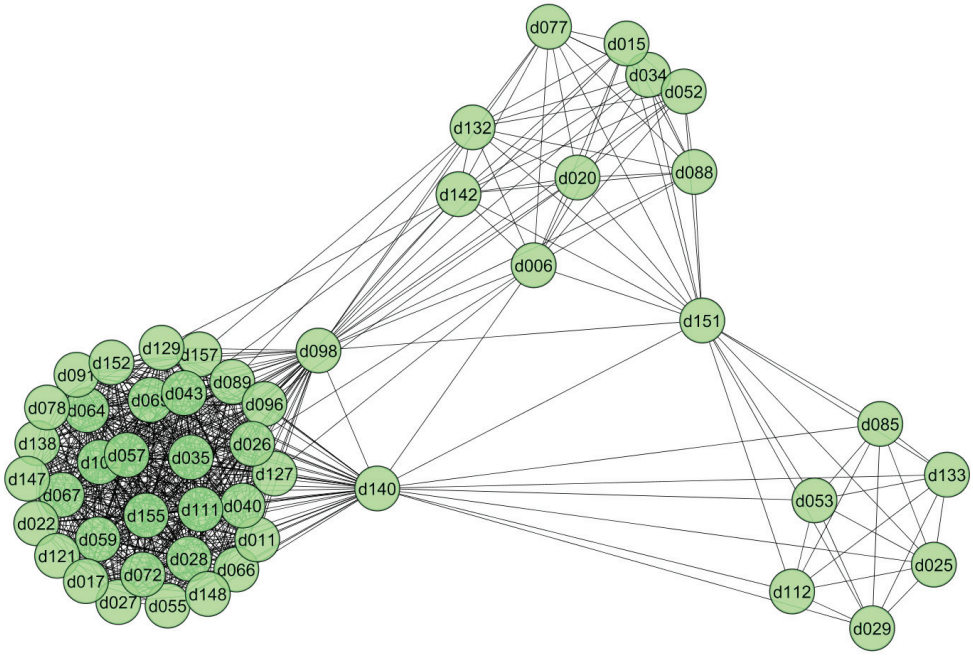


図 3.3　友人関係情報の一例

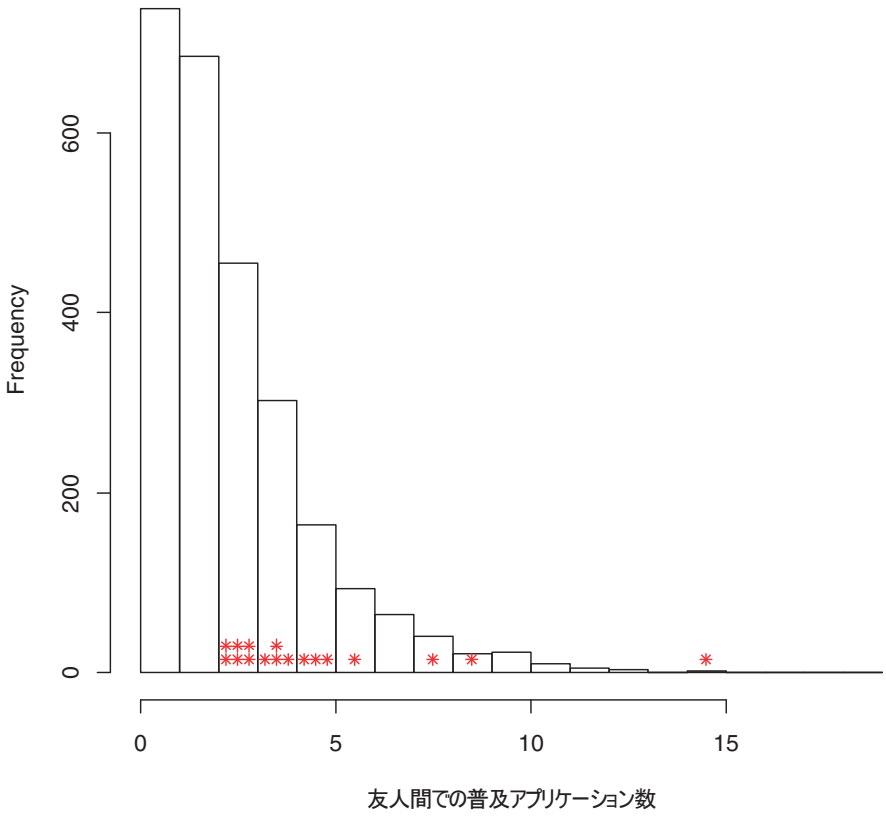


図 3.4　友人間における普及アプリケーション数のヒストグラム

ケーションがあった場合，誰からの紹介であったかを直接尋ねる形で回答を得た．回収できた有効回答は 44 人分（回収率 28%）であり，他の実験協力者からの紹介によりアプリケーションをダウンロードしたことがあったのは 24 人，回答から得られたアプリケーションの紹介者は 18 人であった．この中からアプリケーションの利用履歴を確認し，回答者と紹介者の間に共通のアプリケーションが 2 本以上確認できた紹介者（14 人）を，インフルエンサの正解データとして用いた．

友人関係のある利用者 i, j 間で $i \to j$ に普及が観測されたアプリケーションの数（$|||\{a \in A_{i \to j}\ ;\ t_{ij} = 1\}|||$）を集計したヒストグラムを図 3.4 に示す．図中 $*$ はインフルエンサの正解データとしたアンケートの（紹介者 $\to$ 回答者）の分布を示したものである．インフルエンサであっても普及アプリケーションの数が多い傾向は認められない．直接的インフルエンスによるダウンロードは，これらの一部もしくは全部であり，頻度は大きくないことに注意が必要である．ここでは，アンケート回答の信頼度は高いと考え，紹介者からの回答者への直接的インフルエンスが観測中に存在することを前提とした．また，インフルエンサ調査アンケートへの未回答者が相当数存在するため，正解データはインフルエンサの一部であり他にもインフルエンサは存在する可能性が高いことに注意する．

3.4　実験と考察

3.4.1　利用者の先行性

まず，利用者の先行性について確認する．ここでは，利用者 i の先行性の指標として s_i を式 3.13 により定義する．ただし $r_{a,i}$ は，利用者 i がアプリケーション a をダウンロードした実験協力者中の順位を，u_a はアプリケーション a の実験協力者中の利用者数を，$\|A_i\|$ は A_i に含まれるアプリケーションの数を示す．

$$s_i = \frac{\sum_{a \in A_i}(1 - \frac{r_{a,i}}{u_a+1})}{\|A_i\|} \tag{3.13}$$

利用履歴を用いて各利用者の先行性を算出した結果のヒストグラムを図 3.5 に示す．これにより，アプリケーションダウンロードにおける先行性分布も，Rogers が文献 [60] にて示した採用者分布曲線と符合することが確認できる．図 3.5 には，アンケートにより明らかになったインフルエンサの位置を $*$ にて示している．先行性が高いこととインフルエンサであることが必ずしも一致しないことが分かる．

3.4.2　インフルエンサの推定

3.2.3 節で述べたモデルに基づき, 3.3 節で説明したデータを用いて, 3.2.4 節で示した MCMC 法により，各利用者の周囲への直接的インフルエンスの強さ v_i を推定した．表 3.2 に実行した MCMC 法による推定の詳細条件を示す．未知母数の事前分布は，予備実験を行い，各変数の取りうる範囲を観察し，その結果を勘案し無情報事前分布となるように設定した．事前分布の影響については 3.4.4 節にて議論する．MCMC 法による母数推定には WinBUGS[45] を用い，2.93GHz の Intel® Xeon® CPU のワークステーションにて表 3.2 に示した規模の標本生成に 20 時間程度を要した．

図 3.6 に，MCMC 法により推定された直接的インフルエンスの強さ v_i のヒストグラムを示す．図 3.5 と同様に，インフルエンサの位置を $*$ にて示した．図 3.5 の先行性 s_i の場合と比較して，インフルエンサは v_i の値が大きい部分に分布することが観察され，より良く推定できていることが認められるとともに，仮説として設定した確率モデルについて一定の妥当性が確認できる．

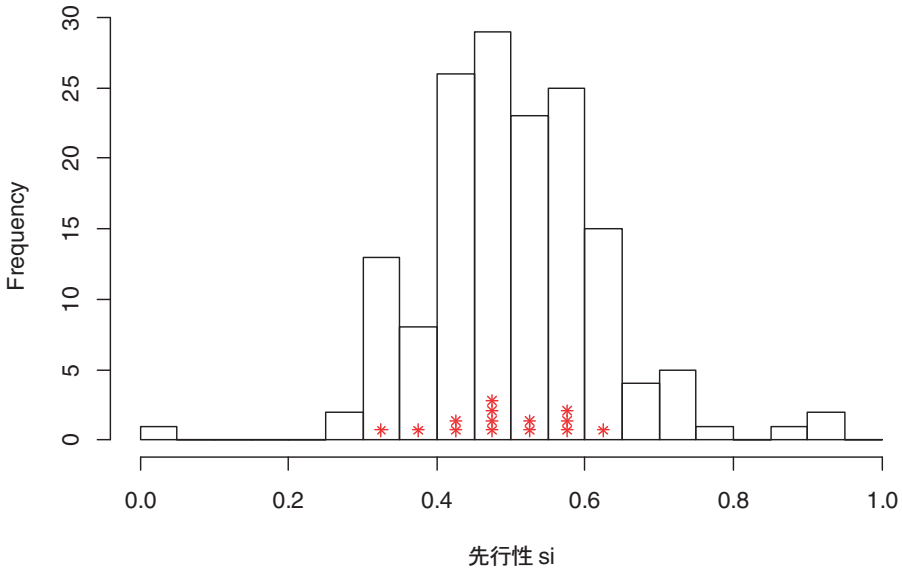


図 3.5　利用者の先行性 s_i のヒストグラム

表 3.2　MCMC 法による母数推定の詳細設定

項　目	値
生成マルコフ連鎖数	3
1 連鎖あたりの総生成標本数	12,000
1 連鎖あたりの burn-in 標本数	6,000
未知母数の事前分布	$v_i \sim N(0, 0.1)$ $\gamma_{q_{i\to *}} \sim Beta(1.0, 19.0)$ $\gamma_{q_{*\to j}} \sim Beta(1.0, 19.0)$ $\gamma_{q_{*\to *}} \sim Beta(1.0, 19.0)$ $\theta_{q_{i\to *}} \sim G(1.0, 0.02)$ $\theta_{q_{*\to j}} \sim G(1.0, 0.02)$ $\theta_{q_{*\to *}} \sim G(1.0, 0.02)$

3.4.3　他手法との比較

本節では，提案手法のインフルエンサ推定能力を，いくつかの既存手法と比較し実験的に評価する．Goyal らは文献 [24] において，個人間の普及のうち友人間のみを取り出し，その時間差が基準より小さいものを直接的インフルエンスと見なす手法を述べるとともに，*influenceability*（影響の受けやすさ）として各利用者の実行アイテムのうち周囲の友人からの直接的インフルエンスによるとみなされるアイテムの比率を提案している．ここでは，この *influenceability* の考え方を応用し，各利用者の実行アイテムのうち友人に対して直接的インフルエンスを与えた比率を *influentiality*（影響の与えやすさ）として次式により算出し比較対象とする．ただし $t_{a,i}$ はアプリケーション a を利用者 i がダウンロードした時刻を表す．

$$influentiality_i = \frac{\| \bigcup_{\{j\ ;\ t_{ij}=1\}} \{a \in A_{i \to j}\ ;\ 0 \leq t_{a,j} - t_{a,i} \leq \tau_{ij}\} \|}{\|A_i\|} \tag{3.14}$$

$$\tau_{ij} = \frac{\sum_{a \in A_{i \to j}} (t_{a,j} - t_{a,i})}{\|A_{i \to j}\|} \tag{3.15}$$

さらに，Goyal らが提案している代表的なモデルである (1) 時間差を考慮しないモデル（Static Model: ST と表記），(2) 時間差を考慮するモデル（Discrete Time Model: DT と表記），を取り上げ，各々について (a) ベルヌーイ試行ベース (B と表記)，(b) Jaccard 係数ベース（J と表記），を実装して比較する．Goyal らのモデルは，個々の友人に対する直接的インフルエンス確率を個別に与える．しかし，インフルエンサを推定する上で，特定の友人に強い直接的インフルエンスを及ぼすことが重要であるのか，あるいは友人全体に平均的に直接的インフルエンスを及ぼすことが重要であるのか明らかでないため，ここではその最大値（MAX と表記）および平均値（MEAN と表記）を算出し，それぞれを直接的インフルエンスを表す指標として扱う．

インフルエンサ推定能力の比較として，提案手法により推定された直接的インフルエンスの強さ v_i，先行性 s_i，$influentiality_i$，Goyal らのモデルによる直接的インフルエンスを表す指標について，上位 k 番目までに正解インフルエンサが被覆される率の推移（recall@k）を算出した．得られた結果を図 3.7，3.8 に示す．提案手法については異なる乱数系列を用いて 3 回推定を行ない得られた被覆率の平均を示している．Goyal らのモデルは，図 3.8 の凡例において上述の表記を組み合わせて表している．

図 3.7 より，$influentiality_i$ は先行性とほぼ同様の特性を示し，直接的インフルエンスをうまく捉えられていないことがわかる．また，Goyal らのモデルは，最大値を用いる場合より平均値を用いた場合の方が良い被覆率を示す傾向が見られるが，それ以外には顕著なモデル間の差は認められない．いずれのモデルに対しても，提案手法がより良い性能を示しており，その優

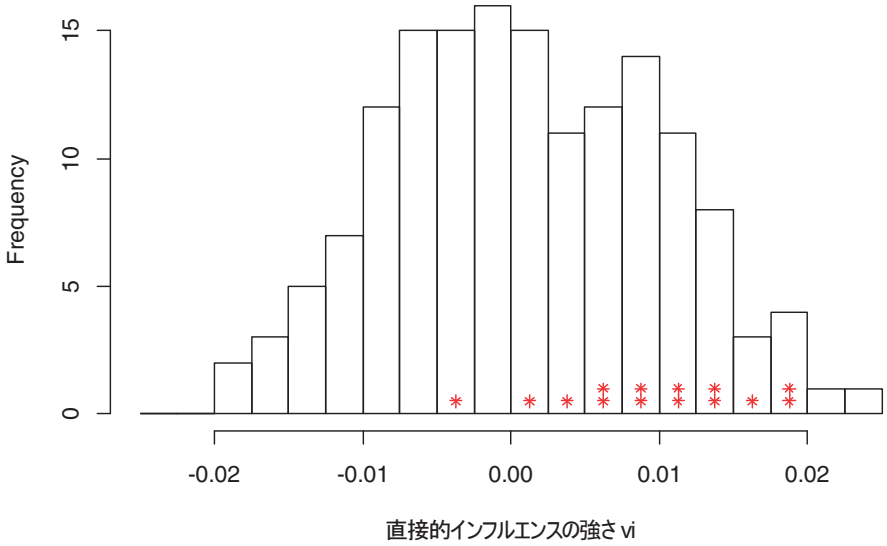


図 3.6 直接的インフルエンスの強さ v_i 推定結果ヒストグラム

表 3.3 事前分布パラメータの設定

設 定	v_i	$\gamma_{q_{i\to*}}, \gamma_{q_{*\to j}}, \gamma_{q_{*\to*}}$	$\theta_{q_{i\to*}}, \theta_{q_{*\to j}}, \theta_{q_{*\to*}}$
表 3.2 の設定	$N(0, 0.1)$	$Beta(1.0, 19.0)$	$G(1.0, 0.02)$
設定 1	$N(0, 0.01)$	$Beta(1.0, 19.0)$	$G(1.0, 0.02)$
設定 2	$N(0, 0.05)$	$Beta(1.0, 19.0)$	$G(1.0, 0.02)$
設定 3	$N(0, 0.2)$	$Beta(1.0, 19.0)$	$G(1.0, 0.02)$
設定 4	$N(0, 0.1)$	$Beta(0.5, 9.5)$	$G(1.0, 0.02)$
設定 5	$N(0, 0.1)$	$Beta(2.0, 38.0)$	$G(1.0, 0.02)$
設定 6	$N(0, 0.1)$	$Beta(1.0, 19.0)$	$G(1.0, 0.01)$
設定 7	$N(0, 0.1)$	$Beta(1.0, 19.0)$	$G(1.0, 0.04)$

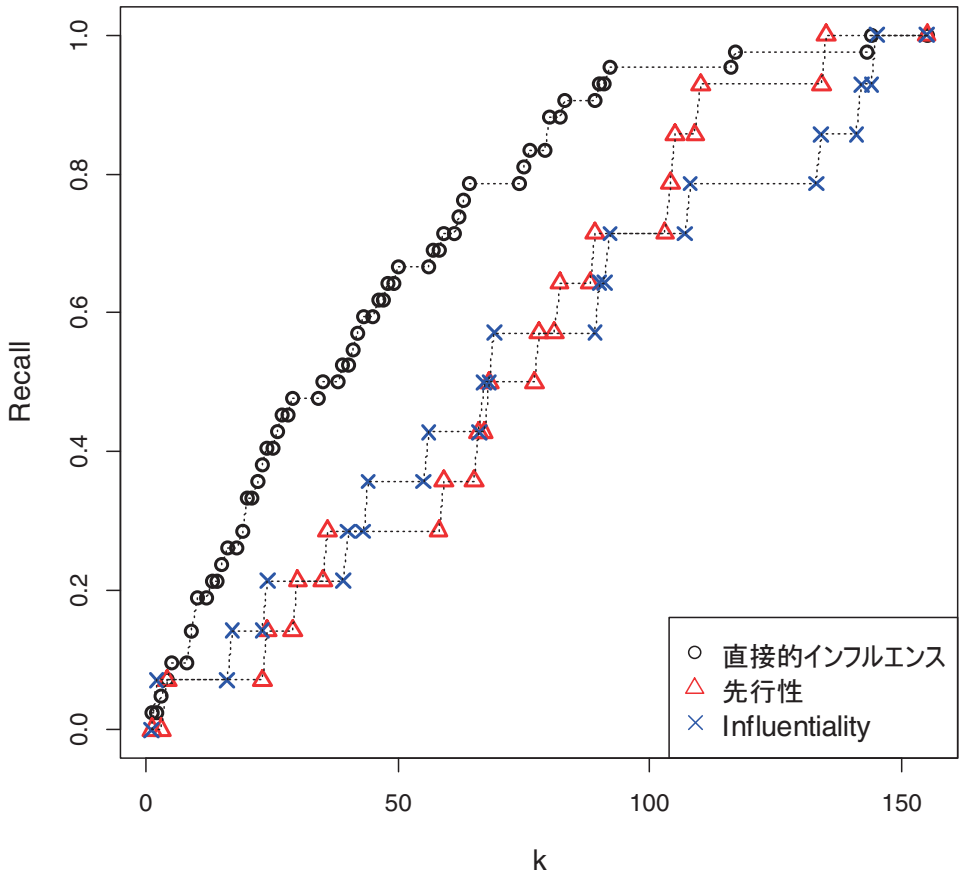


図 3.7　先行性，Influentiality との正解被覆率の比較

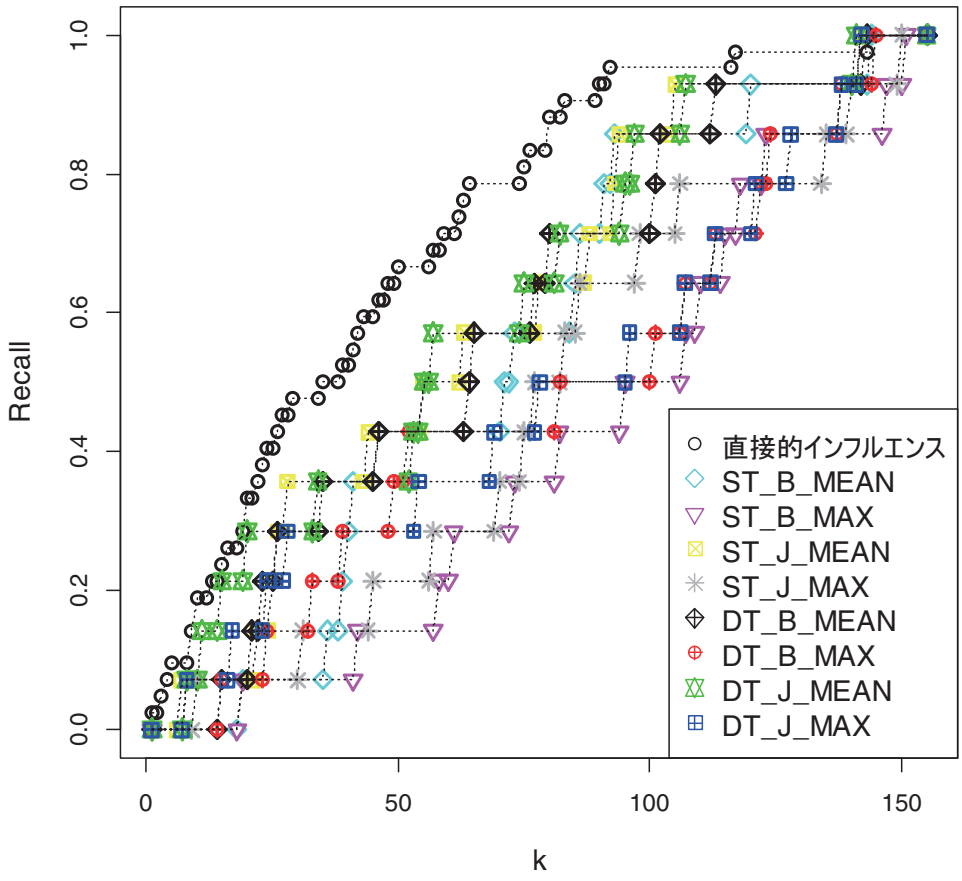


図 3.8　Goyal らのインフルエンス確率モデルとの正解被覆率の比較

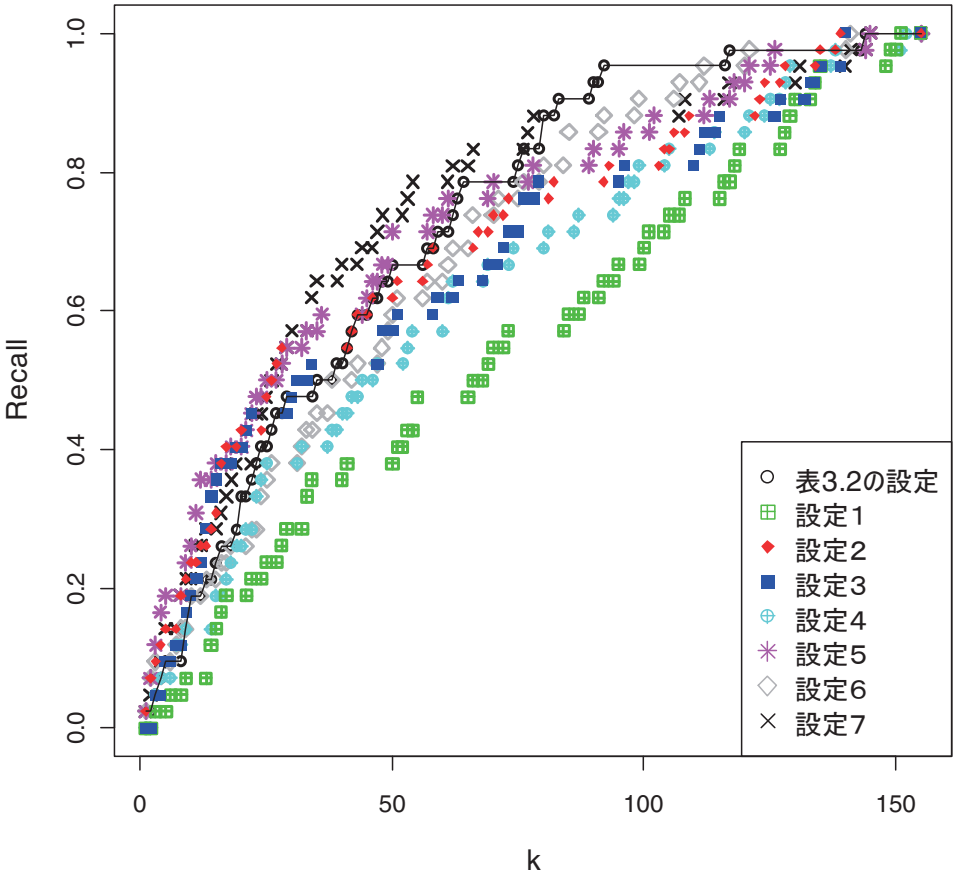


図 3.9　事前分布パラメータの影響

位性が確認できる．ここで比較したモデルはいずれも普及の発生確率を基本として，時刻差の大小および友人関係の有無を用いて直接的インフルエンス候補を選別し，その強さを推定しているため，これらの方法では混在する直接的インフルエンスと拡散的普及を十分に区別できない．これに対して提案手法では，直接的インフルエンスと拡散的普及を分離してモデル化しており，このパラメータを有効に推定できたことが示唆された．

しかしながら提案手法においても，一部のインフルエンサについては ν_i の値が十分大きい値として推定されなかった．原因として，ν_i の値を相手 j によらず一定としてモデル化しているため，友人関係のある相手の中でもごく少数の相手にのみ影響を及ぼしていて，他の相手にはほとんど影響を及ぼさないような場合，結果的に ν_i が大きく推定されなかったことが想定される．これに対し，仮に ν_i を相手 j 毎に異なる値を持つように ν_{ij} としてモデルを拡張すると，未知母数の数が飛躍的に増加してしまい，母数推定が困難となる．これは提案モデルの限界であり，異なるアプローチによる解決が必要である．また，友人関係のある相手が特定の嗜好に偏っているような場合，3.1 章で触れたホモフィリーによる効果が顕在化し，拡散的普及と友人における普及の乖離が大きくなり，直接的インフルエンスは強くないにも関わらず ν_i が大きく推定されてしまう場合も想定される．

3.4.4 事前分布の影響

提案手法では，MCMC 法による母数推定を行う上で，表 3.2 に示すように，未知母数の事前分布をパラメータとして設定する必要がある．事前知識がある場合はそれを事前分布として表現し，事前知識がない場合は事前分布が影響を及ぼさないように無情報となるように設定する（分散に十分大きい値を設定するなど）ことが，MCMC 法においては一般的である [79]. 今回は事前に複数の仮の設定にて推定を行い，その結果を勘案し事前分布が無情報となるようにパラメータを設定した．例として，図 3.6 の ν_i の推定結果ヒストグラムを参照すれば，利用した事前分布である平均 0，分散 0.1 の正規分布は，十分大きな分散を持つ分布であることが確認できる．本節では事前分布の設定により結果がどの程度影響を受けるかを実験的に明らかにし，提案手法の頑健性の一面を評価する．表 3.2 に示した事前分布パラメータを基準として，ν_i の分散，拡散的普及の期待値 γ（$\gamma_{q_{i\to*}}, \gamma_{q_{*\to j}}, \gamma_{q_{*\to*}}$）の事前分布（ベータ分布）の先鋭度，拡散的普及の先鋭度 θ（$\theta_{q_{i\to*}}, \theta_{q_{*\to j}}, \theta_{q_{*\to*}}$）の事前分布（ガンマ分布）のパラメータ β を，表 3.3 に示す様に各々変化させ，MCMC 法を用いてインフルエンサを推定して得られた正解被覆率を図 3.9 に示す．ν_i の分散を 0.01（表 3.2 の値の 1/10）にした場合（設定 1）は，著しく被覆率が悪化した．分散を 0.01 に設定すると，先に見た図 3.6 の分布を勘案すれば，事前分布がもはや無情報と言えない程度に分散が小さくなり，この影響により被覆率が悪化したものと考えられる．これに対して ν_i の分散が 0.05 以上ではほぼ無情報となり，設定 2 および設定 3 では被覆率に大きな差が表れていない．その他のパラメータについても一定の影響が観察される

が，提案手法の優位性を損なうようなものではなかった．以上より，事前分布パラメータは適切に設定する必要があるが，本論文で扱った問題設定においては，表 3.2 に示した事前分布パラメータは適切な設定であることが確認できる．

3.5　むすび

本章では，スマートフォンのアプリケーション利用に関して，周囲に影響を与えるインフルエンサを友人関係情報を利用して，利用履歴のみから推定する手法を提案した．直接的インフルエンスと普及の関係を考察し，個人間の影響度を直接周囲に与えられた影響による分とそれ以外の要因による分の組み合わせで表し，利用者間のアプリケーションダウンロードの連鎖の生起はこれをパラメータとするベータ分布に従う確率モデルとして表した．このモデルに基づき MCMC 法により直接的インフルエンスの強さを推定する手法を提案した．提案手法の可用性を確認するために，約 160 人の大学生による実験で得た 3 か月分の利用履歴データを用いて，直接的インフルエンスの強さを推定し，アンケートによって得た実際のインフルエンサと照合して結果を評価した．他の手法と比較して明らかに良好な被覆率が得られ，提案手法の可用性と効果が確認された．また，現実のマーケティングタスクに対して提案手法を用いることにより有意な効果が得られるか否かは，実施策に適用し評価する必要があり，今後の課題である．

なお提案手法には下記の課題・限界があり，今後の研究の方向性を示している．

- 直接的インフルエンスの強さ v_i を相手によらず一定としているため，相手による違いをモデル化できない
- 時間的変動を考慮していない
- トピックによる普及および直接的インフルエンスの変化を考慮していない
- ホモフィリー効果による影響を考慮していない
- 大規模な設定には計算量的に対応できない

さらに，提案手法では友人関係情報を入力として用いたが，完全な友人関係情報を得ることは現実的には容易でない．第 4 章では，一部の友人関係のみが観測されると仮定して，それ以外の友人関係の有無について推定する手法を提案する．

第 4 章

友人関係の推定

4.1 はじめに

第 3 章では，友人関係情報を用いてインフルエンサを推定する手法について述べた．しかし友人関係情報は容易に得られるわけではない．近年において Facebook®, Twitter®, MySpace®[51] などのソーシャルネットワーキングサービス（SNS）が普及し，関連するデータを分析することでソーシャルネットワーク（友人関係情報）が得られバイラルマーケティング等への応用が期待できるため，ソーシャルデータの分析に注目が集まってきている．

しかしながら，ある市場調査 [32] によれば，バイラルマーケティングで想定しているようなクチコミ効果は，現時点においては SNS におけるようなサイバーな人間関係よりも，大半は実世界における対面の人間関係において起こっていることが示されている．このことは，SNS から得られる友人関係だけでは，目的に対して不十分であることを示唆している．また，通話履歴や電子メールの送受履歴を用いれば，所望の友人関係を得るうえで効果的であると考えられるが，現実にはプライバシーの問題からこれらを大規模に収集し用いることは困難である．さらには，これらを収集したとしても，依然友人関係全体の一部しか反映されない．図 4.1 は，著者らが収集した実際の友人関係ネットワークの実例である（収集の詳細については後述する）．図中の太い実線は実験で収集した 6 ヶ月間の通話履歴から確認された友人関係を，破線はアンケート調査により確認された友人関係（正確には実験期間中に電話したと回答した相手）を示している．この図から，通話に限っても様々なチャネルが使われていることを読み取ることができる．したがって現実的にはコミュニケーション全体のなかの一部分しか観測できないことが想定されるとともに，比較的少数の既知の友人関係のみを用いて未知の友人関係を推定することの重要性が確認できる．このため，本章ではスマートフォンのアプリケーション実行履歴と Web 閲覧履歴を用いて，一部の既知の友人関係を教師情報とした半教師付学習により友人関係の推定を行う手法を提案する [26]．提案手法は，友人同士は共通の興味を共有しており（ホモフィリー [46]），またスマートフォンの利用にもこれが反映されていることを仮

定している．

一部の既知の友人関係を手がかりに友人関係全体を推定するためには，人をノードで表し，友人関係の有無をノード間のリンクの有無で表現するグラフ（ソーシャルグラフ）を想定して，このグラフに対してリンク推定（link prediction）[19] の手法を用いることが考えられる．リンク推定は，ノード間のリンクの有無を推定する手法で，様々な研究がなされている [44]．しかしながら既存のリンク推定手法を用いて友人関係の推定を行う場合には大きく次の課題があり，既存手法では十分な結果を得ることができない．

One-class 設定　すべての行動を完全に監視・把握することは事実上困難であるため，特に関わりが観察されない相手に対しても友人である可能性を否定することができない．つまり観察される事象は，友人関係のあることを示す正例のみであり，負例を学習することができない．

友人関係のスパース性　現実的に，社会においては全体の人数に比して友人である人の数は著しく少ない．One-class 設定であることも関連し，リンクの存在を示す正例の数は少なく，大半のリンクはその有無が未知となり，推定対象となる．

友人関係の有無を示唆する指標が不明　リンク推定では，各ノードに属性等の関連情報が付与されており，対象ノードの関連情報を用いることで，ノード間にリンクが存在しそうな程度を示す指標が利用できると仮定し，これを手がかりに推定することが多い．しかしながら現実に得られる利用者情報（例えばデモグラフィック属性，各種履歴，等）から直接的に信頼度高く友人関係の有無を示唆する指標を設定することは困難である．

既存のリンク推定手法は，利用する情報によって次の 2 種類に分類できる：(1) ネットワーク構造情報（既知のリンクの有無）のみを利用，(2) ノード情報も利用．ここでノード情報とはノードに関する属性等の関連情報であり，ノード間にリンクが存在しそうな程度を示す指標（以降，affinity 指標と呼ぶ）が導出できることを期待している情報である．(1) は対象のノード組に対する共通の隣接ノードやノード間のパスなどのネットワーク構造のみを用いる手法で，Jaccard 類似度 [44] および Adamic/Adar [1] が代表的である．これらの手法は観察されたネットワークがスパースである場合，手がかりが著しく少なくなり，推定が困難となる．一方，(2) の手法はネットワーク構造のほかにノード情報も利用する．この場合，リンクの有無は不明であってもノード情報はすべてのノードにおいて既知であることを前提とする．これによりネットワーク構造情報では一つもリンクのないノードであっても，そのノードに関連するリンクの推定が可能である．先に述べた通り，観察されているリンクの数は友人推定の場合全体からみると少ないので，ここでは (2) のアプローチを用いる．(2) のアプローチの手法としては，以下の 2 手法が最近提案された．

潜在特徴モデル：　Menon らは LFL と呼ぶ潜在特徴モデルを適用した半教師付学習によりリ

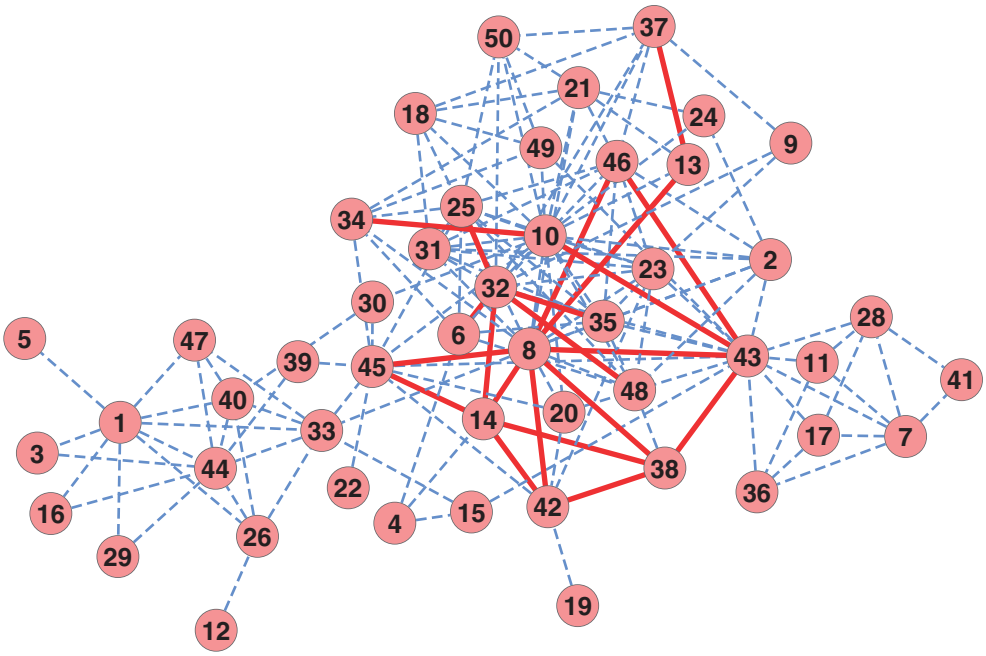


図 4.1　友人関係ネットワークの一例

ンク推定を行う手法を提案した [47, 48]．この手法は，潜在特徴から得るリンク推定結果とノード情報から得られる affinity 指標を線形結合で統合したものによりリンク推定を行うこととして，事前に判明しているリンクの有無との間のロスを最小化するように潜在特徴を調整することで最適化・学習を試みる．この時，事前に有無が分かっているリンクに関連しないノードについては学習に関与しないことに注意する．また，Yang らは FIP（Joint Friendship and Interest Propagation）モデルとして利用者–利用者と利用者–アイテムの関係における潜在特徴を組み合わせて用いる手法を提案した [72].

リンク伝搬：　Kashima らはリンク伝搬（Link Propagation）手法を提案した．この手法は，事前に判明しているリンクを，ノード情報を手がかりに事前に定義したカーネルを用いて伝搬させることによってリンク推定を行う [27]．利用できるノード情報が，高次元データでスパースであるような場合には，適切なカーネルを事前に定義することは必ずしも容易でない.

リンク推定に関連する研究以外にも，友人関係の推定を取り上げている研究が報告されている．これらは用いるデータの内容から友人関係を推定するものであり，次のデータが用いられている．

1. GPS 測位情報等の位置情報： [68, 13, 63],
2. Bluetooth 近接デバイス： [57, 13],
3. 通話履歴： [68, 49].

位置情報による軌跡や近接デバイスデータは友人関係を強く反映すると考えられるが，物理的にコンタクトする友人関係のみが獲得される．また通話履歴は情報の性格からプライバシー情報としての認識が強く，また通話相手の個人情報にも該当するため利用許諾を得ることが容易で無く，この目的に利用することは現実的に困難である．

本検討においては，one-class 設定であることや友人関係ネットワークがスパースであることに対して，豊富な利用者情報をノード情報として活用することによってより良い友人関係の推定を実現することを目指す．ここで，豊富な利用者情報として，スマートフォンのアプリケーション実行や Web 閲覧の履歴を利用する．アプリケーション実行や Web 閲覧は，スマートフォン利用者にとって共通かつ広範な活動であり，これらの履歴が大半の利用者に対して利用可能であることを想定することができる．また標準的オペレーティングシステムにはこれらの履歴を保存する機能が備わっているため，広く展開することを考えても現実的かつ妥当である．さらには，最近の研究により Web 閲覧履歴から利用者の興味が抽出・獲得できる [16] ことや，Web 閲覧履歴を用いると推薦の性能向上が期待できる [40] ことが報告されており，アプリケーション実行および Web 閲覧の履歴には十分な利用者に関する情報が含まれていると考えられる．ただし，利用者の興味等は履歴から直接観測できるものではなく，履歴から目的

に応じて統計的機械学習により抽出する必要があることには注意が必要である．

Yang らは同様のアプローチから FIP モデルを提案した [72]．FIP モデルは，利用者–アイテムの関係を，利用者–利用者のモデリングに統合している．しかしながら，Yang らはアイテムが直接興味を表すキーポイントであると仮定し，得られるノード（利用者とアイテム）に関する情報から潜在特徴を獲得することに注力しており，利用者–アイテムの関係から潜在特徴を得ることをしていない．これは FIP モデルがアイテムの推薦をゴールとして設定しており，友人関係推定を目的としていないためである．

また，Ozaki らの研究 [52] では，利用者情報（アプリケーション実行や Web 閲覧の履歴）を用いて，情報拡散モデルに基づき時間差や利用頻度を勘案した暗黙の影響構造を推定することから友人関係を推定することを試みている．しかし，第 3 章にて述べた通り，暗黙の影響関係には先行性による要因が含まれていることから，影響関係があることは友人であることの十分条件にならず，影響関係と友人関係が必ずしも一致しない．

本検討では，利用者のアプリケーション実行および Web アクセスの履歴に対して，行列分解手法を適用することにより潜在利用者特徴を獲得し，それを友人推定に活用する手法を提案する．行列分解手法は協調フィルタリングにおいて利用者特徴およびアイテム特徴を獲得する上で有効である [37] ことが示されており，友人推定のための利用者–アイテム関係からの潜在特徴抽出にも有望と考えられる．提案手法では，半教師付リンク推定手法を拡張して行列分解を統合し，行列分解とリンク推定を同時に最適化する．さらには，友人関係をよりよくモデル化するために，友人関係と興味マッチングの関係仮説に基づいた新しい affinity 指標を提案する．すなわち提案手法はこれらを統合し，潜在特徴（すべての利用者–アイテム観測を行列分解することにより得る）とリンク推定（既知の友人関係を用いた半教師付学習）を，提案する affinity 指標に基づいて同時に学習・最適化することにより，友人関係の推定を行う．

本検討における主要な貢献は次の通りである．

- 既存の学習リンク推定手法に活動履歴の行列分解による潜在利用者特徴モデルを統合した友人関係推定手法を提案した．事前知識に基づく affinity 指標をあわせて提案した．
- 提案手法の有効性を実際に 50 人から収集したモニターデータにより確認した．

以降，本章の構成は次に示す通りである．4.2 節にて問題の定式化を行う．4.3 節では提案する友人推定手法について述べる．4.4 節にて収集したデータおよび評価結果について説明し，4.5 節で本章をまとめる．

4.2　定式化

4.2.1　半教師付リンク推定

既知のノード情報 $\mathbf{X}$ を用いた半教師付リンク推定は，以下の目的関数を最小化する未知のパラメータ θ を求めることと定式化される．

$$\min_{\theta}\left\{\sum_{(i,j)\in O}\ell(G_{ij},\hat{G}_{ij}(\mathbf{X},\theta))+\lambda\Omega(\theta)\right\} \tag{4.1}$$

ここで，O は既知のリンクの集合を示す．G は $n\times n$ のグラフ構造を示す隣接行列であり，各要素は 0，1，? のいずれかの値を持つ．これらはそれぞれ，リンク無し（既知），リンク有り（既知），リンクの有無は未知を意味する．ここで n はノード数を示す．また，one-class 設定においては G_{ij} の値は，1 または ? のみを取ることに注意する．$\hat{G}$ は推定されたグラフ構造を表す．ℓ はロス関数である．Ω は正則化項（例えば $\ell 2$ ノルム）であり，λ は正則化項に対する重みパラメータである．$\mathbf{X}$ は既知のノード情報であり，利用者に関する属性や履歴情報等が該当する．θ は推定する必要のある未知パラメータの集合である．

学習段階において，未知パラメータ θ を目的関数である式 4.1 を最適化することにより求める．学習が完了した時点で，獲得された θ とノード情報 $\mathbf{X}$ から算出される $\hat{G}$ が，未知リンクに対する推定結果を示す．

4.2.2　行列分解と潜在特徴

ここで $\mathbf{X}$ を利用者–アイテム行列とする．このときアイテムとは，スマートフォンのアプリケーションおよび閲覧先の Web コンテンツとする．$\mathbf{X}$ の各要素は，該当する利用者が該当するアイテムを利用した頻度を表すものとする．$\mathbf{X}$ のサイズは $U\times I$ となる．ここで U は利用者の数，I はアイテムの数である．下式は $\mathbf{X}$ の行列分解を示す．

$$\mathbf{X}\simeq\mathbf{WH} \tag{4.2}$$

行列 $\mathbf{W}$ と $\mathbf{H}$ は，図 4.2 に示される様に行列分解により得られる因子行列である．$\mathbf{W}$ はサイズが $U\times L$ の利用者特徴行列となる．また $\mathbf{H}$ はサイズが $L\times I$ のアイテム特徴行列となる．ここで，L は与える正の整数であり，潜在特徴の次元数となる．式 4.3 に行列分解を行う際に用いられる典型的な目的関数を示す．

$$\min_{\mathbf{W},\mathbf{H}}\left\{\|\mathbf{X}-\mathbf{WH}\|^2+\lambda'\Omega'(\mathbf{W},\mathbf{H})\right\} \tag{4.3}$$

ここで Ω' は正則化項（例えば $\ell 2$ ノルム）であり，λ' は正則化項の重みパラメータである．

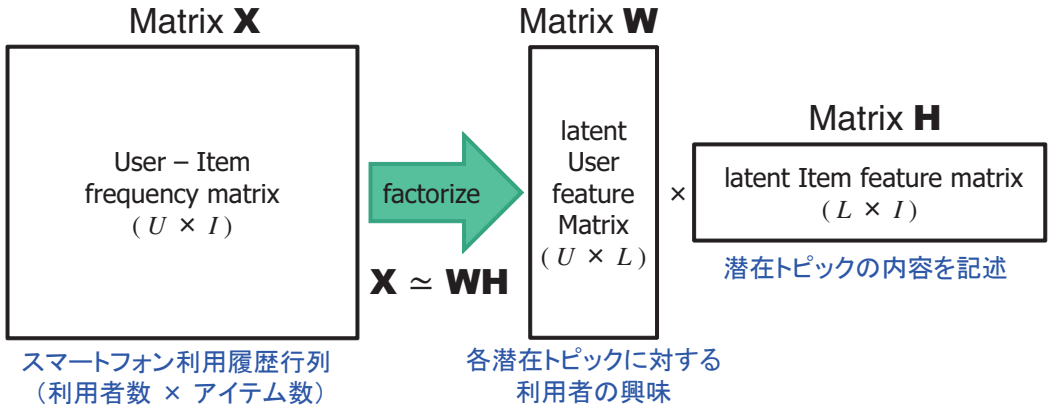


図 4.2　利用者–アイテム行列の行列分解

4.3　提案手法

4.3.1　利用者間の affinity 指標

友人関係をより良く推定するためには，友人関係の有無可能性を示す適切な affinity 指標を設定することが重要であり求められる．類似の属性を持ち類似の行動をする人同士は友人である可能性が高いであろうという発想（ホモフィリー）からノード情報のコサイン類似度が用いられることがあった．しかしながら，友人である可能性が，そのような “類似度” に常に連動することはないと考えられる．例えば，利用者 A は例えば野球，釣り，キャンピングに興味があるとしよう．利用者 A の友人である利用者 B は，必ずしもすべてについて興味がある必要はなく，利用者 B は興味の一部を共有するだけというのが普通である．したがって直感的には，友人である可能性を考える時には，共通の興味トピックの有無が重要と考えられる．さらには，興味の強さが一致しているかどうかは，重要でないことが想定される．このような考えに基づいて，利用者 i と j の間の affinity 指標として興味共有指標 f_{ij} を，シグモイド関数を用いて潜在利用者特徴を変換したベクトルの内積として下式によりモデル化する．

$$f_{ij}(\mathbf{W}) = \frac{\sigma(\mathbf{w}_i) \cdot \sigma(\mathbf{w}_j)}{L} \tag{4.4}$$

ただし，

$$\sigma(\mathbf{w}_i) = \left(\frac{1}{1+e^{-g(w_{i1}-th)}}, \cdots, \frac{1}{1+e^{-g(w_{iL}-th)}} \right) \tag{4.5}$$

ここで，$\mathbf{w}_i$ は利用者 i の潜在利用者特徴を示し，行列 $\mathbf{W}$ の i 行を抜き出したベクトル $\mathbf{w}_i = (w_{i1}, \cdots, w_{iL})$ である．また th と g は，シグモイド関数の閾値とゲインをそれぞれ表すパラメータである．$\sigma(\mathbf{w}_i)$ は，ベクトル $\mathbf{w}_i$ の各要素の値についてシグモイド関数を適用し値を $[0,1]$ に変換するため，$f_{ij}(\mathbf{W})$ は $0 \leq f_{ij}(\mathbf{W}) \leq 1$ の値を取る．

4.3.2　潜在特徴と半教師付リンク推定の統合

本検討で取り扱うタスクは one-class 設定であるので，G_{ij} は任意の $(i,j) \in O$ に対して常に 1 となる．さらには興味共有指標 f_{ij} が友人推定を行う上で $\hat{G}_{ij}$ の共変量として動作する．これらにより，式 4.1 は次式へ変形できる．

$$\min_{\mathbf{W}} \left\{ \sum_{(i,j) \in O} \ell\left(1, f_{ij}(\mathbf{W})\right) + \lambda'' \Omega''(\mathbf{W}) \right\} \tag{4.6}$$

半教師付リンク推定に興味共有指標を組み合わせるためには，式 4.3 と式 4.6 が同時に最適

化される必要がある．これを実現するために，式 4.3 と式 4.6 を結合し，目標関数として次のように定義する．

$$\min_{\mathbf{W},\mathbf{H}} \left\{ \alpha \sum_{(i,j)\in O} \ell(1, f_{ij}(\mathbf{W})) + \beta \|\mathbf{X} - \mathbf{WH}\|^2 + \lambda''' \Omega''' \right\} \tag{4.7}$$

ここで α と β は混合重みパラメータである $(\alpha, \beta > 0)$．式 4.7 から，教師付リンク推定の項（第一項）に対して，行列分解の項（第二項）が潜在特徴に関する追加制約を与える形となっていることがわかる．

さらに，ロス関数 ℓ として，単純な絶対差分値を用いるとすると，式 4.7 は下式に変形できる．

$$\min_{\mathbf{W},\mathbf{H}} \left\{ \alpha \sum_{(i,j)\in O} (1 - f_{ij}(\mathbf{W})) + \beta \|\mathbf{X} - \mathbf{WH}\|^2 + \lambda''' \Omega''' \right\} \tag{4.8}$$

ここで，正則化項 Ω''' には通常標準的な $\ell 2$ ノルムが用いられる．

$$\Omega''' = \|W\|^2 + \|H\|^2 \tag{4.9}$$

4.3.3 パラメータ推定

実験的に提案モデル／手法の妥当性を確認するために，その実装の容易さから最急降下法を適用して式 4.8 を最適化した．表 4.1 に提案モデル／手法におけるパラメータを示す．実験においては，各制御パラメータに対し探索を行い，目的関数（式 4.8）の最適化が最も良好となったパラメータを採用した．

最適化が完了した後，f_{ij} の値が利用者ペア (i, j) に対する友人関係有無の推定結果となる．

4.4 実験による評価

4.4.1 利用するデータ

モニター実験（付録 A 参照）により収集したデータより，50 人の友人グループで応募した実験協力者の履歴データを用いる．この 50 人においては，すべての実験協力者は，実験協力者の中に少なくとも一人以上の友人がいる設定となっている．収集した履歴から本検討では次の 2 種類の履歴情報を利用する．(1) アプリケーション実行履歴，(2) Web ブラウザによるインターネットコンテンツの閲覧履歴．各履歴は，時刻情報，端末 ID，および表 4.2 に示す内容の文字列にて構成される．

表 4.1　提案手法におけるパラメータ一覧

パラメータ	パラメータ種別	説　明
α	制御	学習項の重みパラメータ（式 4.8）
β	制御	行列分解項の重みパラメータ（式 4.8）
λ'''	制御	正則化項の重みパラメータ（式 4.8）
g	制御	シグモイド関数のゲインパラメータ（式 4.5）
th	制御	シグモイド関数の変曲点（式 4.5）
		= 閾値
L	制御	潜在特徴次元数
		= 行列 W の列数
		= 行列 H の行数
W	獲得	利用者特徴行列
H	獲得	アイテム特徴行列

表 4.2　収集した履歴情報の種別

履歴の種別	記録する文字列
アプリケーション実行履歴	実行したアプリケーションのパッケージ名
Web 閲覧履歴	アクセスした URL の文字列

図 4.3 および図 4.4 に当該実験協力者から収集したアプリケーション実行と Web 閲覧に関する履歴情報の統計量を示す．図 4.3 は収集した履歴情報の日毎の分量の推移を示している．また図 4.4 は各履歴情報に表れた利用者数の日毎の推移を示している．モニター期間が進むにつれ，履歴を収集できた利用者の数が漸減していることが読み取れる．

収集した履歴情報から，利用者–アイテム行列 $\mathbf{X}$ を構築する．行列 $\mathbf{X}$ の列数は 4,974 となり，履歴に現れたアプリケーションパッケージ名と URL（ドメイン名のみ）の総数である．行列 $\mathbf{X}$ の各要素の値は，各アイテムの期間中における利用頻度を表す．ここで利用頻度とは，期間中における該当アプリケーションの起動回数もしくは該当する URL を Web ブラウザにてアクセスした回数を示す．

利用履歴の収集に加えて，モニター実験の終了時に友人関係の正解を得るためにアンケート調査を行った．ここでは，3.3.2 節で示した友人関係情報と異なり，アンケートの設問の中から，実験協力者がモニター実験の期間中に，他の実験協力者一人一人を相手として，その相手に利用した電話回線にかかわらず電話で通話したことの有無を尋ねた質問に対する回答を利用する．これは本章においては友人関係の正解を得ることが目的であるため，「話したことのある相手」よりも「電話で通話した相手」の方がより確実に友人関係であろうと推測し，より適切と考えたためである．実際には親しい友人であっても通話しない場合が想定されるが，ここでは現在の世間一般の電話利用状況を鑑み，通話する相手は一定以上の関係があるものと考え，正解の部分集合が得られるので十分であると考えた．図 4.1 は，得られた結果のネットワークを示したもので，破線および実線にて得られた友人関係を示している．回答によれば期間中に 157 組において通話がなされていた．本検討においては，この 157 組を友人関係の正解集合として推定性能評価に用いる．

4.4.2 収集データの解析

まず初めに，収集された履歴データから，コサイン類似度のような類似性指標により，どの程度の友人関係が推定できるかを確認した．ここで，$\mathbf{x}_i$ が行列 $\mathbf{X}$ のなかの利用者 i に関連する行ベクトルを表すこととする．このベクトルの長さは 4,974 である．以下に示す類似性指標をすべての利用者組 (i, j) について算出した．

- コサイン類似度 (CS):

$$CS_{ij} = \frac{\mathbf{x}_i \cdot \mathbf{x}_j}{\|\mathbf{x}_i\|\|\mathbf{x}_j\|} \tag{4.10}$$

- 二値化ベクトルの内積 (DB):

$$DB_{ij} = \mathbf{x'}_i \cdot \mathbf{x'}_j \tag{4.11}$$

ここで,

$$\mathbf{x'}_i = (x'_{i1}, \cdots, x'_{iL}) \tag{4.12}$$

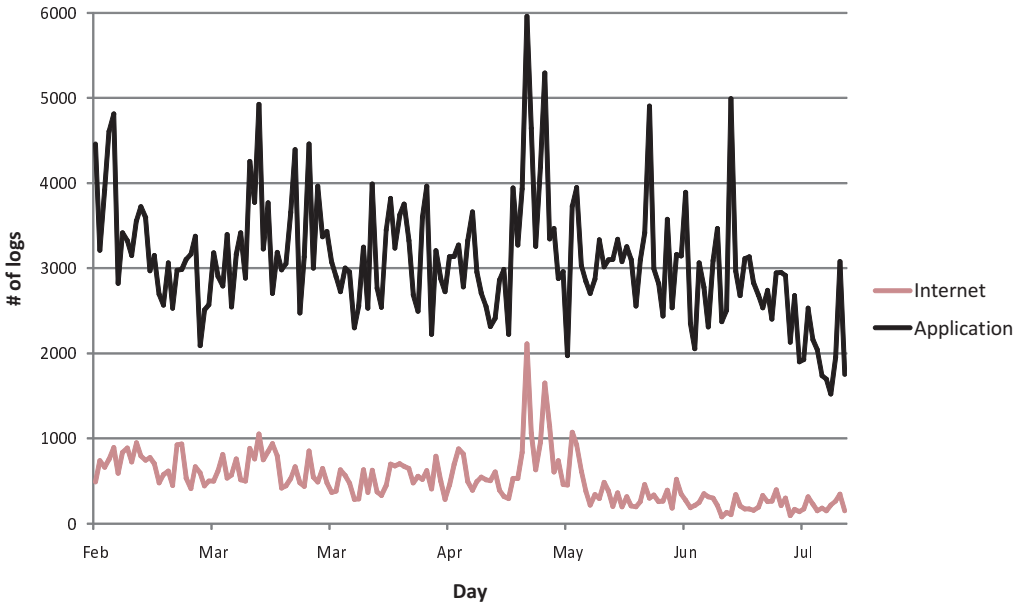


図 4.3　収集した履歴レコード数の日毎推移

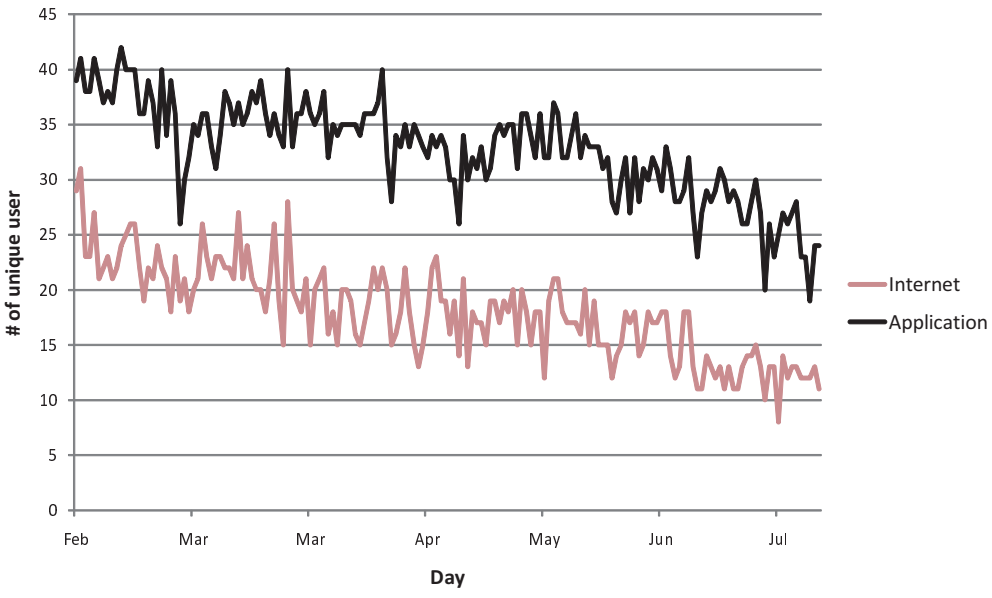


図 4.4　利用者数の日毎推移

$$x'_{ik} = \begin{cases} 1 & \text{if } x_{ik} > 0 \\ 0 & \text{それ以外} \end{cases} . \tag{4.13}$$

- 対数正規化ベクトルの内積 (DL):

$$DL_{ij} = \mathbf{x}''_i \cdot \mathbf{x}''_j \tag{4.14}$$

ここで,

$$\mathbf{x}''_i = (x''_{i1}, \cdots, x''_{iL}) \tag{4.15}$$

$$x''_{ik} = \frac{\log(x_{ik}+1)}{\max_{i'}\{\log(x_{i'k}+1)\}} \tag{4.16}$$

さらに，LSA（潜在意味解析；Latent Semantic Analysis）[12] の考え方を適用し，SVD（特異値分解；Singular Values Decomposition）により利用者特徴を算出し（特徴次元数 l は指定する），算出された利用者特徴のコサイン類似度を算出した．この結果を CS_RD として示す．

すべての利用者組を算出された類似度に従い降順に並び替え，上位 k 番目までに正解の友人関係を持つ利用者組が含まれる再現率を得た．結果を図 4.5 に示す．破線はランダムサンプリング時のベースラインである．図により，上記の類似度尺度には友人関係のある利用者組を判別する能力がないことが明らかである．すなわち，これらの履歴情報から教師付学習をせずに友人関係を推定することは困難であることが示唆される．

4.4.3 評価

提案手法の推定性能を，最近提案されたリンク推定手法である LFL モデル [48] と比較し評価する．Menon らは論文において LFL モデルの中心は既知のリンク情報から潜在特徴を獲得しリンク推定を行うことであり，ノード情報のような観測できる付加的情報を組み合わせる手法については，LFL モデルの拡張として言及している．Menon らのモデルは，潜在特徴からの推定と付加的情報から得る affinity 指標の線形結合を考え，これを最適化する方略であることから，付加的情報単独から一定の判別能力がある affinity 指標が導出できることを仮定しており，これが組み合わせた際の能力向上の源泉となっている．これに対し，4.4.2 節で示した通り，本検討の問題設定においては履歴情報を用いた類似性指標を affinity 指標として用いても予測性能の直接的改善への寄与を期待することができない．LFL モデルでは affinity 指標の獲得は視野に入れていないためである．このためここでは，LFL モデルにおいては履歴情報は特に利用せずに潜在特徴のみを用いた場合の性能により提案手法との比較をすることとする．著者が知る限りにおいて，このようなノード情報から有用な affinity 指標を同時に獲得し用いるリンク推定手法は公知文献として発表されていない．

推定性能を，3-fold 交差検定を行い，上位 k 候補に対する適合率（precision）と再現率（recall）により評価した．3-fold 交差検定において，より現実的な状況による評価とするために，正解

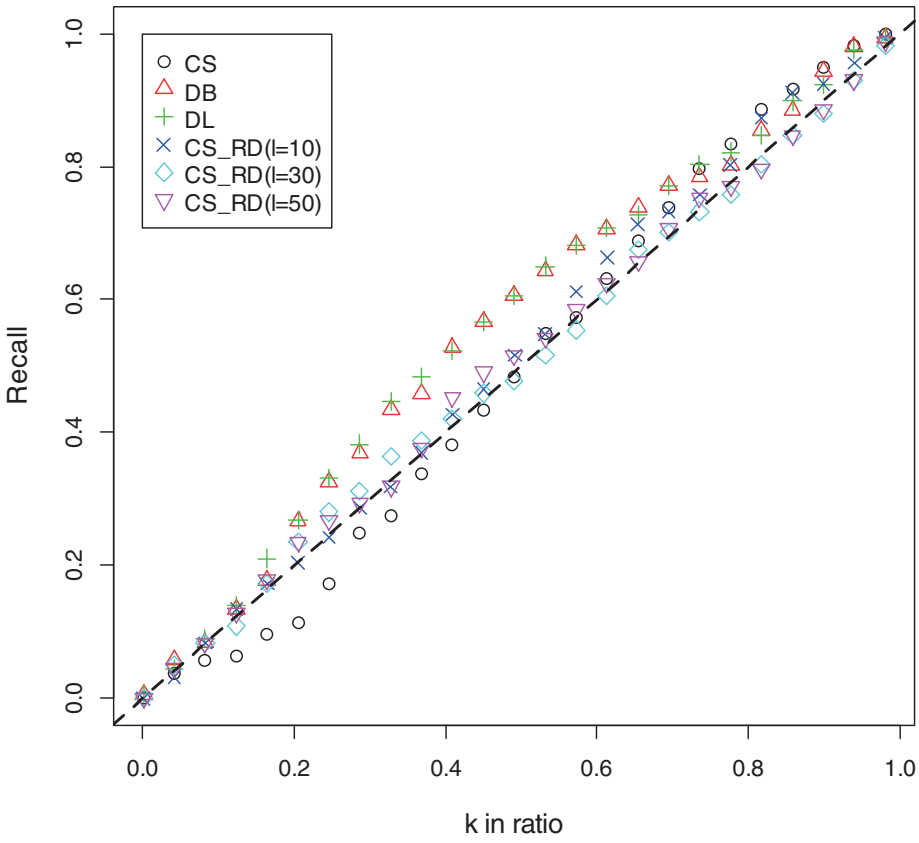


図 4.5　履歴類似性に基づいた Top-k 正解再現率

友人関係集合の 1/3 を既知として学習に用いることとした．これは，学習に用いられた利用者組以外のすべての利用者組は未知として学習の際に扱うことを意味する．正解友人関係集合の残りの 2/3 とすべての非友人の利用者組はテストセットに含まれる．提案手法と LFL モデルの双方において，制御パラメータの探索を行い，最も良い結果となったパラメータを評価に用いた．評価に用いたパラメータの値を表 4.3 に示す．再現率・適合率曲線を描くために，k を 10 から 50 まで変動させた．得られた結果を図 4.6 に示す．図より，提案手法が安定して LFL モデルを上回る性能を示したことが確認できる．この結果から，履歴情報単独では友人関係推定の能力向上に寄与できなかったが，半教師付リンク推定手法と組み合わせて履歴情報から潜在特徴を最適化して抽出し affinity 指標として利用することにより，履歴情報を用いない半教師付リンク推定手法よりも推定性能を向上できることが確認された．

4.5　むすび

本章では，利用者の利用履歴を活用して友人関係を推定する半教師付学習による手法を提案した．提案手法は，利用者–アイテム行列に対して行列分解を適用し，潜在利用者特徴を抽出する．潜在利用者特徴を半教師付リンク推定手法に埋め込むことにより，リンク推定と潜在利用者特徴が同時に最適化されることを実現した．あわせて潜在利用者特徴により友人関係推定を行うための affinity 指標として興味共有指標を提案した．学生モニター実験による実データを用いて実験的に評価を行った．提案手法が既存の手法を上回る性能を示すことが確認された．

今後の発展方向性として，次の項目があげられる．

- 確率的勾配降下法を用いることによりスケーラビリティを改善する
- 異なるデータ種別により評価を行い汎用性・適用可能範囲を確認する
- FIP モデルのような異なる観点の取り込みによるモデルの拡張

表 4.3　性能比較時のパラメータ設定

パラメータ	LFL	提案手法
潜在特徴次元数 L	50	100
Loss function	log	mae
正則化重みパラメータ λ	0.0	0.0

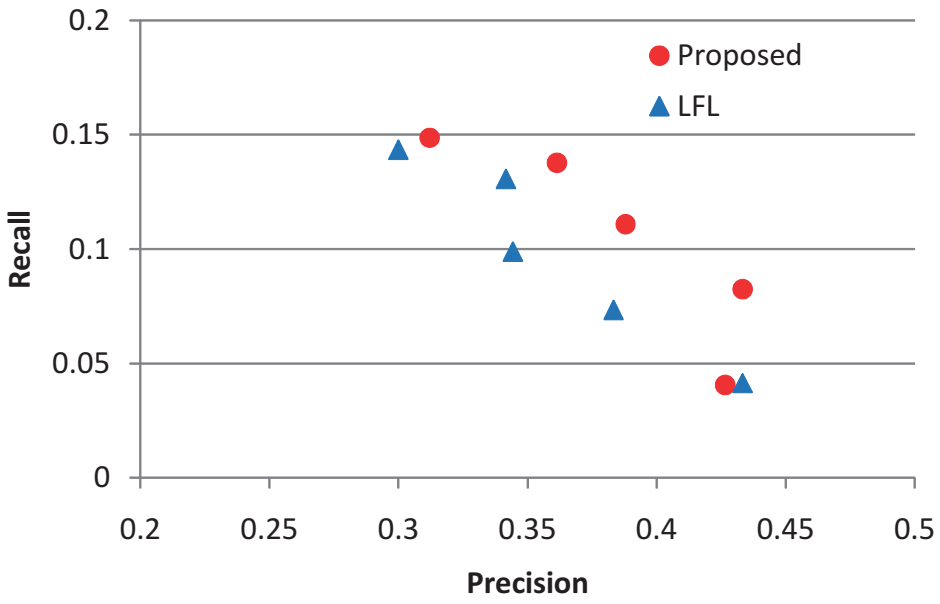


図 4.6　適合率–再現率

第 5 章

結 論

本論文では，スマートフォンの利用履歴をとらえて，コミュニケーションの構造を解析・推定する手法について論じた．以下では，本論文の内容を要約し，今後の展望についてまとめる．

第 2 章では，スマートフォンにおけるアプリケーションの利用順序には，社会構造（人間関係）による影響（インフルエンス）が反映されているとする仮説に基づき，潜在特徴モデルを構築すると，利用順序の予測が高精度に可能となることを示した．また，潜在特徴モデルと予測精度の関係を考察し，影響関係においては潜在グループ構造が存在することを示した．潜在特徴の獲得には非負行列分解（Nonnegative Matrix Factorization; NMF）による低ランク近似を用いた．予測精度の向上には NMF における行列のスパースネス制御が重要であり，結果的にモニター実験のデータでは 5~6 の潜在グループの存在が確認できるとともに，予測精度としては既存の各種の協調フィルタリング手法を上回る精度が得られることを示した．

第 3 章では，スマートフォンのアプリケーション利用履歴と友人関係情報が既知であるとして，実際にクチコミ等によって周囲に直接影響を及ぼすインフルエンサを推定・抽出する手法を提案した．第 2 章に記述したアプリケーション利用予測において得た影響関係は，利用順序を説明するモデルであるため，新しいものを好んで試用する傾向の人が広く影響を及ぼしているように表現される．しかしながら，このような人が必ずしも周囲に直接影響を及ぼしているとは限らないので，ここでは周囲に直接影響を及ぼしている利用者の推定・抽出を試みた．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を，個人間の影響度をパラメータとする確率モデルにより表現した．利用者同士に友人関係が有る場合にのみ情報伝達が起こり直接影響を与え得ることに着目し，個人間の影響度を直接影響とそれ以外の要素の混合として表した．ダウンロードの連鎖の生起確率密度分布をベータ分布にてモデル化し，MCMC 法を用いて母数推定を行い，獲得された直接影響の大きさからインフルエンサを推定した．モニター実験により得たデータを用いて，提案手法の正当性および効果を確認し，行動が早いという先行性指標よりも提案手法が優れていることが確認された．

第 4 章では，スマートフォンの利用履歴を用いて，友人関係ネットワークを半教師付学習に

より求める手法を提案した．第 3 章にて論述したインフルエンサ推定では友人関係情報を入力情報として用いたが，実際には完全な友人関係情報を得ることは容易でない．また，通話履歴やメール履歴等を観測することにより，友人同士であることを確認・推定できるが，観測できるのは全体の一部であり，すべての友人関係を把握することは網羅性の観点から困難である．このことから，半教師付学習により一部の判明している友人関係を利用して全体を推定することを試みた．友人同士においては興味の一部もしくは全部を共有していることを仮説としてモデルを構築した．教師付学習によるリンク推定の手法に対して，利用履歴（アプリケーションの利用履歴とインターネットアクセス履歴を結合したもの）を行列分解することにより得られる潜在特徴を組み合わせ，リンク推定と潜在特徴を同時に最適化することで，半教師付学習を実現した．あわせて利用者同士が友人関係である可能性を示す affinity 指標としてシグモイド関数を用いた極化潜在利用者特徴を提案した．モニター実験により得たデータを用いて，モデルの妥当性を確認し，既存・最新のリンク推定手法と比較して，より良い友人推定の性能が得られることを実験的に確認した．

以上により本論文では，世界的規模で急速に普及しつつあるスマートフォンの利用履歴を活用することで，コミュニケーション構造の推定が行えることを，モニター実験を実施し収集した一定規模の実データを用いて実証することができた．社会を構成するコミュニケーション構造を理解することは，CPS（Cyber Physical System）を実現していくために重要であり，そのための第一歩として本研究が貢献することを期待したい．

最後に，本研究を通しての残された課題ならびに今後の方向性について述べる．本研究においては全体を通してモニター実験により収集したスマートフォンの利用履歴データを基礎として検討を進めた．モニターは大阪大学の学生のみで構成されているため，世間一般の構成を代表しているとは言えない．このため一定の嗜好や特性の偏りがあることが考えられる．また友人関係の観点においても，一般からのサンプリングを行った場合を想定し比較すれば，明らかに極めて濃密な友人関係の存在する集団である．これらの影響について分析することを本研究では取り上げていないが，本論文にて述べた各検討において，共通する困難さの源の一つに，抽出したい関係性のスパースさがあった．この点において，今回の設定は問題が相対的に易しくなる設定であったと言えよう．応用を考えると，このように母集団を絞ることが可能である場合と，そうでない場合がある．可能な場合は母集団を絞ることがより良い結果を得るために有効であることが想定されるが，一般を母集団とした場合への提案手法の適用可能性とともに，実際のデータによる検証を行う必要がある．

また，本研究において提案した手法はいずれも行列分解や MCMC 法を用いているため，扱うデータが大きくなると，そのままでは計算量が著しく大きくなり対応することができない．伝統的な考え方は，代表性を維持したサンプリングデータにて代替するアプローチであるが，近来要求の高まっている個人性を精度良く捉えてサービスやマーケティングに活用するためには，本質的に限界があり適用できない．これに対して，大規模並列分散コンピューティングを

活用することにより，多量のデータを機械学習等により処理して様々な知見を得ることに対する期待が高まりつつある [6]．計算資源については，今後も高集積・高効率化，大規模化，低廉化が進展すると考えられることから，提案した手法についても並列分散処理化によって大規模データを扱える手法が開拓できる可能性がある．これを実現するために，行列分解に対しては，最適解を求めるための勾配法において一部のデータから勾配を計算しパラメータを更新する手続きを繰り返し収束させることでメモリ効率を改善する Stochastic Gradient Descent（確率的勾配降下）手法 [8] の応用や，近年画像認識等の分野で能力が注目されつつある Deep Learning[7, 11] の考え方を適用し多段の符号化器ネットワークを学習させる形とすることにより特徴抽出の高度化に加えて処理の並列化を行いやすくすることなどが，そして MCMC 法においては確率変数の条件付き依存関係を分析してチェーンの生成を最大限分割し同時に並列生成可能とする並列化への挑戦（例えば [23]）が必要であろう．

そして，さらに優れたモデル化の実現を目指して, マルチトピック化，動的な変化への対応，多様な情報（デモグラフィック属性，位置情報履歴，購買履歴，等）の活用を実現することが求められる．

さらには，本研究の成果は，CPS として現実の課題・現象へ適用し効果を上げないと，本当の意味での価値を産み出さない．すなわち，推薦やパーソナライズによるサービスの高度化や，コミュニケーションおよび社会の活性化，そして少子高齢化をはじめとする諸問題の解決に対する方法論の確立が必要である．本研究は，実世界の活動を観察することで，内部に潜在するコミュニケーション構造を推定・解析・理解することが主眼であった．CPS を実現するためには，これに加えて，実世界に働きかける機能が必要である．実世界に働きかけることにより，コミュニケーション構造に変化がもたらされ，その結果がさらにビッグデータに反映され，さらに実世界に働きかけるという大きなループを構築することが必要であろう．これこそが，ビッグデータを用いた CPS である．この働きかける方法論の研究は，サービスサイエンス [77] の課題の一つ（サービスのデザインおよび最適化）として取り組まれ始めたところであり，従来は製品やシステムを対象として取り組まれてきたオペレーションズリサーチや制御工学・システム工学等の知見・手法・枠組みを，サービスを対象として開拓・発展させようとする動きがある．対象が無形であるとともに，その品質は人の体験により評価されるものとなるので，人間工学・心理学・社会学・マーケティング等における知見との融合が必須と考えられるが，今後の発展を期待したい．本研究が，今後の CPS の実現によるより良い社会の構築への礎として貢献することを願う．

謝 辞

本研究は，著者が大阪大学大学院 情報科学研究科 博士後期課程在学中に，同大学 サイバーメディアセンター 竹村治雄教授の指導のもとに行ったものです．竹村教授には懇切丁寧な御指導と御教示を賜りました．ここに謹んで心より感謝の意を表します．また，本論文の審査過程において，内容について，御指導，御助言賜りました大阪大学大学院 情報科学研究科 尾上孝雄教授，清川清准教授に心から感謝いたします．そして，NTT ドコモ 執行役員 研究開発推進部長 栄藤稔博士（元大阪大学サイバーメディアセンター招へい教授，NTT ドコモ サービス＆ソリューション開発部長）には，社会人ドクターとして入学する機会をいただいたところから，研究を進め本論文にまとめるに至るまで，ご多忙にもかかわらず終始熱心かつ懇切丁寧な御指導・激励と数々の御教示を賜りました．心より感謝の意を表します．

また本研究は主に，博士後期課程学生及び NTT ドコモ社員メンバーの双方の立場として参画した，大阪大学サイバーメディアセンターへ 2009 年 10 月より 2012 年 3 月までの期間に設置されたドコモ（コミュニケーション構造解析）共同研究部門において行ったものです．共同研究部門において，多くの議論を通じて多大な御指導をいただきました日本大学文理学部 尾崎知伸博士（元共同研究部門 特任講師），関西大学 データマイニング応用研究センター 佐野夏樹博士（元共同研究部門 特任助教）に感謝いたします．また，共同研究部門の兼任教員として多くの御助言・御示唆・御協力を賜りました，サイバーメディアセンター 助教 間下以大博士，基礎工学研究科 准教授 土方嘉徳博士，東京未来大学 モチベーション行動科学部 教授 大坊郁夫博士（元大阪大学大学院 人間科学研究科 教授）に感謝いたします．そして，共同研究部門に参加された NTT ドコモ サービス＆ソリューション開発部のメンバーである，サービス＆ソリューション開発部 担当課長 吉村 健博士（元大阪大学サイバーメディアセンター招へい准教授），DOCOMO Innovations, Inc. 秋永和計氏（元サービス＆ソリューション開発部 主査），サービス＆ソリューション開発部 主査 藤本 拓氏，サービス＆ソリューション開発部 池部優佳氏には，数々の貴重なアドバイスや，研究／実験を進める上での多くの協力をいただきました．ここに深く感謝致します．また，社会人学生として遠隔からの所属でありましたが，機会を捉えて御討論・御支援をいただきました，サイバーメディアセンター 講師 中澤篤志博士をはじめとした竹村研究室の関係各位に心より感謝いたします．そして，日頃より御討論・御支援・御協力をいただきました NTT ドコモ サービス＆ソリューション開発部 データマイニ

ング担当の皆様に感謝いたします．加えて，本研究に関連した投稿論文に厳しくも建設的なコメントをいただいた，電子情報通信学会，情報処理学会，各国際会議の査読者各位に感謝いたします．

さらには，卒業論文ならびに修士論文の研究指導を通じて研究のイロハを教えてくださり，技術研究の道に導いてくださった大附辰夫 早稲田大学名誉教授，企業研究者としての先輩・同僚として多くのアドバイスや刺激をいただいたドコモテクノロジ 杉村利明氏，NTT ドコモ 先進技術研究所 磯俊樹博士に感謝いたします．

最後に，本論文の執筆を物心両面から辛抱強く支えてくれた，妻晴美，息子遊輝に感謝いたします．

参考文献

[1] Lada Adamic and Eytan Adar. Friends and neighbors on the web. *Social Networks*, Vol. 25, No. 3, pp. 211–230, 2003.

[2] Sinan Aral, Lev Muchnik, and Arun Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, Vol. 106, No. 51, pp. 21544–21549, December 2009.

[3] Liviu Badea. Extracting gene expression profiles common to colon and pancreatic adenocarcinoma using simultaneous nonnegative matrix factorization. In *Pacific Symposium on Biocomputing*, pp. 267–278, Kohala Coast, Hawaii, USA, January 2008.

[4] Matthias Baldauf, Schahram Dustdar, and Florian Rosenberg. A survey on context-aware systems. *International Journal of Ad Hoc and Ubiquitous Computing*, Vol. 2, No. 4, pp. 263–277, June 2007.

[5] Frank M. Bass. A new product growth for model consumer durables. *Management Science*, Vol. 50, No. 12 Supplement, pp. 1825–1832, December 2004.

[6] Ron Bekkerman, Mikhail Bilenko, and John Langford. *Scaling up Machine Learning: Parallel and Distributed Approaches*. Cambridge University Press, 2011.

[7] Yoshua Bengio. Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, Vol. 2, No. 1, pp. 1–127, January 2009.

[8] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In Yves Lechevallier and Gilbert Saporta, editors, *Proceedings of the 19th International Conference on Computational Statistics*, COMPSTAT'2010, pp. 177–187, Paris, France, August 2010. Springer.

[9] Duncan Brown and Nick Hayes. *Influencer Marketing: Who Really Influences Your Customers?* Butterworth-Heinemann, 2007.

[10] Jean-Philippe Brunet, Pablo Tamayo, Todd R. Golub, and Jill P. Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences*, Vol. 101, No. 12, pp. 4164–4169, March 2004.

[11] Jeffrey Dean, Greg Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Quoc Le, Mark Mao, Marc'Aurelio Ranzato, Andrew Senior, Paul Tucker, Ke Yang, and Andrew Ng. Large Scale

Distributed Deep Networks. In P. Bartlett, F.C.N. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pp. 1232–1240, 2012.

[12] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, Vol. 41, No. 6, pp. 391–407, 1990.

[13] Nathan Eagle, Alex (Sandy) Pentland, and David Lazer. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, Vol. 106, No. 36, pp. 15274–15278, 2009.

[14] Facebook. http://www.facebook.com/ .

[15] National Science Foundation. Cyber-Physical Systems (CPS), September 2008. Program Announcements & Information. Document Number: nsf08611. (http://www.nsf.gov/pub_summ.jsp?ods_key=nsf08611).

[16] Hiroshi Fujimoto, Minoru Etoh, Akira Kinno, and Yoshikazu Akinaga. Web user profiling on proxy logs and its evaluation in personalization. In *Proceedings of the 13th Asia-Pacific web conference on Web technologies and applications*, APWeb'11, pp. 107–118, Beijing, China, April 2011. Springer.

[17] Sadaoki Furui. Speech and speaker recognition evaluation. In Laila Dybkjœr, Holmer Hemsen, Wolfgang Minker, and Nancy Ide, editors, *Evaluation of Text and Speech Systems*, Vol. 37 of *Text, Speech and Language Technology*, pp. 1–27. Springer Netherlands, 2007.

[18] Renaud Gaujoux and Cathal Seoighe. A flexible R package for nonnegative matrix factorization. *BMC bioinformatics*, Vol. 11, No. 1, pp. 367+, July 2010.

[19] Lise Getoor and Christopher P. Diehl. Link mining: a survey. *ACM SIGKDD Explorations Newsletter*, Vol. 7, No. 2, pp. 3–12, 2005.

[20] Seth Godin. バイラルマーケティング：アイディアバイルスを解き放て!（大橋禅太郎 訳）. 翔泳社, 2001.

[21] Jacob Goldenberg, Barak Libai, and Eitan Muller. Talk of the Network: A Complex Systems Look at the Underlying Process of Word-of-Mouth. *Marketing Letters*, Vol. 12, No. 3, pp. 211–223, August 2001.

[22] Gene H. Golub and Charles F. Van Loan. The singular value decomposition. In *Matrix computations (3rd ed.)*, pp. 70–71. Johns Hopkins University Press, Baltimore, Maryland, USA, 1996.

[23] Joseph Gonzalez, Yucheng Low, Arthur Gretton, and Carlos Guestrin. Parallel gibbs sampling: From colored fields to thin junction trees. In *Proceedings of the 14th international conference on Artificial Intelligence and Statistics*, AISTATS 2011, Ft. Lauderdale, Florida, USA, May 2011.

[24] Amit Goyal, Francesco Bonchi, and Laks V.S. Lakshmanan. Learning influence probabilities in social networks. In *Proceedings of the 3rd ACM international conference on Web Search and Data Mining*, WSDM '10, pp. 241–250, New York, New York, USA, February 2010. ACM.

[25] Daniel Gruhl, R. Guha, David Liben-Nowell, and Andrew Tomkins. Information diffusion through blogspace. In *Proceedings of the 13th international conference on World Wide Web*, WWW '04, pp. 491–501, New York, New York, USA, May 2004. ACM.

[26] Yuka Ikebe, Masaji Katagiri, and Haruo Takemura. Friendship prediction using semi-supervised learning of latent features in smartphone usage data. In *Proceedings of the 4th international conference on Knowledge Discovery and Information Retrieval*, KDIR 2012, pp. 199–205, Barcelona, Spain, October 2012. SciTePress.

[27] Hisashi Kashima, Tsuyoshi Kato, Yoshihiro Yamanishi, Masashi Sugiyama, and Koji Tsuda. Link propagation: A fast semi-supervised learning algorithm for link prediction. In Haesun Park, Srinivasan Parthasarathy, and Huan Liu, editors, *Proceedings of the 2009 SIAM International Conference on Data Mining*, SDM 2009, pp. 1099–1110, Sparks, Nevada, USA, May 2009. SIAM.

[28] Masaji Katagiri and Minoru Etoh. Social influence modeling on smartphone usage. In *Proceedings of the 7th international conference on Advanced Data Mining and Applications, Part II*, ADMA 2011, pp. 292–303, Beijing, China, December 2011. Springer.

[29] Masaji Katagiri and Minoru Etoh. Implicit influencing group discovery from mobile applications usage. *IEICE Transactions on Information and Systems*, Vol. E95-D, No. 12, pp. 3026–3036, December 2012.

[30] Elihu Katz and Paul Felix Lazarsfeld. *Personal Influence: The Part Played by People in the Flow of Mass Communications*. Foundations of communications research. Free Press, Glencoe, Illinois, USA, 1955.

[31] Noriaki Kawamae, Hitoshi Sakano, and Takeshi Yamada. Personalized recommendation based on the personal innovator degree. In *Proceedings of the 3rd ACM conference on Recommender Systems*, RecSys '09, pp. 329–332, New York, New York, USA, October 2009. ACM.

[32] Ed Keller and Brad Fay. The Role of Advertising in Word of Mouth. *Journal of Advertising Research*, Vol. 49, No. 2, pp. 154–158, June 2009.

[33] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the 9th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '03, pp. 137–146, Washington, DC, USA, August 2003. ACM.

[34] Hyunsoo Kim and Haesun Park. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics*, Vol. 23, No. 12, pp. 1495–1502, May 2007.

[35] Masahiro Kimura, Kazumi Saito, and Ryohei Nakano. Extracting influential nodes for information diffusion on a social network. In *Proceedings of the 22nd national conference on Artificial Intelligence - Volume 2*, AAAI 2007, pp. 1371–1376, Vancouver, British Columbia, Canada, July 2007. AAAI Press.

[36] Yehuda Koren. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pp. 426–434, Las Vegas, Nevada, USA, August 2008. ACM.

[37] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, Vol. 42, No. 9, pp. 30–37, September 2009.

[38] Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, Vol. 401, No. 6755, pp. 788–791, October 1999.

[39] Daniel D. Lee and H. Sebastian Seung. Algorithms for Non-negative Matrix Factorization. In Todd K. Leen, Thomas G. Dietterich, and Volker Tresp, editors, *Advances in Neural Information Processing Systems 13*, pp. 556–562, Cambridge, Massachusetts, USA, April 2001. MIT Press.

[40] Yanen Li, Jia Hu, ChengXiang Zhai, and Ye Chen. Improving one-class collaborative filtering by incorporating rich user information. In *Proceedings of the 19th ACM international conference on Information and Knowledge Management*, CIKM '10, pp. 959–968, Toronto, Ontario, Canada, October 2010. ACM.

[41] LinkedIn. http://www.linkedin.com/ .

[42] Lu Liu, Jie Tang, Jiawei Han, Meng Jiang, and Shiqiang Yang. Mining topic-level influence in heterogeneous networks. In *Proceedings of the 19th ACM international conference on Information and Knowledge Management*, CIKM '10, pp. 199–208, Toronto, Ontario, Canada, October 2010. ACM.

[43] Yucheng Low, Joseph Gonzalez, Aapo Kyrola, Danny Bickson, Carlos Guestrin, and Joseph M. Hellerstein. GraphLab: A new parallel framework for machine learning. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence*, UAI 2010, Catalina Island, California, USA, July 2010. AUAI Press.

[44] Linyuan Lü and Tao Zhou. Link prediction in complex networks: A survey. *Physica A Statistical Mechanics and its Applications*, Vol. 390, pp. 1150–1170, March 2011.

[45] David J. Lunn, Andrew Thomas, Nicky Best, and David Spiegelhalter. WinBUGS a bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, Vol. 10, No. 4, pp. 325–337, October 2000.

[46] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, Vol. 27, No. 1, pp. 415–444, 2001.

[47] Aditya Krishna Menon and Charles Elkan. A log-linear model with latent features for dyadic prediction. In *Proceedings of the 2010 IEEE International Conference on Data Mining*, ICDM '10, pp. 364–373, Sydney, Australia, December 2010. IEEE Computer Society.

[48] Aditya Krishna Menon and Charles Elkan. Link prediction via matrix factorization. In *Proceedings of the 2011 European conference on Machine learning and knowledge discovery in databases - Volume Part II*, ECML PKDD'11, pp. 437–452, Athens, Greece, September 2011. Springer.

[49] Seyed Hamid Mirisaee, Saman Noorzadeh, and Ashkan Sami. Mining friendship from cell-phone switch data. In *Proceedings of the 3rd international conference on Human-Centric Computing*, HumanCom 2010, pp. 1–5, Cebu, Philippines, August 2010.

[50] mixi. http://mixi.jp/ .

[51] MySpace. http://www.myspace.com/ .

[52] Tomonobu Ozaki and Minoru Etoh. Social network inference of smartphone users based on information diffusion models. In *Proceedings of the 7th international conference on Advanced Data Mining and Applications, Part II*, ADMA 2011, pp. 304–317, Beijing, China, December 2011. Springer.

[53] Rong Pan, Yunhong Zhou, Bin Cao, Nathan Nan Liu, Rajan M. Lukose, Martin Scholz, and Qiang Yang. One-class collaborative filtering. In *Proceedings of the 8th IEEE International Conference on Data Mining*, ICDM 2008, pp. 502–511, Pisa, Italy, December 2008. IEEE Computer Society.

[54] Wei Pan, Nadav Aharony, and Alex Pentland. Composite social network for predicting mobile apps installation. In Wolfram Burgard and Dan Roth, editors, *Proceedings of the 25th AAAI Conference on Artificial Intelligence*, AAAI 2011, pp. 821–827, San Francisco, California, USA, August 2011. AAAI Press.

[55] Alberto Pascual-Montano, J.M. Carazo, Kieko Kochi, Dietrich Lehmann, and Roberto D. Pascual-Marqui. Nonsmooth nonnegative matrix factorization (nsNMF). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 28, No. 3, pp. 403–415, March 2006.

[56] Renana Peres, Eitan Muller, and Vijay Mahajan. Innovation diffusion and new product growth models: A critical review and research directions. *International Journal of Research in Marketing*, Vol. 27, No. 2, pp. 91–106, June 2010.

[57] Daniele Quercia, Jonathan Ellis, and Licia Capra. Using mobile phones to nurture social networks. *IEEE Pervasive Computing*, Vol. 9, No. 3, pp. 12–20, July 2010.

[58] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008.

[59] Yossi Richter, Elad Yom-Tov, and Noam Slonim. Predicting customer churn in mobile networks through analysis of social groups. In *Proceedings of the SIAM International Confer-*

ence on Data Mining, SDM 2010, pp. 732–741, Columbus, Ohio, USA, May 2010. SIAM.

[60] Everett M. Rogers. *Diffusion of Innovations*. Free Press, 5th edition, August 2003.

[61] Paat Rusmevichientong, Shenghuo Zhu, and David Selinger. Identifying early buyers from purchase data. In *Proceedings of the 10th ACM SIGKDD international conference on Knowledge Discovery and Data mining*, KDD '04, pp. 671–677, Seattle, Washington, USA, August 2004. ACM.

[62] Kazumi Saito, Ryohei Nakano, and Masahiro Kimura. Prediction of information diffusion probabilities for independent cascade model. In Ignac Lovrek, Robert Howlett, and Lakhmi Jain, editors, *Knowledge-Based Intelligent Information and Engineering Systems*, Vol. 5179 of *Lecture Notes in Computer Science*, pp. 67–75. Springer, September 2008.

[63] Salvatore Scellato, Anastasios Noulas, and Cecilia Mascolo. Exploiting place features in link prediction on location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge Discovery and Data mining*, KDD '11, pp. 1046–1054, San Diego, California, USA, August 2011. ACM.

[64] Xiaodan Song, Belle L. Tseng, Ching-Yung Lin, and Ming-Ting Sun. Personalized recommendation driven by information flow. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pp. 509–516, Seattle, Washington, USA, August 2006. ACM.

[65] David Spiegelhalter, Andrew Thomas, Nicky Best, and Wally Gilks. *BUGS 0.5 Bayesian inference Using Gibbs Sampling. Version 0.5, (version ii)*. MRC Biostatistics Unit, Cambridge, 1996.

[66] Xiaoyuan Su and Taghi M. Khoshgoftaar. A survey of collaborative filtering techniques. *Advances in Artificial Intelligence*, Vol. 2009, pp. 1–19, January 2009.

[67] Twitter. https://twitter.com/ .

[68] Dashun Wang, Dino Pedreschi, Chaoming Song, Fosca Giannotti, and Albert-Laszlo Barabasi. Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge Discovery and Data mining*, KDD '11, pp. 1100–1108, San Diego, California, USA, August 2011. ACM.

[69] Fei Wang, Tao Li, Xin Wang, Shenghuo Zhu, and Chris Ding. Community discovery using nonnegative matrix factorization. *Data Mining and Knowledge Discovery*, Vol. 22, pp. 493–521, May 2011.

[70] Duncan J. Watts and Peter S. Dodds. Influentials, networks, and public opinion formation. *Journal of Consumer Research*, Vol. 34, No. 4, pp. 441–458, December 2007.

[71] Liu Weixiang, Zheng Nanning, and You Qubo. Nonnegative Matrix Factorization And Its Applications In Pattern Recognition. *Chinese Science Bulletin*, Vol. 51, No. 1, pp. 7–18, January 2006.

[72] Shuang-Hong Yang, Bo Long, Alex Smola, Narayanan Sadagopan, Zhaohui Zheng, and Hongyuan Zha. Like like alike: joint friendship and interest propagation in social networks. In *Proceedings of the 20th international conference on World Wide Web*, WWW '11, pp. 537–546, Hyderabad, India, April 2011. ACM.

[73] Junying. Zhang, Le Wei, Xuerong Feng, Zhen Ma, and Yue Wang. Pattern expression nonnegative matrix factorization: Algorithm and applications to blind source separation. *Computational Intelligence and Neuroscience*, Vol. 2008, pp. 1–10 (online), April 2008.

[74] 片桐雅二, 栄藤稔. スマートフォンアプリ実行ログからのインフルエンスグループの発見によるインフルエンサとイノベータの推定. 第 4 回 Web とデータベースに関するフォーラム, WebDB Forum 2011, 東京, November 2011. 情報処理学会.

[75] 片桐雅二, 栄藤稔. スマートフォンのアプリケーション実行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. NTT DOCOMO テクニカル・ジャーナル, Vol. 20, No. 2, pp. 54–58, July 2012.

[76] 片桐雅二, 栄藤稔, 竹村治雄. スマートフォンアプリケーション実行ログからのインフルエンサ推定. 情報処理学会論文誌 データベース, Vol. 5, No. 3, pp. 75–85, September 2012.

[77] 上林憲行. サービスサイエンス入門 : ICT 技術が牽引するビジネスイノベーション. オーム社, 2007.

[78] 池田謙一（編）. クチコミとネットワークの社会心理 : 消費と普及のサービスイノベーション研究. 東京大学出版会, 2010.

[79] 豊田秀樹（編）. マルコフ連鎖モンテカルロ法. 統計ライブラリー. 朝倉書店, 2008.

[80] 中沢潤, 大野木裕明, 南博文. 心理学マニュアル — 観察法. 北大路書房, 1997.

[81] 川前徳章, 山田武士, 上田修功. Relative innovator の発見によるパーソナライズ手法の提案. 情報科学技術レターズ, Vol. 6, pp. 99–102, August 2007.

付 録

A　スマートフォン利用モニター実験の概要

A.1　目的

本モニター実験は，コミュニケーション構造の解析研究を遂行するために必要なデータを収集することを目的として，特にスマートフォンをモニターすることで得られる利用履歴および位置情報を必要なアンケート回答とともに 150 人以上の規模にて 6 ヶ月程度収集する．

A.2　概要

実験実施主体

大阪大学 サイバーメディアセンター ドコモ（コミュニケーション構造解析）共同研究部門

実験実施期間

2011 年 1 月 ~ 2011 年 7 月（約 6 ヶ月間）

モニター実験協力者数

157 名

モニターの募集と採用

モニターの募集は，大阪大学キャンパス内の掲示板および学内向け電子掲示板（KOAN[*1]掲示板）に募集案内を掲出することにより行った．応募受付のために専用 Web フォームを開設し，希望者は学内の情報教育端末 PC[*2]から応募する形とした．また，受付サイトでは，実験内容・趣旨・取得データの取扱いおよび研究目的の利用などに関する説明を提示し，許諾が得られる場合にのみ受付専用フォームに記入するように設計した．応募条件として下記を設定

[*1] 大阪大学教務情報システム：Knowledge of Osaka-university Academic Nucleus

[*2] キャンパス内に情報教育向けに設置されている共有 PC，講義等で利用されていない場合には一定の条件のもと学生は自由に利用できる

した；

(1) 大阪大学の学生であること
(2) 平成 22 年度 ~ 平成 23 年度の間に在籍予定であること
(3) 利用規約に同意すること
(4) 3 人以上 50 人以下のグループで応募すること，ただしグループ構成者全員が条件 (1)~(3) を満足すること

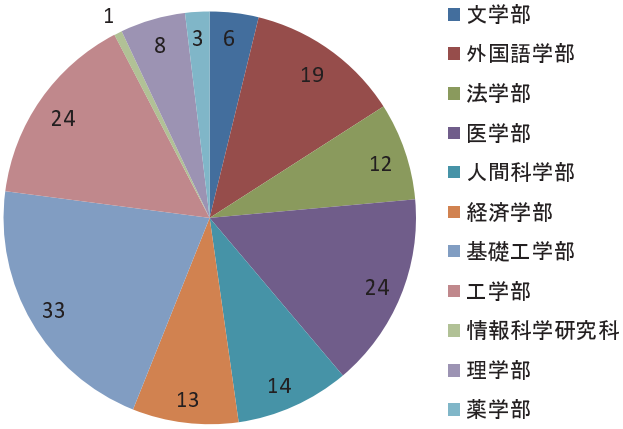
応募の受付は 2011 年 1 月 6 日から 1 月 21 日の 16 日間行い，75 グループ，総計 1,139 人からの応募があった．この中から抽選にて，6 グループ，157 人を採用した．採用者の所属学部構成を図 A.1(a) に，学年構成を図 A.1(b) に，性別構成を図 A.1(c) に示す．また各応募グループ毎の学部／学年／性別構成を図 A.2~A.4 に示す．

実験実施に係る倫理審査

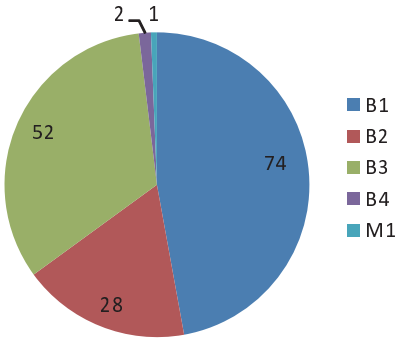
本実験の実施については，2010 年 12 月 16 日付にて大阪大学サイバーメディアセンター教授会において倫理審査を受け，承認されている．

スマートフォン利用履歴収集システム

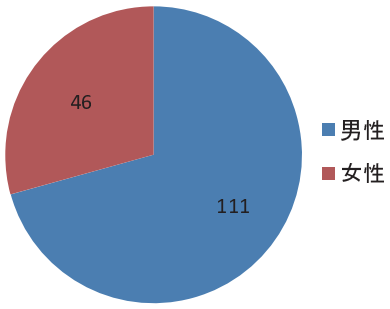
利用履歴の収集は，図 A.5 に示すシステムにより行った．各スマートフォン端末には，履歴収集用の専用アプリケーションをインストールし，逐次履歴を収集する．収集された履歴情報は，専用アプリケーションが一定時間間隔で専用サーバへアップロードする．収集する履歴項目およびサーバへのアップロード間隔については，サーバから端末の専用アプリケーションを制御できる．



(a) 所属学部構成



(b) 学年構成



(c) 性別構成

図 A.1　実験協力者の構成

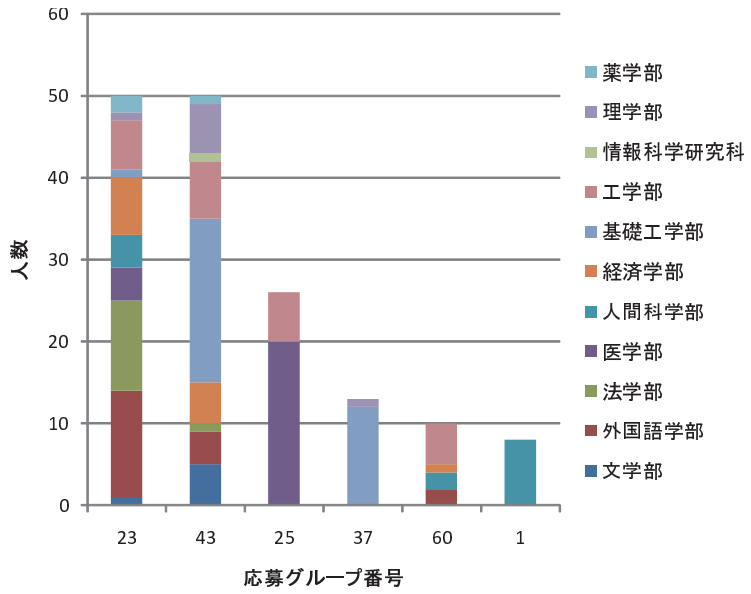


図 A.2　実験協力者の所属学部構成（応募グループ別）

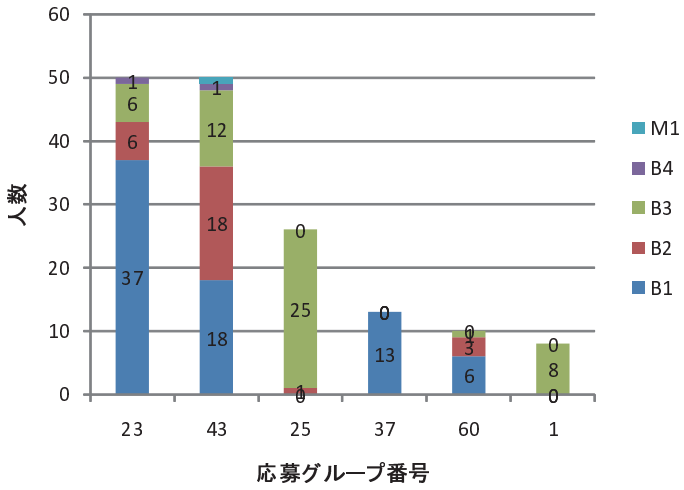


図 A.3　実験協力者の学年構成（応募グループ別）

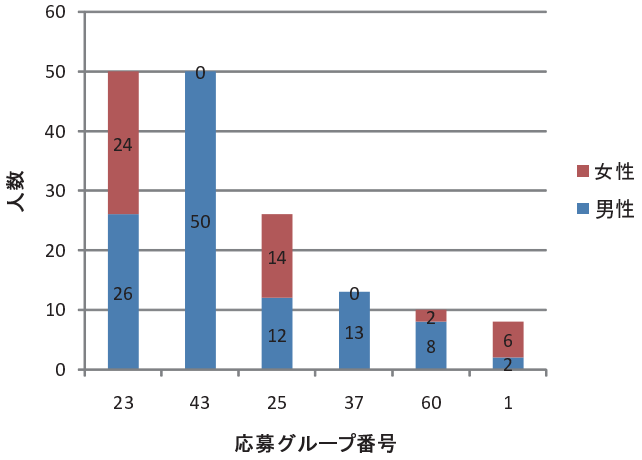


図 A.4　実験協力者の性別構成（応募グループ別）

図 A.5　利用履歴収集システム

A.3 履歴データ

本実験においては，表 A.1 に示す情報を，履歴データとして収集した．

実験期間に収集された履歴データの概要として，収集された利用履歴のレコード数の期間中日毎推移を図 A.6 に示す．実験開始当初の 2 週間程度は，各種の試用のため利用が多かったと考えられる．それ以降は，ほぼ一定量のレコード数にて推移した．次に，利用履歴が収集された利用者の人数の期間中日毎推移を図 A.7 に示す．開始当初より 10% 程度の端末から履歴が収集できていないことがわかる．また期間が進むにつれ，収集できた利用者数は漸減している．原因としては，端末の電源断や履歴収集アプリケーションの停止もしくは意図しないアンインストール等が考えられる．

収集された利用履歴レコード総数に関する表 A.1 の機能毎内訳を図 A.8 に示す．アプリケーションの利用履歴が全体の 2/3 以上を占め，スマートフォンの利用実態を表している．通話の利用履歴数は 9,235 で，1 人あたり平均では 9.8 回/月 であった．また通話の利用者には，一部片寄りがみられた．期間限定の実験用貸与端末であることも，通話利用が少なかったことの要因である可能性がある．また位置情報については自動的・定期的に測位し履歴を残す形であるが，屋内での GPS 測位は GPS 衛星からの電波受信が困難で測位できない場合が多いことを反映し，大半は測位不可のレコードであった．

図 2.5，図 3.1 は，すべてのアプリケーション利用履歴から，初回実行分のみを抽出して集計したものである．また，図 2.6，表 2.2，図 2.7 は，初回実行分の集合からさらに学習セットとして設定した条件（期間および利用者数）を満たす履歴のみの集計に基づいている．さらに図 4.3，図 4.4 については，応募グループ 43 に所属する利用者に関する履歴を抽出し集計したものである．

A.4 アンケートデータ

履歴情報の収集と合わせて，アンケート調査を行った．実験終了時（貸与端末の返却時）に行ったアンケート内容を図 A.9 に示す．さらに 2011 年 12 月に電子メールを用いて，図 A.10 に示す内容の追加アンケート調査を実施した．

実験終了時アンケートの回収率は 156/157（99.4%）であった．また，電子メールによる追加アンケートの回収率は 44/157（28.0%）であった．

---

*3 起動またはバックグラウンドからの復帰

*4 個人情報保護のため，通話先電話番号は個人が特定でいないよう一方向ハッシュ関数により置換・匿名化して記録する

*5 個人情報保護のため，ログ収集後秘匿パラメータによるオフセット処理を行い，匿名化した上で研究目的に利用する

表 A.1　履歴情報

履歴の種別	記録する内容
アプリケーション実行履歴	アプリケーション名，パッケージ名，起動種別*3
Web 閲覧履歴	アクセス先 URL
通話履歴	発信／着信相手の電話番号*4，応答の有無，通話時間
位置情報（GPS）	緯度*5，経度*5，計測誤差

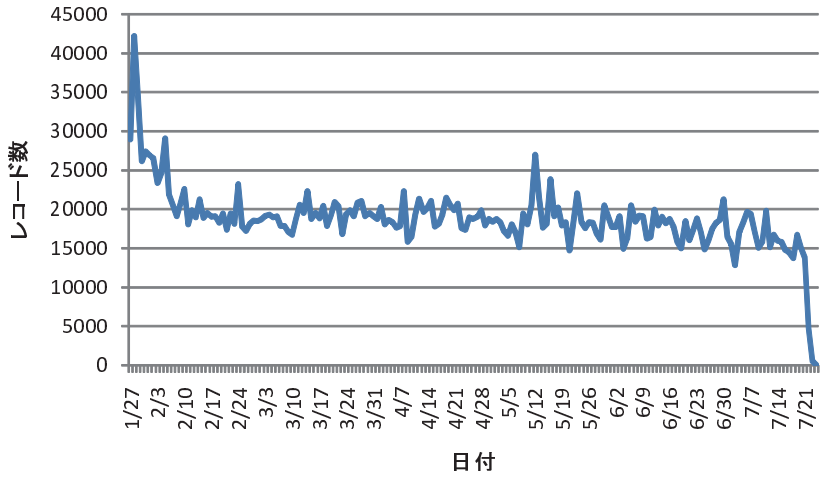


図 A.6　利用履歴レコード数の日毎推移

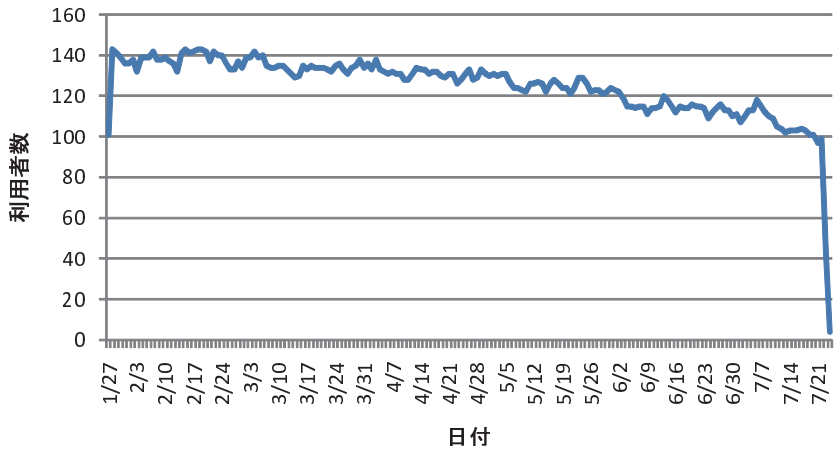


図 A.7　利用者数の日毎推移

図 3.3，図 4.1 は，このアンケート結果の一部を集計・可視化したものである．また，友人関係アンケートの回答を集計して得た平均友人数を表 A.2 に示す．表中において，「会話」，「メール」，「電話」は，それぞれ (1) 会話をしたことがある，(2) メールをしたことがある，(3) 電話をしたことがある友人の平均数である．ただし，この友人数は各自が所属する応募グループ内の友人数であり，応募グループの構成人数が一定でないことに注意が必要である．

友人から聞いたことがきっかけでダウンロードしたアプリケーションの占める割合，およびきっかけを作った友人の数について、アンケート回答集計結果を図 A.11 示す．この結果より，友人から影響を受けて使い始めるアプリケーションは一定数見込めるが，実際に使うアプリケーションのなかに占める割合は多くなく，また影響を受ける友人数も数人程度と少ないことが読み取れる．直接的な「クチコミ」の効果については実際のところ，ここで示された程度の割合と考えるのが妥当であろう．

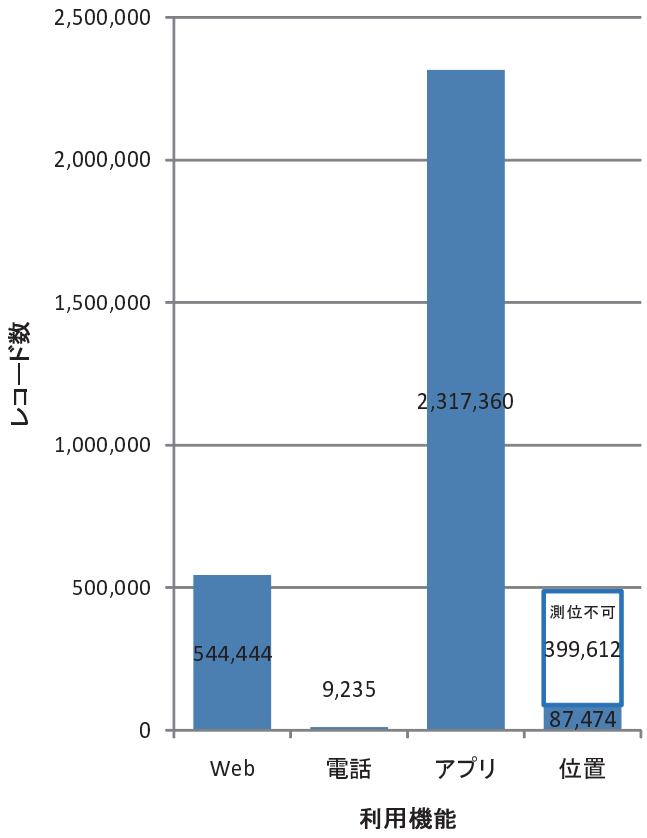


図 A.8　利用履歴レコードの内訳

表 A.2　友人関係アンケートの回答集計結果

応募グループ	人数	平均友人数		
		会話	メール	電話
23	50	13.4	9.3	7.0
43	50	23.8	9.5	6.4
25	26	13.0	6.9	4.0
37	13	8.6	5.3	3.7
60	10	8.0	5.9	4.0
1	8	7.0	5.9	3.4
全体	157	15.6	8.2	5.7

アンケート

実験にご参加いただきありがとうございました。最後にアンケートのご協力をお願いいたします。
※収集した情報は匿名化し、研究目的にのみ利用します。

学籍番号 ＿＿＿＿＿＿＿＿＿＿

名前 ＿＿＿＿＿＿＿＿＿＿

1．実験に参加された理由は何ですか？（複数回答可）

Xperiaに興味があったから　携帯電話がタダで使えるから　友人に誘われて
その他（　　）

2．Xperiaの使い心地はいかがでしたか？

大変よい　よい　普通　いまいち　大変使いにくい

3．Xperiaを主に何に利用しましたか？上位3つに〇をつけてください。

mixi　twitter　メール　電話　音楽プレーヤー　ナビ（地図）
ゲーム　動画視聴　カメラ　その他（　　）

4．実験期間中、パソコンの利用時間(授業を除く)とXperiaの利用時間の割合はどのくらいでした

パソコン：Xperia ＝ 0:100　20:80　50:50　80:20　100:0

5．実験期間中の電話利用のうち、ご自身の携帯電話とXperiaの利用割合はどのくらいでしたか

ご自身の携帯：Xperia ＝ 0:100　20:80　50:50　80:20　100:0
電話をしていない

6．いくつぐらいのアプリケーションをダウンロードしましたか？

0　1～10　11～20　20以上

7．（6の回答が「0」以外の方）
そのうち、友人から聞いたことがきっかけでダウンロードしたアプリケーションの割合は？

0%　20%　50%　80%　100%

8．（7の回答が「0%」以外の方）その友人は何人ぐらいですか？

0人　2,3人　10人　10人以上

9．実験期間中、アプリケーションについて話をした実験参加者は何人ぐらいいましたか？

0人　2,3人　10人　10人以上

10．Xperiaの使い方を教え合える人は、実験参加者に何人ぐらいいましたか？

0人　2,3人　10人　10人以上

うらに続く ↓　↓　↓

11．机の上のメンバリストを見て、該当IDの欄に〇×をつけてください。

Q1：話したことがある人に〇、話したことがない人に×
Q2：実験期間中に電話をした人に〇、していない人に×
Q3：実験期間中にメールをした人に〇、していない人に×

※電話やメールはXperiaの利用に限りません
※メールはメーリングリストなどの一斉送信は含みません
※ご自身の欄は空白としてください

ID	Q1 話し	Q2 電話	Q3 メール
1			
2			
3			
4			
5			
6			
7			
8			
9			
10			
11			
12			
13			
14			
15			
16			
17			
18			
19			
20			
21			
22			
23			
24			
25			
26			

12．ご意見、ご感想などありましたらご記入ください。

ご協力ありがとうございました。

図 A.9　実験終了時アンケート（応募グループ 25 用）

○質問１：
スマートフォン実験にて、他の誰かから紹介されたことがきっかけで
アプリケーションをダウンロードして使ってみたことがあったと
大多数の方からお答えいただきました。

・質問１－１：
紹介されたことがきっかけでダウンロードして使ってみたのはどんな
アプリケーションだったでしょうか。
アプリ名を記入してください（複数可）
アプリ名が分からない場合は、何をするアプリかをご回答ください。

・質問１－２：
具体的にどなたからの紹介であったかを教えてください。
（お名前をフルネームでご回答ください；複数可）

・質問１－３：
あなたが、他の方へアプリケーションを紹介したことはありますか。
ある場合は、どのようなアプリを紹介したのか教えてください。
（複数可）

○質問２：
スマートフォンの利用について教えてください。

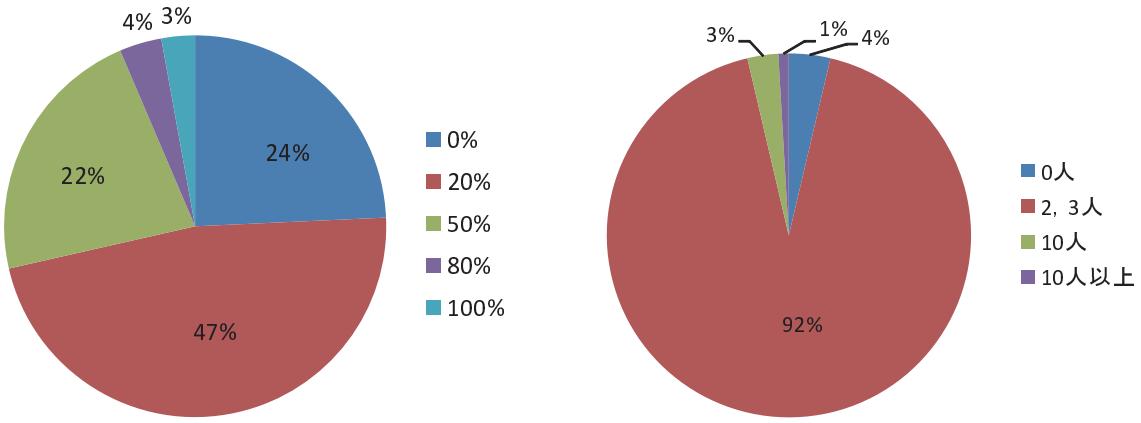
・質問２－１：
実験終了から４か月ほど経過しましたが、その後スマートフォンを
お使いでしょうか？

１：実験以前から使っている
２：実験中または実験後に使い始めた
３：今後使ってみたいと思っているが、まだ使っていない
４：当面スマートフォンを使うつもりはない

・質問２－２：
スマートフォンをお使いの方（質問２－１で１，２とお答えの方）は
お使いの端末の種類を教えてください。

１：iPhone
２：アンドロイド
３：その他（回答欄に種類を回答ください；例：Windows，Blackberry，他
４：スマートフォンは使っていない

図 A.10　追加アンケート



(a) 友人から聞いてダウンロードしたアプリケーションの割合

(b) アプリケーションをダウンロードするきっかけとなった友人の数

図 A.11　アプリケーションダウンロードのきっかけに関するアンケート結果

コミュニケーション研究のための スマートフォン利用モニター募集

実験の目的

スマートフォンの利用状況から、コミュニケーションの構造解析を行う

実験の内容

NTTドコモの製品Xperiaを貸出します
通信料、通話料はこちらで負担します
半年間、自由に利用してください
利用のログを収集させていただきます

OSはAndroid2.1 最新バージョンです！

利用ログは匿名化処理した上で研究目的で各種統計分析を行います

募集期間

2011年1月6日～21日

詳しくはこちら

http://192.168.177.201/cgi-bin/entry.cgi

大学設置の情報教育端末からのみアクセス可能です

大阪大学 サイバーメディアセンター
コミュニケーション構造解析共同研究部門
TEL:: 090-5[redacted]-[redacted]
E-mail: xperia_support@dcm.cmc.osaka-u.ac.jp

図 A.12 学内に掲出したモニター募集ポスター