行動ログを用いた
人の行為選択のモデル化と予測

佐藤 晃矢

グローバル教育院

エンパワーメント情報学プログラム

筑波大学

2020年3月

概要

Web の誕生により情報の比較が容易となったことで，個々の選択肢はこれまで以上に競争的な環境にさらされている．SNS 内で用いられるタグや SNS あるいは企業は競争的な環境にさらされている選択肢の例であり，様々な選択肢との比較を通して人々から選択される．このような競争的な環境において，これらの選択肢が選択される，あるいは選択され続けるためには，複数の選択肢からの選択と，一旦選択した選択肢からの離脱という，2 つの側面から人の行為選択の詳細なメカニズムを明らかとする必要がある．

これを明らかにする上で人の行動ログの分析は有益である．行動ログは個人あるいは集団による人の行為選択の結果を反映したものである．実際に，SNS では多数のユーザがサービスに参加し投稿や他のユーザとのやりとりがサーバ上にタイムスタンプとともに保存されている．取得される行動ログは Web の世界だけにとどまらず，企業における勤怠情報も電子的に取得，保管されている．

そこで本研究では，タグや SNS あるいは企業の選択に関する行動ログを用いて，人の行為選択のモデル化と予測により，人の行為選択の詳細なメカニズムを明らかにする．本研究では特に，タグ付けを対象に複数の選択肢からの選択と，SNS や企業に新たに参加した人の振る舞いを対象に一旦選択した選択肢からの離脱，2 種類の行為選択のモデル化と分析を行なう．これにより，競争的な環境においても人から選択されるための，あるいは選択され続けるための効果的な介入方法を明らかにすることが可能となる．今回扱った具体的な研究課題は以下の 2 つである．

複数の選択肢からの選択では，SNS におけるタグ付けを対象に分析を行った．ここでは，人の選択行動をモデル化する上で用いられる Yule–Simon 過程をベースに，タグ同時利用を考慮可能なように拡張を加え分析を行った．Yule–Simon 過程では，新たな種類のタグは生み出され続ける，既存のタグの選択は優先的選択性が働く，という 2 つのメカニズムによりタグ付けを記述する．このような 2 つの単純なメカニズムで，タグ同時選択における，共起の偏りと新規タグの生み出され方をどの程度再現可能であるのかを確かめた．その結果，共起の偏りに関してはタグ共起の階層性を含む，大部分が Yule–Simon 過程により説明可能であることを確かめた．一方で，新規タグの生み出され方に関しては Yule–Simon 過程の持つランダムに生み出されるという点からは外れ，その生み出され方はユーザのタグを付ける動機により異なることを明らかにした．

一旦選択した選択肢からの離脱では，時系列データとして取得される SNS と企業に新規参加する人の振る舞いから，早期離脱を予測するモデルを構築した．再帰的ニューラルネットワークを基に新たなモデルを構築し，その有効性を代表的な機械学習手法と比較することで確かめた．更に特徴量の分析を行い，SNS の場合には最初期のフォローやライクといった自分自身の振る舞いが早期離脱に対して与える影響が大きいことを明らかにした．また企業の場合には，採用チャンネルと性別が早期離脱に対して与える影響は少ないことを明らかにした．

目 次

図 目 次

表 目 次

第1章　はじめに

World Wide Web (Web) の誕生により，多種多様な情報に対して素早くアクセスすることが可能となった．これにより，様々な情報の比較を行った上で選択を行うことが一般的になりつつある．Web 上に共有される多数の情報，つまり選択肢の中から自由な選択が可能となったことで，それぞれの選択肢はこれまで以上に競争的な環境にさらされている．ソーシャル・ネットワーキング・サービス (SNS) で用いられるタグや SNS あるいは企業は競争的な環境にさらされている選択肢の例であり，様々な選択肢との比較を通して人々から選択される．実際に，SNS 内で人々から付けられるタグを見ると，多種多様なタグが選択されている事がわかる．また，SNS という単語をキーワードに Web 検索を行うと，使い勝手や機能が少しずつ異なる，多種多様な SNS が検索にヒットする．更に，求人という単語をキーワードに Web 検索を行うと，同じような業種，あるいは規模の企業の採用ページが多数検索にヒットする．

SNS や企業の場合は特に，それぞれの選択肢の継続的な発展に対して，そのコミュニティを選択した人々の果たす役割は大きい．例えば SNS の場合には，人と人とをつなげるサービスであるという性質上ある程度多くの人がそのサービスに参加し，継続的にサービスを利用していることに価値がある．また．企業の提供するサービスは基本的には複数の従業員の手で提供される，このため，それぞれの選択肢は人々から選択される，継続的に選択され続ける，という 2 種類の行為選択に対する介入を目的に様々な取り組みが行われ，それらの取り組みに対する注目も年々高まっている．実際に，それぞれの選択肢が選択されるために行う活動に広告活動があり，広告費の年間推移を見てみると，2011 年の東日本大震災の発生した年を除いて年々増加を続けている [1]．また，一旦獲得した参加者の継続利用を促すためのコストである，従業員支援プログラム (例えば，コンサルタティング，従業維持など) に対してかけられるコストも年々増加を続けている [2]．

このような競争的な環境で，タグあるいは SNS や企業といった選択肢が選択される，あるいは選択され続けるためには，複数の選択肢からの選択と，一旦選択した選択肢からの離脱という，2 つの側面から人の行為選択の詳細なメカニズム明らかとする必要がある．これを明らかにすることは，競争的な環境においても人々から選択されるための，あるいは人々の継続利用を促すための，コストを抑えつつも効果の高い介入方法を検討する上で役に立つことが考えられる．

人の行為選択のメカニズムを理解するための古典的かつ典型的な手法に少数の人を対象としたランダムサンプリングによるサーベイ調査がある [50]．これは，社会科学において用いられる典型的な調査方法ではあるものの，この方法で取得されるデータは一般的には以下のような欠点を持つ．ある時刻における人の行動を反映したデータであり，時系列データとして

取得されることは稀である．このため，人の行為選択を時間的に分析することは難しい．また，まれな振る舞いを示す人をサーベイ調査で集中的に集めることは困難である．このため，早期離脱のようなまれな振る舞いに対する分析も難しい．このような理由から，時間的な側面から，あるいはまれな行為選択に対する分析は十分に行われているとは言えない．上記の 2 つの視点から人の行動や社会現象の分析を行なうことは，人の行為選択を理解する上でまだ手つかずの重要な研究課題が多数残されている．

行動ログは，個人あるいは集団による人の行為選択の結果を反映したデータであることから，人の行為選択の詳細なメカニズムを明らかにする上で有益なデータであり，人の行動ログから意味のある情報を抽出し分析を可能とすることは人の行為選択に見られる詳細なメカニズムを明らかにすることに繋がる．例えば SNS では，多数のユーザがサービスに参加し投稿や他のユーザとのやりとりがサーバ上にタイムスタンプとともに保存されている．このため，SNS における他のユーザとのやり取りは，交友関係や人の振る舞いを反映したデータであると言える．取得される行動ログは Web の世界だけに限ったものではない．実世界における人々の行動ログの取得，及び記録も進んでいる．企業における人材管理は電子的に行なわれつつある．いつ出社しいつ退勤したのかという勤怠情報は，勤務場所に設置された端末に個人を識別する ID カードを出勤時と退勤時にをかざすことで記録される．これは，従業員の属性情報 (例えば，年齢，性別，役割) と結び付けられ，デジタルデータとしてそれぞれ管理されている．企業における勤怠情報はどのように企業と関わっていくのかという，人の企業との関わり方を反映したデータである．

行動ログは一般的にタイムスタンプとともに保存される．また，多数の人々の集合からなるデータでありそこにはまれな行動のログも当然含まれている．このような大規模な行動ログの持つ特性により，社会科学では扱われてこなかった時間的あるいはまれな行動に対する分析を行うことが可能になりつつある．このような大規模なデータとその分析を可能とする手法を用いて，これまで以上の広さと深さから人の行動や社会現象を明らかにすることを目的に新たな研究分野が生まれた．これは計算社会科学と呼ばれ近年注目を集めている．計算社会科学は，社会科学において明らかとなっている人の行動や社会現象において観測される知見の補強や新たな理論の構築を可能とする．

そこで本研究では，人の行為選択の詳細なメカニズム明らかにすることを目的に，タグや SNS あるいは企業の選択に関する行動ログを用いて，人の行為選択のモデル化と予測を行う．本研究では行為選択の中でも特に，複数の選択肢からの選択と，一旦選択した選択肢からの離脱という，2 つの行為選択を対象にモデル化と分析を行なう．まずはじめに，SNS におけるタグ付けデータを用いることで，複数の選択肢（タグ）の中から人はどの様なメカニズムで選択を行なうのかを明らかにする．続く研究課題では，一旦選択した選択肢の利用をどの様にやめてしまうのか，SNS あるいは企業に新たに参加した人の振る舞いから明らかにする．これにより競争的な環境においても人から選択されるための，あるいは選択され続けるための効果的な介入方法を明らかにすることが可能となる．今回扱う具体的な研究課題は以下の 2 つである．

- SNS におけるタグ付けデータを利用して，人の同時選択メカニズムをモデル化と分析に

より明らかにする．

- SNS，あるいは企業に新たに参加した人のそれぞれのコミュニティ内での振る舞いから，コミュニティからの早期離脱を予測するモデルの構築と分析により人の早期離脱メカニズムを明らかにする．

1 つめの研究課題では，SNS におけるタグ付けデータを用いることで人の同時選択メカニズムを明らかにする．SNS 上に投稿されるコンテンツは投稿を行なうユーザ自身の手で自由に選択される複数個のタグが付与されることで管理されている．この時に付与されるタグはその後の情報検索に利用される．このような，タグ付けにより投稿の管理を行うようなシステムはソーシャルタギングシステム (STS) と呼ばれる．STS ではサービスを利用するすべての人が自由に分類のためのタグを付与できる点が図書館のような中央集権的な管理手法とは大きく異なる．また，同様の情報であっても同じタグが付けられるとは限らず，ユーザそれぞれが独自に持つ語彙や好み，他者への共感やサービスの雰囲気，あるいは何のためにタグを付けるのかといった動機に大きく依存してタグが付けられる．

このようなタグ付け行為は一種の集団現象であり，タグ付けデータを用いることで人の選択行動に見られる共通の法則を明らかにする試みが広く行われている．タグ付けデータを用いて人の選択行動に見られる共通の法則を明らかにする試みでは，大きく分けて以下の 2 つのことが議論されている．

- ユーザはいつ・どのように新たな種類のタグ (新規タグ) を生み出しているのか．
- 新規タグは出現後どのような使われ方を見せるのか．

これまでにわかっていることは，新規タグは常に生み出され続けるという進化可能性があるということ．もう 1 つはこれまでに使われた回数の多いタグほど使われる回数が多くなる，優先的選択性あるいは正のフィードバックがあるということである [10, 8]．実際に，タグの総数とタグの種類数の関係がベキ分布するという Heaps 則や，タグ選択頻度の順位と選択頻度の関係がベキ分布するという Zipf 則といった統計的な振る舞いが説明できることがわかっている [10, 8]．これらの 2 つの性質は STS を含む人の選択行動一般に広く見られる統計則である．このため，人のタグ選択のモデル化や分析において得られた知見は，より広い意味での人の選択行動においても適用可能であることが考えられる．

このような人のタグつけ行動にみられる，2 つのメカニズムを説明することが可能なモデルに Yule–Simon 過程がある．Yule–Simon 過程はもともと，生物の属の増加構造がベキ分布になるという性質を説明するために Yule により提案された [65, 70]．その後 Simon により離散的なモデルへと修正され，生物以外のシステムで広く観測可能なベキ分布を説明することにも応用できることが示されたモデルである [55]．Yule–Simon 過程では新規タグはポアソン的に一定の確率で生み出され，既存のタグの選択は優先的選択性に基づく．その結果として，Yule–Simon 過程により上記で述べた Heaps 則や Zipf 則といったタグ付け行動をよく再現する．また Yule–Simon 過程は Barabási–Albert モデルと呼ばれる代表的なネットワークモデルのより一般的なモデルとして知られている [54]．

STS では 1 つのコンテンツ (例えば，写真，テキスト) に対して複数のタグが同時に付けられることが一般的である．一方で，Yule–Simon 過程の枠組みではタグは 1 つずつ独立に付けられることからタグの同時利用は考慮されない．同時につけられるタグ同士は意味的なつながりをもち，共起には偏りが生じることが考えられる．タグ同時利用に見られる共起の偏りを利用して，タグの持つ意味的な階層構造の復元やタグの推薦が可能となることから，タグ共起の偏りが持つ意味は大きい．またこのような同時選択の偏りは，タグ以外の人の選択行動においても頻繁に生じるものである．例えば，実際の購買行動ではいくつかの商品が買い物かごに入れられ同時に購入されている．特に，おむつを買う人はビールも同時に購入しやすいという，米国小売りチェーン店において見られた商品の共起例は有名である [64].

そこで，タグ同時選択のメカニズムを明らかにするために，Yule–Simon 過程に拡張を加えタグの同時選択を考慮可能なモデルを構築する．これにより，実データの持つタグ同時選択の振る舞いを Yule–Simon 過程の枠組みでどの程度再現可能であるのかを確かめることが可能となる．ここでは特に，2 つの視点からタグ同時選択の分析を行う．タグ選択における共起の偏りがどの程度 Yule–Simon 過程の枠組みで説明可能なのか，モデルと実データの共起の偏りに着目した分析から明らかにする．新規タグは突然変異のようなランダムな振る舞いにより一定の確率で引き起こされると仮定されているが，タグの同時利用において新規タグがどのように生み出されているのかを明らかにする．

2 つめの研究課題では，SNS，あるいは企業に新たに参加した人のコミュニティ内での振る舞いから，コミュニティからの早期離脱を予測するモデルの構築と分析により，人の早期離脱メカニズムを明らかにする．

SNS では，新規ユーザの多くがサービスから早期離脱していることがこれまでに知られている [3, 14]．特に近年では，多種多様な SNS が存在していることから人々はより早期離脱しやすい状況にある．多様な SNS の存在はすでに参加している SNS に固執する必要性を減じさせ，人々はより気軽に参加している SNS から離脱を選択することを可能とする．また，人々の同時に参加する SNS の数は増えた一方で，それぞれの SNS に対して費やすことが可能な時間はほとんど変化していない．1 日が 24 時間という点は依然として変わらず，我々の情報処理能力もほとんど変化していない．その結果，自分が興味を持った SNS に新たに参加する場合にはすでに属している SNS に費やす時間を削る必要がある．その時間がとれなくなってしまうことで人々は既存の SNS から離脱する．

早期離脱者が予め分類可能となる，あるいは早期離脱の要因が明らかとなることは，新規ユーザのサービス継続を促す効果的な介入方法の検討につながることから重要な課題である．これまでの研究では，SNS における離脱者の予測や離脱要因の分析にはロジスティック回帰やランダムフォレストといった機械学習による手法が広く利用されてきた [3, 14]．これらの機械学習手法を用いることで，高い精度で離脱者の分類や要因分析が可能であることがこれまでに知られている．また，SNS への新規参加者の離脱率は高いことがこれまでに知られている．このため離脱者の中でも，特に早期離脱者に着目して予測や要因分析を行なった研究もいくつか存在する．

早期離脱するユーザとそうでないユーザを識別する上で，それぞれのユーザの時間的な振

る舞いの変化を考慮することでより精度良く両者の分類が可能となることが考えられる．例えばSNSにおける新規ユーザを例にある期間で同一の投稿数の2人のユーザを考える．あるユーザはサービスに参加した直後だけ投稿を行うユーザである．もう1人のユーザは観測期間を通してまんべんなく投稿を行うユーザである．直感的には，前者のユーザはサービスに参加した直後に離脱している可能性が高いことが考えられる，反対に，継続的にサービスの利用を行っている後者のユーザの離脱率は低いことが考えられる．この様な仮設に基づき，Yangらは入力される時系列間の相関を考慮可能な再帰的なニューラルネットワーク (RNN) を用いることで，時系列間の相関を考慮しない機械学習手法よりも高い精度でSNSにおける早期離脱者の予測が可能であることを示している [68].

時系列間の相関を考慮可能なRNNを利用することで，これまでよりも高い精度で早期離脱者の予測を行うことが可能である．しかしながら，RNNでは明示的には考慮されていない情報がある．入力される時系列の各時刻ごとの新規ユーザの早期離脱予測に対する影響度の違いである．これまでの研究から，SNS内外での人の振る舞いは間欠的な振る舞いを示すことがわかっている [4, 67]．これは，ある時刻において頻繁に活動を行い，それ以外の時刻ではほとんど活動を行わないことを意味する．例えば，SNSに参加するユーザの場合にも投稿や他者へのいいねは，ユーザのログイン時に短期間でまとまって行われていると考えることが自然である．新規ユーザも同様の振る舞いを示すと仮定すると，早期離脱者を予測する上で影響の大きな時刻とそうでない時刻が存在することが考えられる．このような時間毎の早期離脱への影響度の違いを明示的に考慮することで，より高い精度で早期離脱者の予測が可能となることが考えられる．

そこで，SNSにおける新規ユーザの時間的な振る舞いを入力として，各時刻ごとの影響度の違いを考慮して早期離脱者の予測を行う新たなモデルを提案する．更に，提案モデルを用いることで早期離脱者の分類へ与える影響の違いを時間的，特徴量の面から明らかにする．最後に，特徴量の分析結果を踏まえたSNSにおける早期離脱の防止に効果的な介入方法を検討する．

更に，SNSにおける早期離脱ユーザの予測において得られた知見やモデルを利用することで，飲食チェーン店における早期離脱従業員の予測と要因分析を行なう．飲食チェーン店の場合にもSNSと同様に早期離脱率は高いことがこれまでに知られている [66]．しかしながら，飲食チェーン店を対象として早期離脱者の予測を行った研究は，離脱者に着目した研究の中でも少ない．これには以下の2つの理由が考えられる．早期離脱者のある程度まとまったデータの入手が難しく，そもそも分析することができなかった．採用時のフィルタリングにより，早期離脱可能性の高い候補者を雇用しないという戦略が取られてきた [57]．近年になり大規模な行動ログの取得が可能となったことで，ある程度まとまった早期離脱者のデータを取得することが現実的になりつつある．また，従業員不足となっている飲食チェーン店の場合には採用時のフィルタリングは現実的ではない．この様な理由から，飲食チェーン店における従業員の早期離脱予測も，SNSにおける新規ユーザの離脱予測と同様に重要な研究課題であると言える．

本論文は全部で7章からなる．1章は序論である．ここでは本論文の目的とその意義につい

て詳細に述べた．2 章は関連研究である．これまでに行なわれてきた研究を俯瞰することで，本研究との差異を強調する．ここでは特に，2 種類の行為選択に関連する研究を重点的に紹介する．つまり，SNS におけるタグ選択メカニズムに関する研究と，企業あるいは SNS に新たに参加する人の早期離脱に関する研究である．3 章では，SNS におけるタグの同時選択メカニズムに着目して，Yule–Simon 過程を土台にタグの同時選択を考慮可能なモデルを提案する．更に，タグ共起の偏りの観点からどの点で実データを再現するのかを確かめる．これにはタグ共起ネットワークを構築し，実データとモデルの差異を見ることで確かめる．4 章ではタグ同時選択のメカニズムの中でも特に，新規タグの選ばれ方に着目して分析を行なう．また，STS を採用するサービスにおけるユーザのタグ付けの動機がどのように新規タグの生み出され方に影響を及ぼすのかを明らかにする．5 章は，SNS における新規ユーザの早期離脱予測に関する研究結果について述べた章である．6 章では飲食チェーン店における従業員の早期離脱予測に関する研究結果について述べる．7 章はまとめであり，本論文の成果を総括し結論についてを述べる．更に，今後の展望及び課題について述べ結びとする．

第2章　関連研究

ここでは，これまでに行なわれてきた研究を俯瞰することでそれぞれの章の導入を容易とするとともに，本研究との差異を強調する．ここでは特に，以下の 2 つのテーマに関連する研究を重点的に紹介する．SNS におけるタグ選択メカニズムに関する研究と，企業あるいは SNS に新たに参加する人の早期離脱に関する研究である．

2.1　タグ選択のモデル化と分析

Web の誕生により膨大な情報が Web 上で保存，及び共有されるようになりつつある．膨大な情報の管理，検索を行なうために，Web の世界では従来の中央集権的な情報管理手法から，新たな情報管理手法が用いられるようになった．ボトムアップの情報管理手法であり，その情報を投稿した，あるいはその情報に興味があるユーザ自身の手で情報の管理を行う，という情報管理手法である．従来の中央集権的な情報管理を行う典型的な例として，図書館がある．図書館では司書が管理される図書に対してラベル付けすることで情報管理を行う．一方で，このような従来の中央集権的な情報管理手法では，膨大な情報が管理され，新たに追加されていくような Web 上の情報管理には適しているとは言えない．

Instagram，YouTube，Twitter，Facebook などの Web 上のオンラインコンテンツ共有サービスでは，ユーザが任意の文字列を付与することによって投稿されるコンテンツの管理を行う Social Tagging System (STS) が採用されている [27]．例えば，Delicious は，Web ページをタグ付けにより，管理共有を目的としたサービスであり，ブックマークした URL に対してサービスの利用者がタグ付けを行う．同様に Flickr や Instagram，RoomClip では写真，YouTube では動画，そして Twitter や Facebook では文章や写真へタグを付与する．これにより投稿される大量のコンテンツの検索・管理を可能としている．

STS を採用するシステムは，そのタグを付ける機能により，2 つの種類により細かく分類することが可能である [63]．それぞれ，Broad tagging system と Narrow tagging system と呼ばれる．Broad tagging system は各情報に対して様々なユーザがタグを付ける事が可能な仕組みを採用している．それとは対象的に，Narrow tagging system では，投稿された情報に対してタグを付与する権利があるのは，その情報を投稿したユーザだけである．Broad tagging system を採用する典型的なサービスには，Delicious がある．Flickr や Instagram，RoomClip などのサービスは Narrow tagging system を採用するサービスである．

様々なユーザが任意のタグを付与することが可能であるという STS の仕組み上，投稿されるコンテンツと付与されるタグの間に明確な規則は存在しない．このため，タグ付け行為は，

人の選択行動を捉えたデータであると言える．このため，タグ付けデータを用いることで，人の選択行動を明らかにすることに注目が集まっている．それだけでなく，新しい種類のタグがどの程度生み出され，どのように使われているのかといったタグの振る舞いを知ることは，効率的なデータベースの設計や情報ナビゲーションを実現するために重要な課題でもある．このような理由から，STS におけるタグ付けを対象として，これまでにいくつかのモデル化や分析が行われている．

オンラインサービスにおけるタグの振る舞いをモデル化する最初の試みには，Golder と Huberman による”Polya urn”モデルがある [22]．彼らは，タグの選択には早い時期に作成され，選択回数の多いタグほどより選ばれやすいという「優先的選択性」があることを発見し，それを説明するモデルを提案した．Polya urn モデルでは，いくつかの種類のタグがそれぞれ同数個壺の中に存在する状況を仮定する．その壺の中からランダムに 1 つを取り出し，取り出されたタグが選択される．選択されたタグは，壺の中に再び戻される．このとき，選択された種類のタグをもう 1 つ生み出し，合計で 2 つのタグ (選択されたタグと新たに追加された同種類のタグ) が壺の中に戻される．同様のやり方で，再び壺の中からランダムに 1 つのタグを選択していく．このような選択されたタグと同一種類のタグを追加して壺の中に戻す仕組みにより Polya urn モデルでは，これまでのタグの選択回数に応じて各タグの選択確率は変化する．このようなタグの選択回数に応じてかけられる，正のフィードバックは優先的選択と呼ばれる．

Polya urn モデルによりタグ選択の偏りを記述することが可能となる一方で，タグの種類数の増え方に関しては実際のタグ付けを模したものであるとは言えない．つまり，Polya urn モデルでは，タグの種類数は最初に壺の中に存在するタグの種類数から変化せず，時間的な種類数の増加は起こらない．一方で実際のタグ付けを見てみると，様々な種類のタグが新たに生み出され選択されている．

Catutto らは，タグ使用頻度の順位と頻度数のみならず，新しい種類のタグの増加する傾向がベキ分布に従うことを示した [10, 8]．彼らは，これらのベキ分布を Yule–Simon 過程を導入することで説明した．Yule–Simon 過程では，優先的選択性と，新しい種類のタグは常に生み出され続けるという 2 つの仕組みによりタグ選択を記述するモデルである．Yule–Simon 過程はもともと，生物の属の増加構造がベキ分布になるという性質を説明するために Yule により使用された [65, 70]．その後 Simon により離散的なモデルへと修正され，生物以外のシステムで広く観測可能なベキ分布を説明することにも応用できることが示されたモデルである [55]．Yule–Simon 過程に改良を加えることで，より実際のタグ付けに近い構造や特徴を表現する提案もこれまでに多数おこなわれている．例えば，Halpin らはタグの語用的な特徴を考慮する改良を加えたモデルを提案している [28]．また，Cattuto らは，使われたタグの数の増加に従って新たなタグの生成確率が減少することを考慮する改良や [8]，最近使われたタグほど使われる確率が上がるような記憶を考慮する改良を加えている [10]．このように Yule–Simon 過程はタグの振る舞いをモデル化する手法として広く用いられている，代表的な手法である．Yule–Simon 過程は古典的なモデルであるものの，人の選択行動に見られる社会現象を記述する一般的なモデルとしても知られている．例えば，代表的なネットワークモデルである Barabáshi–Albert

モデルは Yule–Simon 過程のサブクラスであると考えることが可能である．

STS では一般的に 1 つの投稿に対して，複数のタグが同時に付与される．このため，投稿に対して同時に付与されるタグ間の共起関係は情報検索において有益な情報となる．例えば，ユーザに適切なタグを提案するためにあらかじめ共起しやすいタグを求めておくことで，あるタグに対してより共起しやすいタグを推薦できることがこれまでに知られている [53]．また，情報を効率良く分類するために，情報を階層的に整理することがしばしば有効であるが，タグの共起関係を利用することでタグの階層構造を自動的に抽出することが可能であることがこれまでに知られている [60]．このことから，タグ共起のモデル化がこれまでに行われている．Cattuto らは出現するタグの可能性空間としてネットワークを構築し，その上をランダムウォークすることで，人のタグ同時選択をモデル化した．タグの可能性空間のネットワークのノードはタグの種類に対応し，エッジは各タグ間の意味的なつながりを表す．これにより，タグ共起の階層性を含む様々な統計値を再現可能であることが確かめられている [9]．一方で，新たな種類のタグは生み出され続ける，既存のタグは優先的選択性に基づいて選択されるという Yule–Simon 過程もつシンプルなメカニズムで，どの程度実データのタグ共起を再現可能であるのかは明らかではない．また，Yule–Simon 過程をそのまま利用するだけでは，個々のタグはそれぞれ独立に生成または選択されるため，そのままではタグ共起の分析も行うことはできない．

Yule–Simon 過程はタグ付けにおいてみられる Zipf 則や Heaps 則といった，マクロな統計的特徴をうまく説明することがこれまでによく知られている [10, 8]．Yule–Simon 過程では，新規タグの生成確率が一定，あるいは時間的に減衰するような関数であることを仮定している．つまり，Yule–Simon 過程は，新たな種類のタグが作成されるメカニズムはランダムな確率により記述され，そのメカニズムについては何も述べていない．Tria らは Kauffman による隣接可能性を用いることで，新規タグ間の相関関係の存在について議論した [35]．Tria らは，新規タグ間には相関が存在し，新規タグの生成は他の新規タグの生成を促す役割を果たすことを明らかにした [62]．よりミクロなスケールで新規タグの生成を見た場合には，同様に単純なポアソン過程のようなランダムなプロセスからは逸脱することが考えられる．例えば，同じ投稿内で新規タグが作成される場合には，同時に選択される，あるいは新たに生成される新規タグ間には何らかの相関が生じることが考えられる．しかしながら，同じ投稿内でどのように新規タグが生成されていくのかという点はこれまでに明らかにされていない．

2.2 コミュニティからの離脱

2.2.1 SNS に参加するユーザの早期離脱予測

サービスからのユーザの離脱予測や離脱の要因分析は比較的長い歴史を持ち，これまでに様々なサービスを対象に研究が行われてきた [14]．古いものには，無線通信産業や銀行等のサービスを利用するユーザを対象とした研究がある．これらのサービスに参加するためには煩雑な登録手続きを経る必要がある．煩雑な登録手続きを経て登録を行ったユーザは，そのサービスを利用することに対して高い動機を持つ可能性が高い．このため，これらのサービ

スを利用するユーザはある程度サービスを継続して利用し，新規ユーザの早期離脱が問題になることは少ない．

一方 SNS では，これまで扱われてきた実世界でのサービスに比べて，参加への心理的あるいは手続き的な障壁は低く，SNS に参加するためにはメールアドレスの登録だけで済むことが多い．また，離脱した場合のペナルティも存在しない．このため，SNS における新規ユーザの早期離脱率は非常に高いことがわかっている．例えば，オンラインのディスカッションサイトである，Usenet newsgroups では 7 割近くの新規ユーザは 1 回目のポストを行った後にサービスから離脱している [3]．また，オンラインの Q&A サービスである，Yahoo! Answers では 5 割ほどの新規ユーザが早期離脱している [14]．

これまでの研究から，より個人的な趣味嗜好を捉えた情報を新規ユーザに対して提示することで早期離脱率が減少することが確かめられている．例えば，ある新規ユーザの他のサービスの利用履歴から，自分自身のプロフィール画面の候補や友達の推薦を初期に行うことで，早期離脱率が減少することがこれまでに知られている [19]．また，自分のした投稿に対して，類似度の高いユーザから何らかの反応があることで，早期離脱確率が減少する事がわかっている [11]．ここで，早期離脱可能性の高いユーザと低いユーザを正確に区別可能となることで，各ユーザごとに早期離脱に応じて介入方法を変えることが可能となる．例えば，新規離脱可能性の高いユーザに対して，よりパーソナライズされた人の手による高コストの介入を行い，そうでないユーザに対しては機械的な低コストの介入を行うというようなやり方である．

SNS における新規ユーザを対象に，サービスからの早期離脱を予測するモデルに近年注目が集まりつつある．Dror らは Yahoo! Answers における新規ユーザを対象に複数の機械学習モデルを用いて早期離脱予測を行い，Random Forest が高い性能を発揮することを示した [14]．Pudipeddi らは Stack overflow を対象に早期離脱ユーザと，サービスの利用をある程度継続したベテランユーザの離脱予測を行い，それぞれのユーザ群の予測に対して決定木が高い識別性能であることを示した [45]．

近年になり，深層学習を用いてユーザの早期離脱を予測するモデルも提案されている．Yang らは時系列情報から新規ユーザの早期離脱予測を行うために，再帰的ニューラルネットワーク (RNN) をベースにしたモデルを提案している [39, 68]．RNN は再帰的な構造により，時間的な相関を学習可能なモデルとして知られている．例えば，動画ではフレーム間の関係性が，文章では単語間の関係性が何らかの意味を持つと考えるのが自然であり，動画分類や音声認識などの分野において，RNN を利用した手法が最も高い性能を発揮することがこれまでに知られている [15, 24]．Yang らは同時に，複数の RNN を学習に利用し各モデルからの出力をアテンションにより調整することで，時系列情報とユーザの個性を用いた早期離脱者の予測モデルも提案している [68]．Yang らの研究では，アテンションは各モデルの出力に対して考慮されており，時刻毎の早期離脱への影響度の違いを考慮することを目的とした，各時刻での出力に対してかけられるものでは無い．筆者の知る限り，各時刻の予測に対する影響度の違いを考慮して SNS における早期離脱者の分類を行うモデルは存在しない．

オンライン空間内外における人の振る舞いは間欠泉的であることがこれまでに広く知られている [4]．つまり，ある時刻にはユーザの活動が集中し，それ以外の多くの時刻では活動が

ほとんど行われていないというものである．SNS におけるユーザの振る舞いも間欠泉的な振る舞いを示すことが確かめられている [67]．実際に我々が SNS を利用する状況を考えると，サービスにログインした時に複数のいいねやコメントなどを連続して行い，ログアウトすることが一般的である．また，ユーザがサービスに参加してからの最初期の振る舞いがユーザの長期的な振る舞いに影響を及ぼす可能性がこれまでに示唆されている．例えば，Pal らは SNS におけるコミュニティ内で将来的に熟練したユーザになる可能性の高いユーザをはじめの数週間のふるまいから予測できることを示した [42]．更に，Karumur らは MovieLens に新たに参加したユーザの最初のセッションにおけるアクションの多様さが，早期離脱率に大きな影響を及ぼすことを示している [34]．このことは，時刻毎の予測への影響度の違いを考慮した学習が可能となることで，早期離脱者の予測精度が向上する可能性があることを表している．また，時刻毎の影響度の違いが明らかになることで，どの時刻に介入を行うべきかという具体的な介入時刻の決定に役に立つものとなることが考えられる．

2.2.2 企業で働く従業員の早期離脱予測

従業員の離脱に着目した研究はこれまでに多数行われている．予め離脱する従業員が明らかとなることで，従業員の離脱を見越して，新規従業員の募集や離脱可能性の高い従業員に対する面談といった介入が可能となる．このため予測を通した介入を目的に，さまざまな機械学習手法がこれまでに提案されている．ロジスティック回帰，ランダムフォレスト，サポートベクトルマシンなどの分類問題を扱う代表的な機械学習手法は，従業員の離脱を高い精度で予測可能であることがこれまでに知られている．また，多層ニューラルネットワークなどの深層学習モデルを用いた離脱者の予測も近年になり行われつつある [57].

これまでの研究の多くはすべての従業員を対象としており，新規従業員の早期離脱に着目した研究は少ない．Strohmeier らのレビュー論文によると人材配置 (離脱予測と採用の親カテゴリ) に関する研究の 1/3 を採用に関する研究が占めており，離脱可能性の高い人材はとらないという戦略のもと研究が行われてきたと考えるのが自然であり，このことが新規従業員の早期離脱に着目した研究は少ない理由として考えられる [57]．しかしながら，少子高齢化に伴う労働人口の減少により新たな人材の確保が難しくなりつつある．特に，飲食業界などでは離脱率が高い [66]．このような業種では採用時のフィルタリングは現実的ではなく，サービスを提供するためには，応募してきた候補者をなるべく多く採用する必要がある．このため，一旦採用を行った新規従業員の離脱を予測することで，早期離脱を防ぐための効果的な介入を行う必要がある．

近年になり，人材不足の業界を対象として，従業員の早期離脱に着目した研究もこれまでにいくつか行われている．Geun らは外国人看護師の早期離脱率が高い点に着目し，新たに雇用された外国人看護師を対象にアンケート調査を行い，早期離脱につながる要因を明らかにした [21]．また，Peg らは飲食業界を対象に分析を行い，早期離脱と盗難の関係を明らかにしている [59]．一方で，これらの研究では早期離脱要因の推定に着目したものであり，予測を意図したものではない．また，両者の研究は時系列情報を利用したものでは無いという点で，

今回の研究とは異なる．我々の知る限りでは，新規従業員の勤怠時系列と属性情報から新規従業員の早期離脱を予測するモデルは存在しない．

飲食業における離脱要因を明らかにするために，これまでにいくつかの研究が行われてきた．Cantrell らは，飲食業における離脱要因は，給料や勤怠時間とシフトの自由度が低いことであることを示した [7]．また，Mcfillen らは，仕事の満足感と貢献が離脱に与える影響が大きいことを示した [38]．勤怠時系列が持つ情報は勤務日時だけではなく，誰といつ一緒にどこで働いたかといった各従業員の詳細な働き方や，企業あるいは仕事との関わり方が記録されている．このため，勤怠時系列を利用することで，従業員の早期離脱予測をこれまでよりも高い精度で予測可能となることが考えられる．

第3章　タグ同時選択のモデル化とタグ共起の観点からの分析

SNS 上に投稿されるコンテンツは，投稿を行なうユーザ自身の手で自由に選択された複数個のタグが付与されることで管理されている．STS では，サービスを利用するすべての人が自由に分類のためのタグを付与できる．このため，同様の情報であっても同じタグが付けられるとは限らず，ユーザそれぞれが独自に持つ語彙や好み，他者への共感やサービスの雰囲気，あるいは何のためにタグを付けるのかといった動機に大きく依存してタグが付けられる．このようなタグ付け行為は一種の集団現象であるとみなすことが可能である．

タグ付けデータを用いることで，人の選択行動に見られる共通の法則を明らかにする試みが広く行われている．これまでにわかっていることは，新規タグは常に生み出され続けるという進化可能性があるということ．もう 1 つは，これまでに使われた回数の多いタグほど使われる回数が多くなる，優先的選択性あるいは正のフィードバックがあるということである [10, 8]．実際に，タグの総数とタグの種類数の関係がベキ分布するという Heaps 則や，タグ選択頻度の順位と選択頻度の関係がベキ分布するという Zipf 則といった統計的な振る舞いが説明できることがわかっている [10, 8]．

このような人のタグつけ行動にみられる，2 つのメカニズムを説明することが可能なモデルに，Yule–Simon 過程がある．Yule–Simon 過程はもともと，生物の属の増加構造がベキ分布になるという性質を説明するために Yule により使用された [65, 70]．その後 Simon により離散的なモデルへと修正され，生物以外のシステムで広く観測可能なベキ分布を説明することにも応用できることが示されたモデルである [55]．

Yule–Simon 過程が説明するのは新しい種類のタグの増え方と既存のタグの選び方であり，タグは一つずつシーケンシャルに発生する．したがって，一つのコンテンツに対してタグが複数使われるような，ソーシャルタギングに典型的なタグの同時利用は考慮されていない．そこで，この章では複数のタグが同時に発生する Yule–Simon 過程を考え，それを Windowed Yule–Simon 過程と呼ぶ．このタグの同時利用を取り込んだ新しいモデルをタグ生成の null モデルとし，実データの振る舞いをどのような側面で再現し，またどのような点でズレが生じるのかをタグ共起の観点から検証する．

はじめに，Yule–Simon 過程の説明を行う．次に，この章で扱う 4 つのデータセットの説明と，タグ選択をモデル化する上で，Yule–Simon 過程を利用する妥当性の検証を行う．その後，新たにタグの同時利用という概念を加えた Windowed Yule–Simon 過程の説明を行う，最後に，STS を採用する SNS である，RoomClip のデータを用いて，実際のデータが持つタグ共起の偏りをどの程度再現可能であるのかを分析により明らかにする，ここでは，実データとモデ

ルによるタグ付けデータから，タグ共起ネットワークを構築し，それぞれのネットワークが持つ相関や階層性の観点から分析を行い比較する．タグ共起ネットワークは各ノードがタグの種類に対応し，エッジが同時利用関係を表すネットワークである．

3.1 Yule–Simon 過程によるタグ付けのモデル化

3.1.1 データセットの説明

図 3.1: RoomClip に実際に投稿される写真とタグのイメージ．RoomClip に投稿されたタグの例として，投稿された写真に対して，ディスプレイ，小物，PC，My Desk の 4 つのタグが付与されている．

まずはじめに，分析に利用したデータセットの説明を行う．ここでは，STS を採用する代表的な 4 つの Web サービス Delicious，Flickr，Instagram，RoomClip [1]から取得した 4 つのタグ付けのデータセットを用いて分析を行った．Delicious において投稿されるコンテンツは Web ページの URL である．このため，ユーザはお気に入りの URL の管理，検索を行うことを目的に，投稿した URL に対してタグ付けを行う．Flickr，Instagram，RoomClip は写真の保管と共有のためのサービスであり，投稿した写真に対してタグ付けを行う．図 3.1 に RoomClip において投稿された写真とタグの例を示す．

それぞれのサービスでは投稿に対してタグをつけて情報の管理を行うという点は共通するものの，Delicious は Flickr，Instagram，RoomClip とはサービスの性質が大きくことなる．Delicious は broad tagging system [63] に基づいており，複数のユーザが同じコンテンツに対し

[1]https://roomclip.jp/

表 3.1: タグ同時選択の分析に利用した，4 つのサービスの基本統計．

Service	Users	Vocabularies	Annotations	Entries	Range
Delicious	532,924	2,481,108	140,126,555	47,257,452	Jan. 2003 - Dec. 2006
Flickr	319,686	1,607,879	112,900,000	28,153,045	Jan. 2004 - Dec. 2005
Instagram	2,110	271,490	8,201,542	1,047,774	Oct. 2010 - Feb. 2014
RoomClip	32,852	194,881	3,141,524	692,459	Apr. 2012 - May. 2015

てタグを付けることが可能である．例えば，あるユーザがある Web ページにタグを付けていたとしても，他のユーザが同一の Web ページに対してタグを付けることが可能な仕組みとなっている．このような場合には，各投稿はユーザ ID と投稿 ID のペアによって識別が行われ，同一のコンテンツではあるもののそれぞれ別の投稿として扱われる．一方で，Flickr，Instagram，RoomClip は Narrow tagging system を採用しており，投稿者であるユーザだけが投稿したコンテンツに対してタグを付けることが可能である．また，Flickr，Instagram，RoomClip，3 つのサービスは投稿される写真に対してタグを付けるという性質は共有するものの，Flickr は Instagram や RoomClip とはサービスの性質が大きくことなる．Flickr では，投稿可能枚数に制限が存在する．一方，Instagram や RoomClip では投稿可能枚数に制限はない．情報共有を目的としたサービスでは，一般的には投稿可能枚数には制限は存在しない．そのため，Flickr ではオンラインの写真保管サービスとしての側面が強いサービスであるといえる．反対に，Instagram や RoomClip ではそのような制限は存在しないため，より他のユーザとの情報共有を目的としたサービスであるといえる．Instagram と RoomClip の間にも微妙な差異が存在する．Instagram ではほぼすべての写真が投稿対象であるが，RoomClip では室内の写真に投稿が制限されている．

4 つのデータセットは以下に述べる方法でそれぞれ取得されたものである．Delicious，Flickr データは，Görlitz らにより取得されたデータであり，これは大規模なスクレイピングにより収集されたものであり，ある期間までの全タグ付けデータを取得したものである [23]．Instagram データは，Ferrara らによってスクレイピングにより収集されたデータである．Ferrara らはサービス内で開催された，特定のイベントに参加したユーザの中からランダムに 2,000 人以上のユーザを選択し，対象としたユーザのサービスに参加してからのすべての投稿に対して付与されたタグを取得した [17]．RoomClip データは，RoomClip Inc. から直接提供を受けたものであり，匿名化されたデータである．収集されたデータは，サービスが生まれてから，今までのすべてのユーザのタグ付けデータを保持している．RoomClip には運営者により定義された代表的な 9 つのタグが存在する．ユーザは投稿を行う際に，これらのタグの中から 1 つを選択する必要がある．このため，前処理として運営者により定義されたタグは分析から除いた．表 3.1 に各データセットの基本統計をのせる．

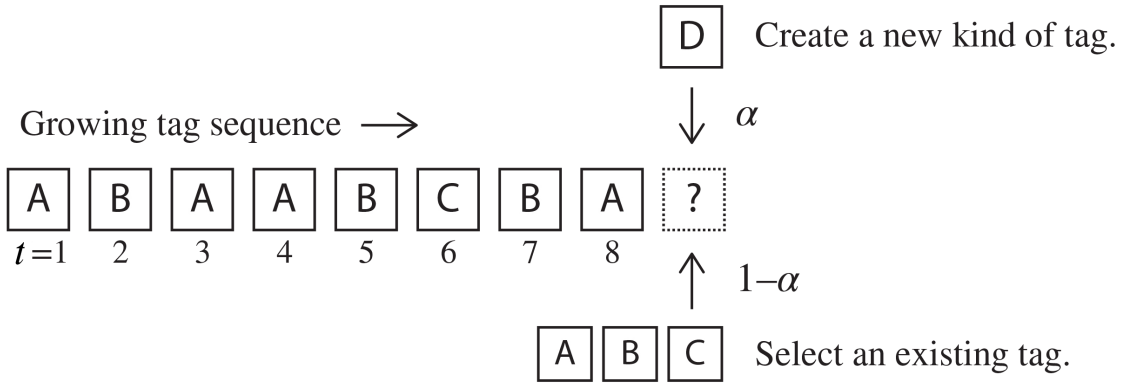


図 3.2: Yule–Simon 過程の概念図．Yule–Simon 過程では 2 つのメカニズムにより人のタグ付け行動をモデル化する．それぞれ，新しい種類のタグは常に生み出され続ける，既存のタグはこれまでの選択回数に応じて選ばれる (優先的選択) というメカニズムである．各試行ごとに確率 α で今までに使用されていない新たな種類のタグを生成する．新しい種類のタグの生成確率 α は，モデルに与えるパラメーターであり，一定あるいは時間減衰する関数として与えられる．それ以外の確率 $(1-\alpha)$ でこれまで使われたタグの中からランダムに選択を行う．確率 $1-\alpha$ で既存のタグが選択される場合，これまでに選択されたタグの中からランダムに 1 つのタグを選択する．つまり，タグ A ($n=4$)，B ($n=3$)，C ($n=1$) がそれぞれ選択される確率は $\frac{4}{8}$，$\frac{3}{8}$，$\frac{1}{8}$ である．

3.1.2 Yule–Simon 過程

ここでは，Yule–Simon 過程の詳細な説明を行う．Yule–Simon 過程は人々の社会的なふるまい全般で観測される，べき分布を表現可能なモデルである．Yule–Simon 過程では，試行ごとに新たな種類のタグを生成，またはこれまでに使われたタグの中から選択を行なう．Yule–Simon 過程における新たな種類のタグの生成確率は α である．それ以外の確率 $1-\alpha$ でこれまでに使われたタグの中からランダムに 1 つのタグを選択する．使われたタグの総数を t．これまでにタグ i が出現した回数を n_i とすると，$t+1$ 回目の試行でこれまでに使われたタグの中からタグ i を選択する確率 $P(i)$ は，

$$P(i) = (1-\alpha)\frac{n_i}{t} \tag{3.1}$$

で与えられる．

Yule–Simon 過程の概念図を図 3.2 に示す．図 3.2 では，[A，A，B，A，B，A，B，C] の合計 $t=8$ の試行が行われている．続く 9 回目の試行において，確率 $1-\alpha$ で既存のタグが選択される場合，タグ A ($n=4$)，B ($n=3$)，C ($n=1$) がそれぞれ選択される確率は $\frac{4}{8}$，$\frac{3}{8}$，$\frac{1}{8}$ となる．実際にこれまでに使用された回数の多いタグほど選ばれる確率が高くなる．このように，Yule–Simon 過程は出現回数 (n) の値が大きいタグほど選択される確率があがる．これは，優先的選択性と呼ばれる．

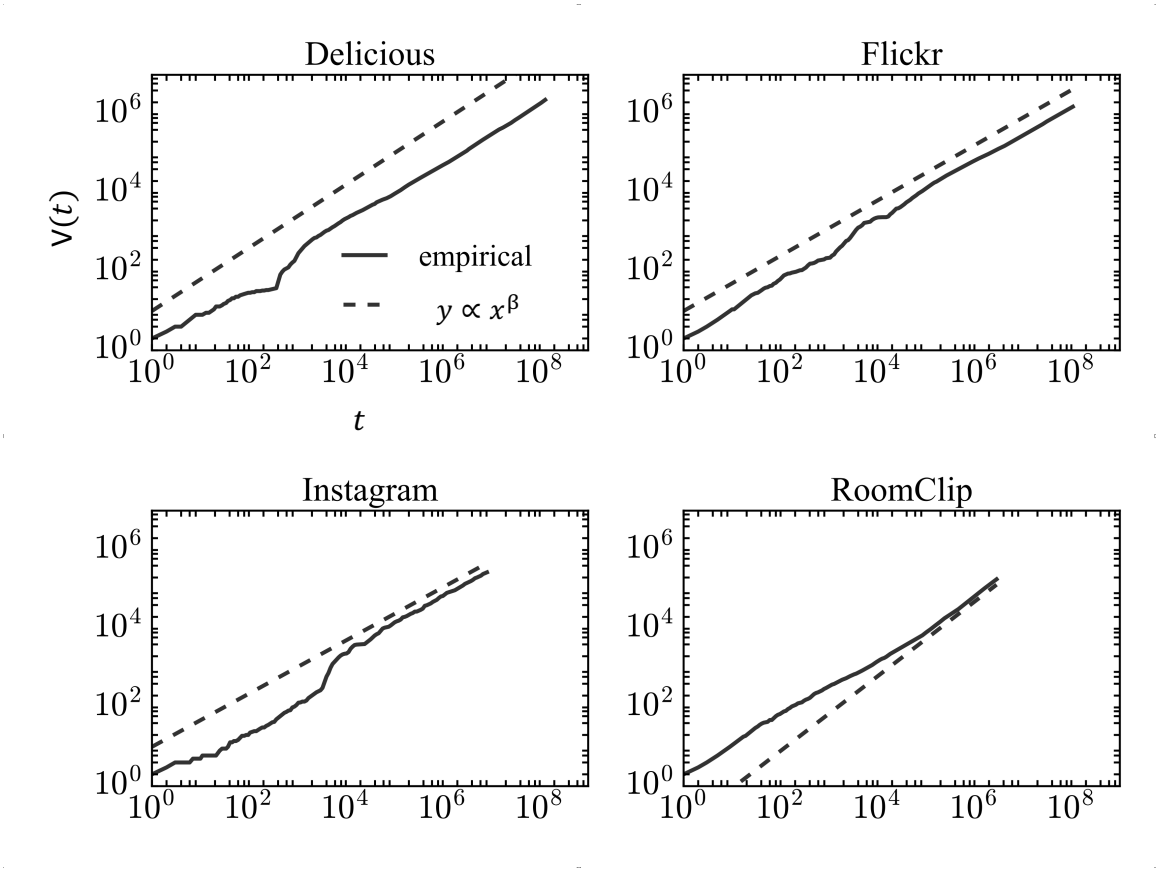


図 3.3: STS を採用する 4 つのサービスにおける Heaps 則の計算結果．横軸がタグの総数 (t) であり，縦軸がタグの種類数 $V(t)$ である．また，点線は $y \propto x^{\beta}$ の補助線である．

3.1.3 Yule–Simon 過程によるモデル化の妥当性の検証

ここでは，今回利用する 4 つのデーセットにおいてみられる，人のタグ選択行動を Yule–Simon 過程によりモデル化することの妥当性を検証する．検証には，新たな種類のタグの生み出され方，既存のタグの選択確率，各タグの選択頻度と選択頻度順にならべた順位との関係，以上の 3 つの視点から確かめる．

新たな種類のタグの生み出され方

まずはじめに，新たな種類のタグの生み出され方が Yule–Simon 過程により記述可能であるのかを確かめる．これまでの研究から，文書において出現した単語の総数と語彙の種類数の関係がべき分布になることが知られている．このように，単語の総数と語彙の総数がべき分布となる法則は Heaps 則と呼ばれる [30]．STS を採用する SNS におけるタグも，同様のベキ分布に従うことがこれまでに報告されている [37].

$V(t)$ を試行 t におけるタグの種類の数とすると，Heaps 則は

$$V(t) \propto t^{\beta} \tag{3.2}$$

と表現される．ここで，β は指数である．Yule–Simon 過程では新たな種類のタグの生成確率は一定，あるいは時間減衰する関数として与えられる．両者の関係がべき分布する場合には Yule–Simon 過程によるモデル化は妥当であると言える．

図 3.3 に STS を採用する 4 つのサービスにおける Heaps 則の計算結果を示す．横軸がタグの総数 (t) であり，縦軸がタグの種類数 $V(t)$ である．また，点線は $y \propto x^{\beta}$ の補助線である．それぞれのサービスでその傾きは異なるものの，4 つのサービスでおおよそべき分布することが明らかとなった．このことから，新規タグの生成という観点に関しては，Yule–Simon 過程によりモデル化することは妥当であるといえる．

既存のタグの選択確率

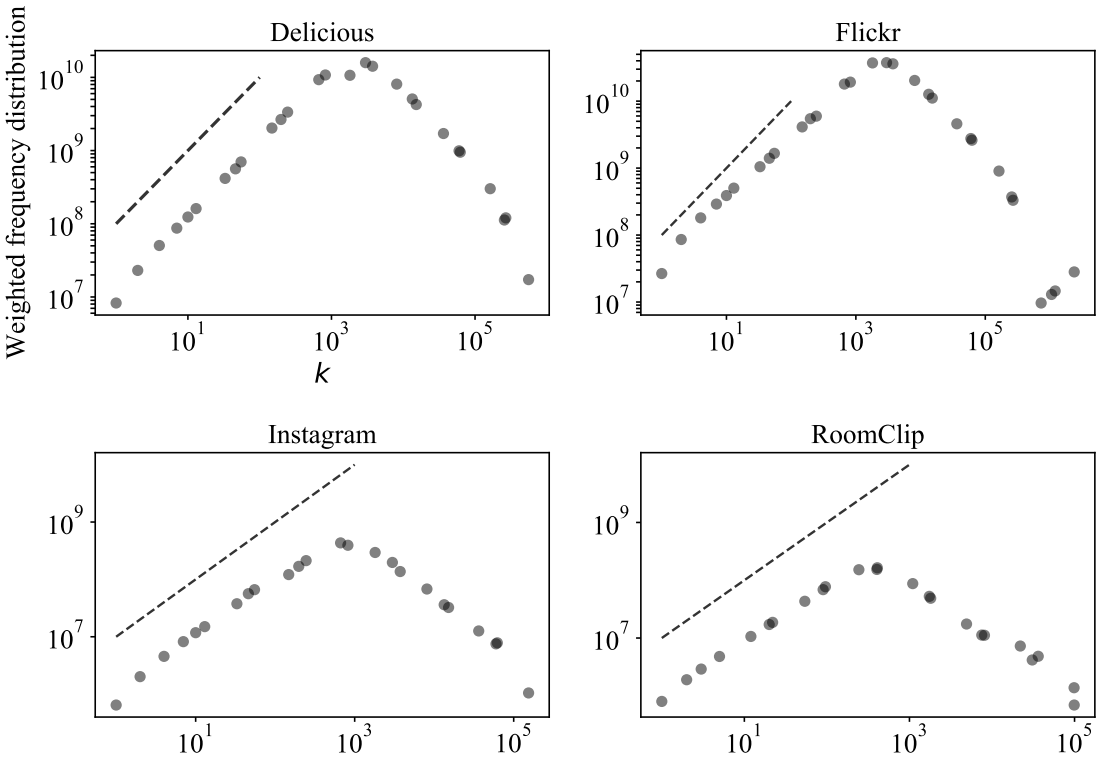


図 3.4: STS を採用する 4 つのサービスにおける重み付きの頻度分布の計算結果．横軸がタグの選択回数であり，縦軸がその選択回数における重み付きの選択頻度である．点線は傾き 1 の補助線である．

つぎに，既存のタグの選ばれ方が Yule–Simon 過程と同様に優先的選択に従っているのかを確かめる．つまり，これまでの選択回数に応じて，その後の選択確率が上昇していくのかを確かめる．実際のタグ選択において優先的選択が働いているのかは，Newman の手法を用いることで計算することが可能である [40].

T_n を選択回数 n のタグが選択される相対的な確率とする．タグの総使用回数 t の時点で，選択回数 n のタグが選択される確率は

$$P_n(t) = \frac{T_n N_n(t)}{t} \tag{3.3}$$

と記述することが可能である．ここでは，$N_n(t)$ は t の時点で選択回数 n のタグの種類数である．T_n を実データから求める場合には，各選択ごとに選択回数 n のタグが選択された頻度を $\frac{N_n(t)}{t}$ の逆数で重み付けすることで計算可能である．このときに，実データから求めた T_n と n の重み付きの頻度分布を見たときに，n と無相関である場合には優先的選択に従わないことをあらわす．反対に，n に比例して T_n が増加する場合には優先的選択が働いているとみなすことが可能である．

4 つのデータセットで重み付きの頻度分布を計算したときの結果を図 3.4 に示す．横軸が選択回数であり，縦軸がその選択回数における重み付きの選択頻度である．点線は傾き 1 の補助線である．横軸である n が小さな値を示す場合には，ほぼ補助線と同様の傾きで重み付きの選択確率は上昇していくことがわかる．選択回数が大きな場合には，タグの数が非常に少ないため，そもそも重み付きの選択頻度を計算することが難しい．このため，選択回数が多いものの傾きは補助線から外れてしまう．タグ選択において優先的選択が働いているという点でも，タグ選択を Yule–Simon 過程によりモデル化することは妥当であるといえる．

各タグの選択頻度と選択頻度順にならべた順位との関係

各タグの選択頻度と選択頻度順にならべた順位との関係がべき分布になることは，Zipf 則 [71] と呼ばれ，Heaps 則と同様に他のオンラインソーシャルネットワークサービスでも見つかっている [10, 53]．タグの使われる頻度の順位を r，順位 r でのタグの使われる頻度を $K(r)$ とすると，Zipf 則は

$$K(r) \propto \frac{1}{r^{\gamma}} \tag{3.4}$$

となる．ここで，γ は指数を表し，γ が 1 となる場合を厳密には Zipf 則とよぶ．また，Yule–Simon 過程が生み出す指数 γ は解析的にも求めることが可能であり，$\gamma = 1 - \alpha$ となる [55]. Yule–Simon 過程における α の値は小さな値，あるいは時間減衰する関数であるため，おおよそ 1 となる．

図 3.5 に 4 つのサービスで Zipf 則を計算したときの結果を示す．横軸はタグの使われる頻度の順位であり，縦軸がタグの使われる頻度を表す．また，点線は補助線である．Delicious や RoomClip の場合には，実際のデータの傾きは補助線とほぼ一致する．一方で Flickr や Instagram では r の値が大きい場合には傾きが変化している．しかしながら，Yule–Simon 過程に対して記憶構造を考慮するような拡張を加えることで，このような傾きの変化を記述可能であるこ

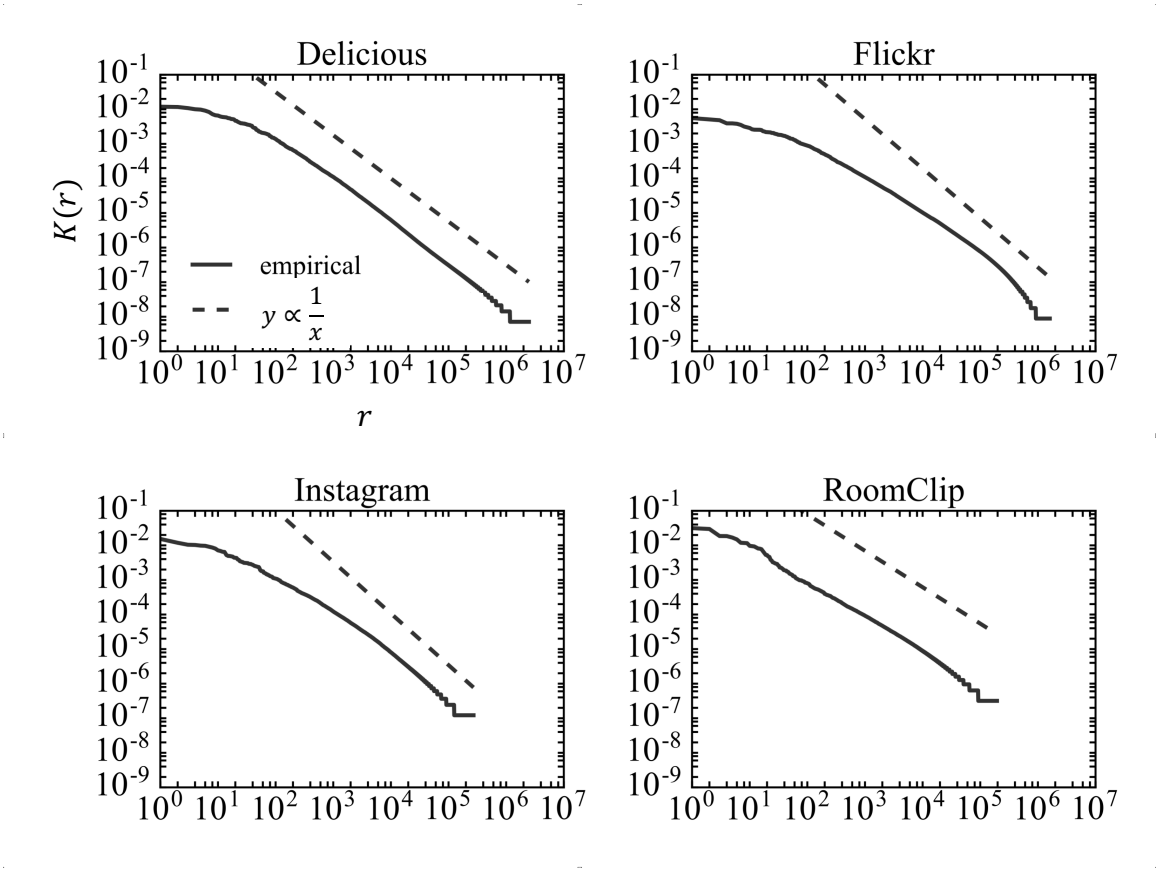


図 3.5: STS を採用する 4 つのサービスにおける Zipf 則の計算結果．横軸はタグの使われる頻度の順位 (r) であり，縦軸がタグの使われる頻度 ($K(r)$) を表す．また，点線は $y \propto \frac{1}{x}$ の場合の補助線である．これは，Yule–Simo 過程をシミュレートした場合の解析解である．

とがこれまでに知られている [10]．このため，タグの使われる頻度の順位と頻度という観点からも，タグ選択を Yule–Simon 過程によりモデル化することは妥当であるといえる．

以上の 3 つの視点から，Yule–Simon 過程をタグ選択のモデルとして利用することは妥当であることを確かめた．

3.2 Windowed Yule–Simon 過程

ここでは，Yule–Simon 過程を拡張することで，タグの同時利用を考慮に入れた Windowed Yule–Simon 過程の提案を行う．

実データの持つタグの振る舞いは Yule–Simon 過程を利用することで同様の傾向を示すことがこれまでにたしかめられている．しかし前述したように，Yule–Simon 過程自身は複数のタグがセットで用いられる，タグの同時選択については考慮されていない．一般的に，投稿さ

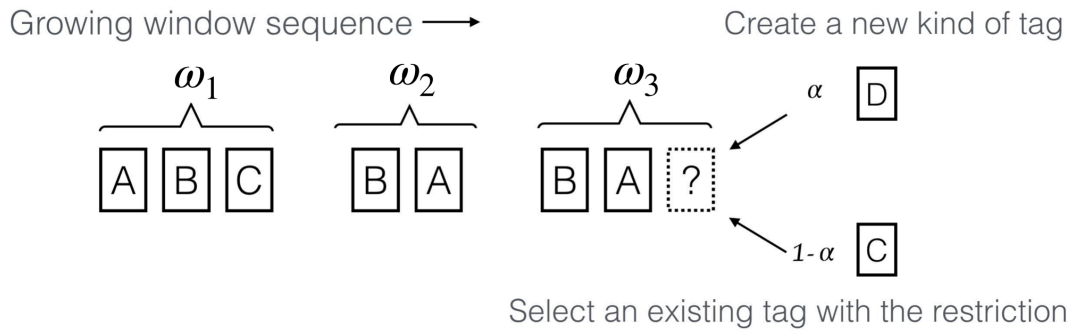


図 3.6: Yule–Simon 過程に対してタグの同時利用を考慮可能なように拡張を加えた，Windowed Yule–Simon 過程の概念図．各ウィンドウは写真や文章といったユーザによる投稿を表す．各ウィンドウには何らかの確率分布によりタグの同時利用個数を表す，サイズが与えられる．新たな種類のタグを生成する確率は，Yule–Simon 過程と同様に α である．既存のタグから選択する場合は Yule–Simon 過程とは異なり，ウィンドウ内でのタグの重複を許さない制約を与える．その制約のもと，既存のタグの中からウィンドウを無視してランダムに 1 つのタグを選択する．例えば，既に 2 つのウィンドウが存在し，3 つ目のウィンドウで，3 つ目のタグを選択する場合を考える．既に存在するタグ A，B，C のうちタグ A，B は既に同じウィンドウ内で使用されているため選択される確率はそれぞれ 0 となる．以上の制約の結果，タグ C が選択される確率は $1 - \alpha$ となる．

れたコンテンツに対してユーザは複数のタグを付与する．通常，それらのタグの間には何らかの関係があることが多い．

そこで，ここでは一つのコンテンツに対して複数のタグを同時に生成する Windowed Yule–Simon 過程を提案する．このタグの同時利用を取り込んだ新しいモデルをタグ生成の null モデルとし，実データの振る舞いをどのような側面で再現し，またどのような点でズレが生じるのかを検証することが可能となる．ここでウィンドウとは同時に付けられるタグのセットを表し，一つのウィンドウが一つのコンテンツに対応する．j 番目に投稿されたコンテンツに対して付与されるタグの数をウィンドウサイズ (ω_j) とする．また，一つのコンテンツに対して同じタグが 2 度使われることはないので，ウィンドウ内でのタグの重複を許さない制約を与える．このような排他則を加えたが，排他則の効果は極めて小さく，タグの使われ方は Yule–Simon 過程とほとんど変わらないことが考えられる．実際に，Yule–Simon 過程では既存のタグの選択はこれまでの選択回数に応じたクラスで選ばれる．クラスの中からどのタグが選択されるかには任意性が残されている．従って十分なタグの種類が存在する場合には排他則の効果を避けながら優先的選択によりタグを選択することは容易である [55].

Yule–Simon 過程と提案モデルの概念図を図 3.6 に示す．新たな種類のタグを生成する確率

は，Yule–Simon 過程と同様に α であり，既存のタグから選択する場合の各タグの選択確率は，次の手順で与える．図 3.6 では，既に 2 つのウィンドウが存在し，3 つ目のウィンドウで，3 つ目のタグを選択する場合を考える．新たな種類のタグ (今回の場合は便宜上 D とした) が生成される確率は α，既に存在するタグ A，B，C のうちタグ A，B は既に同じウィンドウ内で使用されているため選択される確率はそれぞれ 0 となる．以上の制約の結果，タグ C が選択される確率は $1-\alpha$ となる．Algorithm1 に詳細な Windowed Yule–Simon 過程のタグ選択手順を述べる．

Algorithm 1 Windowed Yule–Simon 過程

for $i = 0$ to $V(0)$ **do**
  add i to Tags list
end for
for $j = 0$ to number of contents **do**
  TTags to empty list
  generate ω
  for $i = 0$ to ω **do**
    add random number to $T\alpha[0, 1]$
    if $T\alpha \leq \alpha$ **then**
      add the new kind of tag to Tags list
      add the generated tag to TTags list
    else
      while the selected tag is in TTags list **do**
      end while
      add the selected tag to Tags lis$t$
      add the selected tag to TTags list
    end if
  end for
end for
remove first $V(0)$ tags from Tags list
return Tags list

Algorithm1 中の $V(0)$ は初めに存在するタグの種類の数を表している．$V(0)$ は便宜的に ω の平均値よりも十分に大きい値とする．排他則を加えたため，最初期に確率 $1-\alpha$ で既存のタグを選択する場合，選択可能なタグがなくなってしまうことを防ぐことを目的として，前もって用意しておく初期タグ群である．つまり，α，$V(0)$ はモデルに任意に与えられるパラメーターであり，ω は実データに即した，なんらかの分布関数により生成される．

3.3　タグ共起の観点からの分析

Windowed Yule–Simon 過程により，実データの持つ共起のパターンをどの程度再現可能であるのかを確かめる．分析には，実データと Windowed Yule–Simon 過程それぞれからタグ共起ネットワークを構築し，それぞれのタグ共起ネットワークが持つ相関や階層性の観点から分析を行い比較する．ここでは，実データとして RoomClip のデータを用いて分析を行う．

3.3.1　パラメーター設定

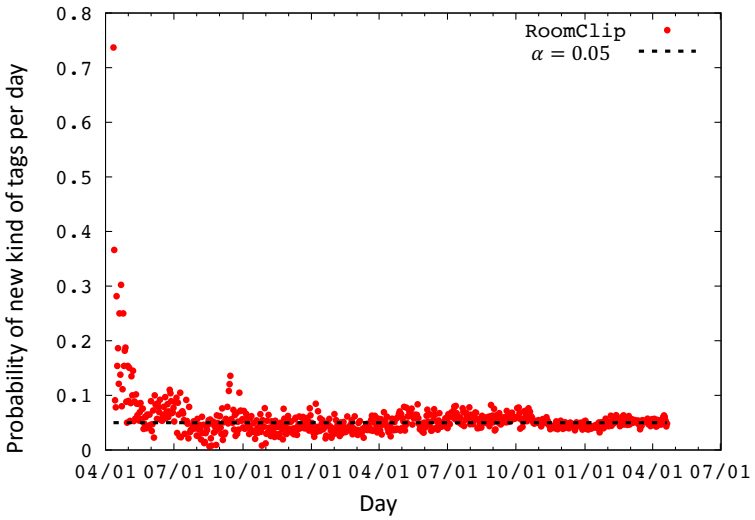


図 3.7: RoomClip における 1 日毎の新たな種類のタグの生成確率の平均．縦軸が新規タグ生成確率の平均であり，横軸が日付である．赤点が RoomClip の結果であり，黒い点線が $\alpha = 0.05$ の場合の補助線である．

Windowed Yule–Simon 過程をシミュレートするためには，新たな種類のタグの生成確率 α に加え，はじめに存在するタグの種類の数 $V(0)$，ウィンドウサイズ ω の分布をパラメーターとして与える必要がある．そこで，Windowed Yule–Simon 過程をシミュレートする上で必要なパラメータを，実データの分析を通して検討する．

はじめに，新たな種類のタグの生成確率 α を実データから検討する．図 3.7 に実データの 1 日毎に平均した新たな種類のタグの生成確率 α の変化を示す．縦軸が新規タグ生成確率の平均であり，横軸が日付である．赤点が Roomclip の結果であり，黒い点線が $\alpha = 0.05$ の場合の補助線である．図 3.7 によると，新たな種類のタグの生成確率はサービスの開始初期は $\alpha = 0.7$ を超え，新しい種類のタグが多数生成されていることが分かる．その後，サービスの後半では $p = 0.05$ 付近で揺らいでいる．そこで，サービスの最初の期間を除いた新たな種類のタグの生成確率平均をとり，モデルに対して与えるパラメーターとした ($\alpha = 0.05$) .

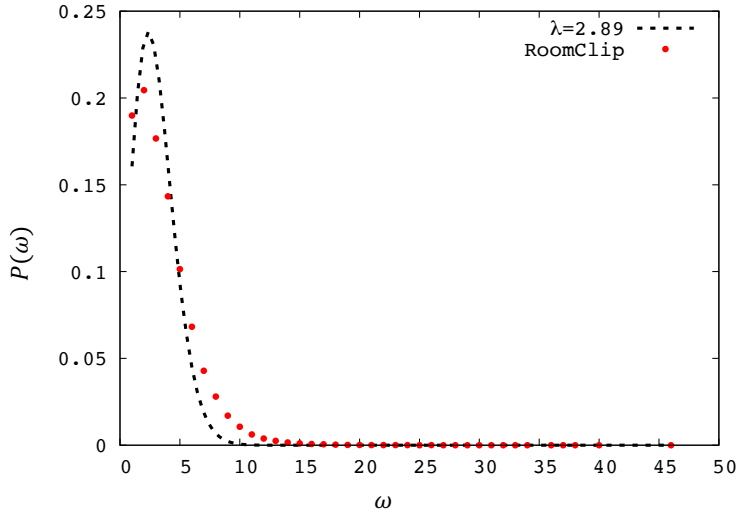


図 3.8: RoomClip におけるウィンドウサイズ (ω) の分布と，最小 2 乗誤差法でポアソン分布にフィッテイングした結果．横軸はある投稿に対して同時に付けられるタグ数 (ω) であり，縦軸はその割合である．赤点が RoomClip の結果であり，黒線が最小 2 乗誤差法でポアソン分布にフィッテイングした結果 ($\lambda = 2.89$) である．

コンテンツに対して同時に付与されるタグ数の分布 ω は，実データの分布がポアソン分布 $P(X = q) = \frac{\lambda^q e^{-\lambda}}{q!}$ に従うと仮定し，パラメーターを実データより求める．後述するモデルの解析解を求める上で，ポアソン分布の性質が望ましかったためである．図 3.8 に RoomClip におけるウィンドウサイズ (ω) の分布と，最小 2 乗誤差法でポアソン分布にフィッテイングした結果を示す．横軸はある投稿に対して同時に付けられるタグ数 (ω) であり，縦軸はその割合である．赤点が RoomClip の結果であり，黒線が最小 2 乗誤差法でポアソン分布にフィッテイングした結果 ($\lambda = 2.89$) である．実際の分布とポアソン分布でフィッティングを行った結果は，似た分布となることがわかる．このため，ω の分布は，最小 2 乗誤差法でフィッティングした分布 ($\lambda = 2.89$) を用いることとする．

はじめに存在するタグの種類の数 $V(0)$ としては，ω の平均値よりも十分大きな $V(0) = 10$ 種類のタグをそれぞれのモデルに対して与える．

3.3.2 Windowed Yule–Simon 過程における個別のタグの振る舞い

パラメーターを設定したことにより，モデルを動かすことが可能となる．はじめに，Windowed Yule–Simon 過程が Yule–Simon 過程と同様に，Heaps 則，Zipf 則をそれぞれ示すことを確認する．これにより，排他則を導入したことによる影響が少ないことを確認する．

Windowed Yule–Simon 過程の Heaps 則を Yule–Simon 過程，実データの結果と共に図 3.9 に示す．縦軸がタグの種類数 ($V(t)$) であり，横軸が総タグ数 (t) である．それぞれ，青線が

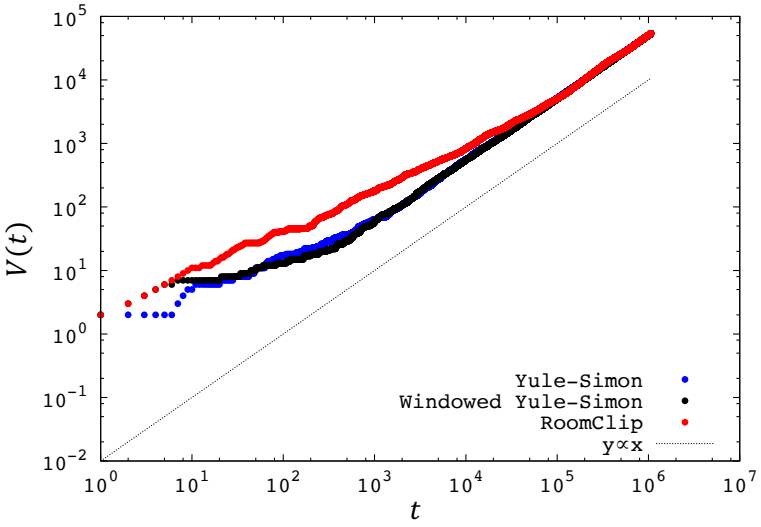


図 3.9: Yule–Simon 過程，Windowed Yule–Simon 過程，実データが示す Heaps 則の計算結果．縦軸がタグの種類数 $V(t)$ であり，横軸が総タグ数 (t) である．それぞれ，青線が Yule–Simon 過程の結果，黒線が Windowed Yule–Simon 過程の結果，赤線が RoomClip の結果である．また，点線は補助線である．

Yule–Simon 過程の結果，黒線が Windowed Yule–Simon 過程の結果，赤線が RoomClip の結果である．また，点線は補助線である．今回は新たな種類のタグの出現確率 α を固定するため $\beta = 1$ のベキ分布となり，それぞれのモデルのシミュレーション結果とも一致する．

Windowed Yule–Simon 過程，Yule–Simon 過程，実データ，それぞれの Zipf 則を計算した時の結果を図 3.10 に示す．横軸はタグの使われる頻度の順位 (r) であり，縦軸がタグの使われる頻度 $(K(r))$ を表す．それぞれ，青線が Yule–Simon 過程の結果，黒線が Windowed Yule–Simon 過程の結果，赤線が RoomClip の結果である．また，点線は補助線である．Zipf 則に関しても，Windowed Yule–Simon 過程と Yule–Simon 過程は同様の分布を示すことを確認した．Windowed Yule–Simon 過程には排他則の効果が入ってくるため，Simon らの解析解とは厳密には異なる．しかし，ここでは用いたウィンドウサイズが小さいため，排他則の効果は無視できる程度に小さいことを確認した．

以上の 2 つの結果から，Windowed Yule–Simon 過程が Yule–Simon 過程と同様に，Heaps 則，Zipf 則をそれぞれ示すことを確認した．

Windowed Yule–Simon 過程も同様に Heaps 則と Zipf 則を示すことが確かめた．次に，実データにより示されるタグの共起構造を Windowed Yule–Simon 過程により再現できていることを，共起ネットワークを構築し共起ネットワークの相関や階層性の観点から分析を行うことで確かめる．

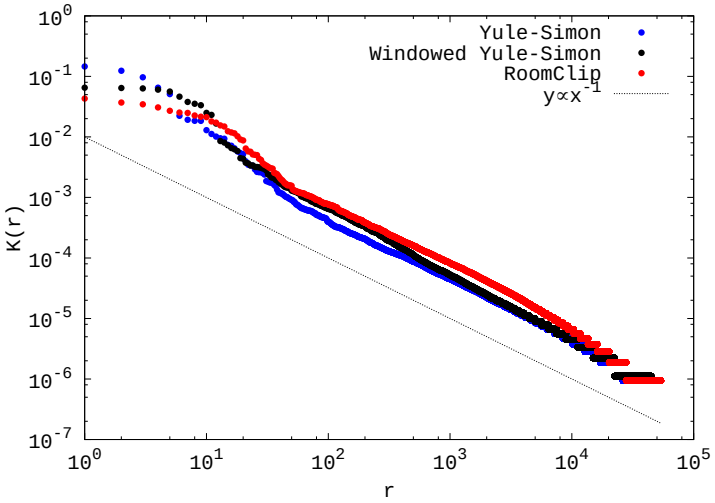


図 3.10: Yule–Simon 過程，Windowed Yule–Simon 過程，実データが示す Zipf 則の計算結果．横軸がタグの使われた頻度の順位 (r) であり，縦軸がタグの使われた頻度 ($K(r)$) である．それぞれ，青線が Yule–Simon 過程の結果，黒線が Windowed Yule–Simon 過程の結果，赤線が RoomClip の結果である．また，点線は補助線である．

3.3.3 タグ共起ネットワーク

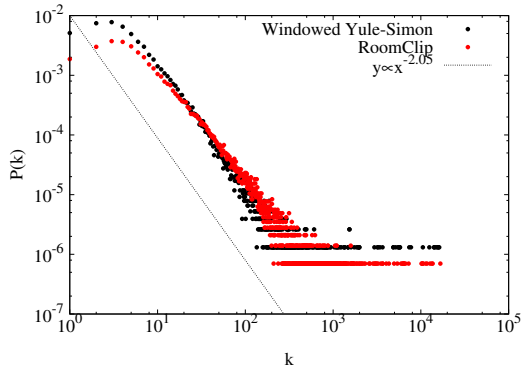
Windowed Yule–Simon 過程により生成されるタグの共起構造と実データのタグ共起構造を，タグの共起関係から作る共起ネットワークを用いて分析・比較する．タグの共起構造を分析するために，タグ共起ネットワークを作成する．タグ共起ネットワークのノード i はタグの種類に対応し，エッジは同一のウィンドウ内で出現したタグ間で貼られる．ノード i とノード j のリンクの重み w_{ij} $(= w_{ji})$ は，リンクでつながる 2 つのタグの共起回数で与える．ノードの重み s_i は，

$$s_i = \sum_{j \in G(i)} w_{ij} \tag{3.5}$$

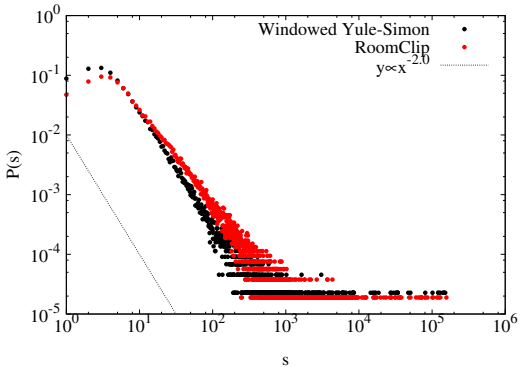
とする．ここで，$G(i)$ はノード i の隣接ノードセットを表す．

実データから作られるタグの共起構造を Windowed Yule–Simon 過程により再現できているかを確かめる．具体的には，共起ネットワークに対して，ネットワークの階層性を含む次の 3 つの分析を行うことで確かめる．

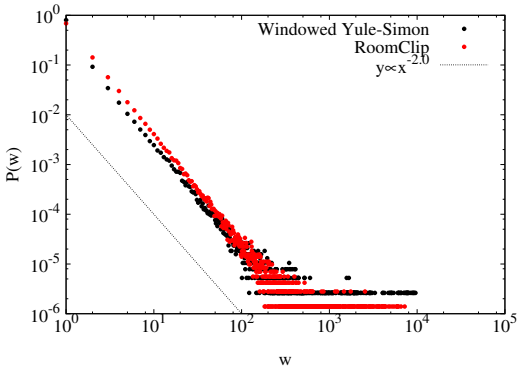
1. ノードの次数 (k_i)・重み (s_i)，エッジの重み (w_{ij}) の分布
2. 次数-クラスタ係数相関
3. 次数-次数相関



(a) ノードの次数 (k) 分布



(b) ノードの重み (s) 分布



(c) エッジの重み (w) 分布

図 3.11: RoomClip と Windowed Yule–Simon 過程から構築したタグ共起ネットワークにおけるノードの次数 (k)，ノードの重み (s)，エッジの重み (w) の分布の結果．(a)， (b)，(c) の横軸は，それぞれノードの次数 (k)，ノードの重み (s)，エッジの重み (w) であり，縦軸はその割合である．黒い点が Windowed Yule–Simon 過程の結果であり，赤い点が RoomClip の結果である．また，点線は補助線である．

3.3.4 次数・重みの分布

Windowed Yule–Simon 過程と実データのタグ共起ネットワークからそれぞれノードの次数 (k_i)・重み (s_i)，エッジの重み (w_{ij}) の分布を計算した結果を図 3.11 に示す．(a)， (b)，(c) の横軸は，それぞれノードの次数 (k)，ノードの重み (s)，エッジの重み (w) であり，縦軸はその割合である．黒い点が Windowed Yule–Simon 過程の結果であり，赤い点が RoomClip の結果である．また，点線は補助線である．図 3.11 をみてわかるように，ノードの次数分布，ノードの重み分布，エッジの重み分布は全てベキ分布を示し，Windowed Yule–Simon 過程と実データは似たような傾向を示すことが明らかになった．このようにノードの次数がベキ分布となるようなネットワークは Web をはじめとした様々なネットワークにおいて観測される [5]．また，Windowed Yule–Simon 過程のタグ共起ネットワークの次数分布がベキ分布に従うことは，

ウィンドウサイズの平均値が定義できる場合には，

$$p(k) \quad \propto \quad k^{-1/(1-\alpha)-1} \tag{3.6}$$

となり，解析的に示すことができる (詳しい式の導出は付録に示す)．今回解析に使用した α は 0.05 であり，図 3.11 に示すように Windowed Yule–Simon 過程から示される次数分布は指数が -2.05 のベキ分布となることがわかる．

3.3.5 次数-クラスタ係数相関

次数-クラスタ係数間の相関を実データと Windowed Yule–Simon 過程それぞれに関して計算する．

重みなし無向グラフのクラスタ係数は

$$c_i = \frac{2e_i}{k_i(k_i-1)} \tag{3.7}$$

で定義される．ここで e_i はノード i の隣接ノード同士で貼られるエッジ数である．次数 k_i とクラスタ係数 c_i の関係は，

$$C(k) = \frac{1}{N_k}\sum_i \delta_{k,k_i} c_i \tag{3.8}$$

で表される．N_k は次数 k のノード数とする．重みなしの次数-クラスタ係数相関に，負の相関が現れる場合にはネットワークが階層構造になっていることを示唆する [46, 47].

さらに，エッジの重みを考慮に入れたクラスタ係数は

$$c_i^w = \frac{1}{s_i(k_i-1)} \sum_{j,h\in G(i),j\in G(h)} \frac{(w_{ij}+w_{ih})}{2} \tag{3.9}$$

で定義される [6]．重みありの場合の次数 k_i とクラスタ係数 c_i^w の関係は，

$$C^w(k) = \frac{1}{N_k}\sum_i \delta_{k,k_i} c_i^w \tag{3.10}$$

で計算する．重みなしのクラスタ係数 $C(k)$ と重みありのクラスタ係数 $C^w(k)$ の間に $C(k) > C^w(k)$ という関係が成り立つとき，重みの小さなエッジを介したクラスタが生成されやすいということを表す．反対に，$C(k) < C^w(k)$ という性質が成り立つとき，重みの大きなエッジを介しクラスタが形成されやすいということを表す．図 3.12 に Windowed Yule–Simon 過程と RoomClip のタグ共起ネットワークから計算した重みなしの場合の次数-クラスタ係数相関の結果を，図 3.13 に重みありの場合の結果をそれぞれ示す．横軸が次数 (k) であり，縦軸が重みなし，あるいは重みありのクラスタ係数である．黒い点が Windowed Yule–Simon 過程の結果であり，赤い点が RoomClip の結果である．重みなし・重みありの結果は共に，Windowed Yule–Simon 過程と実データは同様の分布を示し，負の相関が現れる．これは実データと Windowed

Yule–Simon 過程から生成される共起ネットワークは階層構造を持つことを示唆している．重み付きの場合，次数 k_i の値が小さい場合はエッジの重みとクラスタの現れやすさの間には関係は現れない．反対に，次数 k_i の値の大きい場合には重みなしのクラスタ係数と重み付きのクラスタ係数の間に $C(k) < C^w(k)$ の関係が成立し，重みの大きなエッジを介しクラスタが形成されやすいという性質が現れることを示す．

3.3.6 次数-次数相関

次にネットワークの階層構造の更なる分析として，次数-次数相関の計算を行う．ノード i の重みなしの隣接ノードの平均次数を

$$k_{nn,i} = (1/k_i) \sum_{j \in G(i)} k_j \tag{3.11}$$

とすると，次数 k と平均隣接ノード次数の関係 $k_{nn}(k)$ (次数-次数相関) は

$$k_{nn}(k) = \frac{1}{N_k} \sum_i \delta_{k,k_i} k_{nn,i} \tag{3.12}$$

と定義される [43]．一般的にインターネットなど人工的に生成されたネットワークには負の相関がみられ，共著関係など社会的なネットワークの場合には正の相関が見られることが知られている [41]．さらに，重みを考慮に入れた重みありの隣接ノードの平均次数を

$$k^w_{nn,i} = \frac{1}{s_i} \sum_{j \in G(i)} w_{ij} k_j \tag{3.13}$$

とすると，次数 k と平均隣接ノード次数の関係 $k^w_{nn}(k)$ は，

$$k^w_{nn}(k) = \frac{1}{N_k} \sum_i \delta_{k,k_i} k^w_{nn,i} \tag{3.14}$$

と定義される [6]．ここで，重みなし相関と重みあり相関の間に $k^w_{nn}(k) > k_{nn,i}$ という関係が成り立つ場合は，大きな重みを持つエッジは次数の大きな隣接ノードを指しやすいことを示す．逆に，$k^w_{nn}(k) < k_{nn,i}$ という関係が成り立つ場合は，大きな重みを持つエッジは次数の小さな隣接ノードを指しやすいことを示す．

重みなしの場合の次数-次数相関を図 3.14 に，重みありの場合の次数-次数相関を図 3.15 にそれぞれ示す．横軸が次数 (k) であり，縦軸が重みなしの次数 ($k_{nn}(k)$)，あるいは重みありの次数 ($k^w_{nn}(k)$) である．黒い点が Windowed Yule–Simon 過程の結果であり，赤い点が RoomClip の結果である．重みなしの場合の全体の傾向としては負の相関を示すということがわかる．ただし，次数の小さな場合には実データと Windowed Yule–Simon 過程では大きな差が見られた．この差は実データにおける出現頻度の低いタグは出現頻度の低いタグとモデルから期待される以上に共起しやすいことを示している．

このような相関の強さは，

$$A = \frac{S^{-1}\sum_j m_o l_o - [S^{-1}\sum_i \frac{1}{2}(m_o+l_o)]^2}{S^{-1}\sum_o \frac{1}{2}(m_o^2+l_o^2) - [S^{-1}\sum_o \frac{1}{2}(m_o+l_o)]^2} \tag{3.15}$$

によって定義される同類度 A (Assortativity) で定量的に示すこともできる [41]．ここで，S はネットワーク上の総エッジ数，m_o と l_o はエッジ o の両端の次数を示す．Windowed Yule–Simon 過程と実データに対して同類度を計算したところ，それぞれ $A = -0.296$，$A = -0.185$ となった．このことからも両者のネットワークは次数の大きいノードと次数の少ないノードがリンクを持ちやすいという性質を持つことがわかる．

重みありの場合は，Windowed Yule–Simon 過程も RoomClip も同じような分布を示すが，相関は重みなしの場合より小さい．この傾向は次数 k の値が小さいところほど顕著である．次数 k の値が大きくなると $k^w_{nn}(k) > k_{nn,i}$ の関係が成り立ち，大きな重みを持つエッジは次数の大きな隣接ノードを指しやすくなることがわかる．

3.4 タグ共起分析のまとめ

ここではタグの共起構造に注目し，Yule–Simon 過程を拡張した Windowed Yule–Simon 過程を提案した．提案モデルの Zipf 則，Heaps 則は Yule–Simon 過程と同様のベキ分布を示した．次に，提案モデルが実データにおけるタグ共起の特徴をどの程度再現するかを確かめるために，両者のタグ共起ネットワークが持つ構造の比較を行った．Windowed Yule–Simon 過程のタグ共起ネットワークは実データで観測されるものと様々な面で近い傾向を示した．それは次数分布，リンクの重みの分布，ノードの重みの分布がベキ分布を示すという一次の統計量や，次数-クラスタ係数相関や次数-次数相関といった二次の統計量である．二次の統計量の分析から，Yule–Simon 過程を拡張した今回の簡単な確率モデルはタグ共起ネットワークが持つ階層性を再現することがわかった．

ソーシャルタギングにおいて，実際に利用されるタグはそれぞれが固有の意味を持ち，意味に基づいて共起しやすいタグやしにくいタグが存在する．今回我々が提案したモデルはこのような意味的なバイアスを含まないため，モデルと実データが示すネットワーク構造には大きな違いが見られることが考えられた．にもかかわらず，両者は定性的に似た傾向を示した．この点が本研究の興味深い点である．一方で，このような簡単な確率モデルで本当の意味でのタグの持つ意味を捉えることができているとは思えない．タグの持つ意味がもたらす使われ方の偏りが，共起のどういった特徴量に現れるのかという点でさらなる分析が必要である．例えばネットワークのモジュール性に注目した分析は有望な候補であろう．

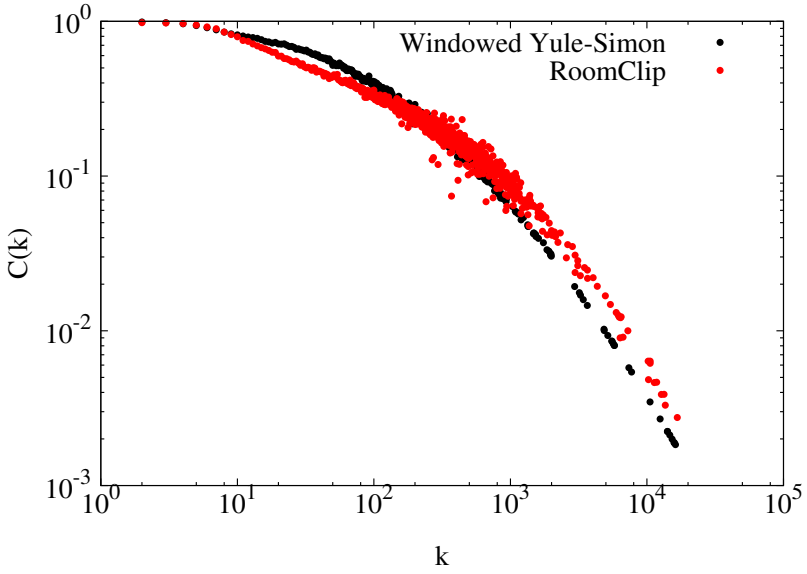


図 3.12: RoomClip と Windowed Yule–Simon 過程から構築した，タグ共起ネットワークにおける重みなしの場合の次数-クラスタ係数相関．横軸が次数 (k) であり，縦軸が重みなしのクラスタ係数 ($C(k)$) である．黒い点が Windowed Yule–Simon 過程の結果であり，赤い点が RoomClip の結果である．

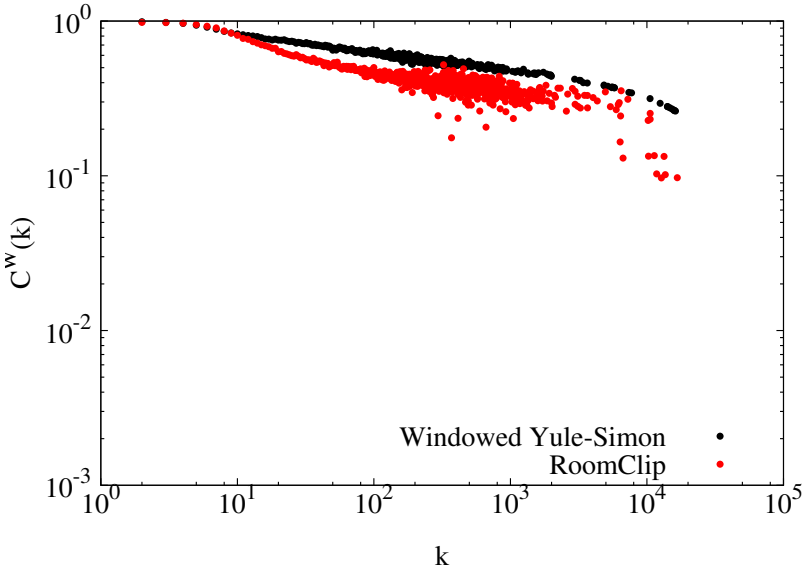


図 3.13: RoomClip と Windowed Yule–Simon 過程から構築した，タグ共起ネットワークにおける重みありの場合の次数-クラスタ係数相関．横軸が次数 (k) であり，縦軸が重みありのクラスタ係数 ($C^w(k)$) である．黒い点が Windowed Yule–Simon 過程の結果であり，赤い点が RoomClip の結果である．

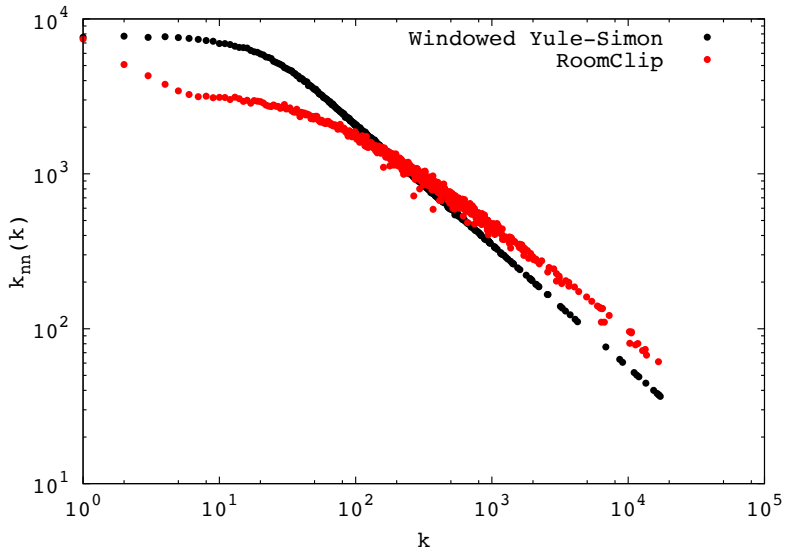


図 3.14: RoomClip と Windowed Yule–Simon 過程から構築した，タグ共起ネットワークにおける重みなしの場合の次数-次数相関．横軸が次数 (k) であり，縦軸が重みなしの次数 ($k_{nn}(k)$) である．黒い点と赤い点はそれぞれ，Windowed Yule–Simon 過程と RoomClip の結果である．

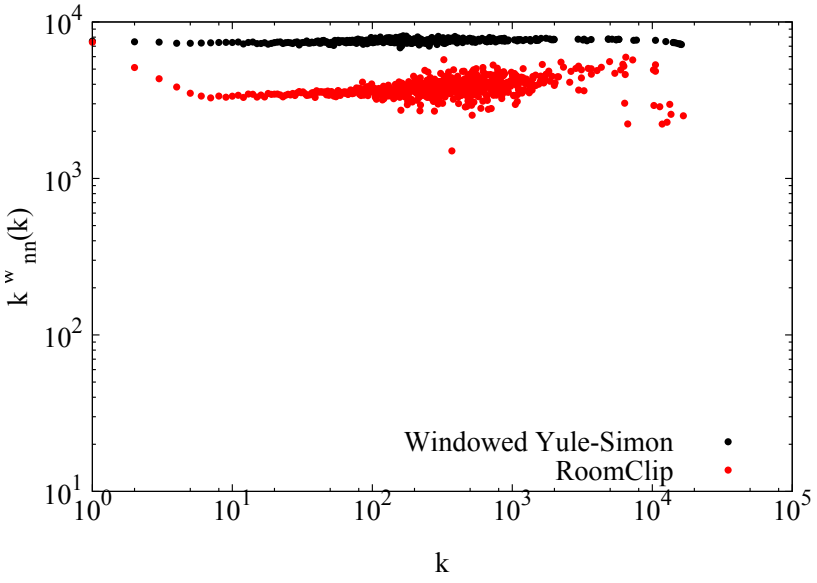


図 3.15: RoomClip と Windowed Yule–Simon 過程から構築した，タグ共起ネットワークにおける重みありの場合の次数-次数相関．横軸が次数 (k) であり，縦軸が重みありの次数 ($k^w_{nn}(k)$) である．黒い点と赤い点はそれぞれ，Windowed Yule–Simon 過程と RoomClip の結果である．

第4章　タグ同時選択における新規タグの生み出され方の分析

Yule–Simon 過程は，タグ付けにおける Zipf 則や Heaps 則といった，マクロな統計的特徴をうまく説明することがこれまでに知られている．一方で，Yule–Simon 過程はミクロな観点，つまり新たな種類のタグの作成においてみられる，不規則性や何らかの相関といった性質については特に何も述べてはいない．実際に Yule–Simon 過程や，今回提案した Windowed Yule–Simon 過程では，新規タグの生成確率が一定，あるいは時間的に減衰するような関数であることを仮定している．つまり，Yule–Simon 過程では新たな種類のタグが作成されるメカニズムはランダムな確率により記述され，その詳細なメカニズムに関しては特に考慮していない．

よりミクロなスケールで新規タグの生成を見た場合には，単純なポアソン過程のようなランダムなプロセスからは逸脱することが考えられる．例えば，同じ投稿内で新規タグが作成される場合，同時に選択される，あるいは新たに生成される新規タグ間には何らかの相関が生じることが考えられる．これは、意味的なつながりや，文脈，タグを通じたコミュニケーション方法に関連している可能性がある．同様の概念に基づいて，Tria らは Kauffman による隣接可能性を用いることで，新規タグ間の相関関係の存在について議論しており，新規タグの生成は他の新規タグの生成を促す役割を果たすことを明らかにしている [62]．ここでは，別の観点から新規タグ生成についての分析を行う．つまり，ある投稿に対して複数のタグが同時に付与された場合に，どのように新規タグ生成されるのかということである．

実際のソーシャルタギングにおける新規タグの生成確率 α は，ウィンドウサイズごと，あるいは投稿するユーザごとに異なることが考えられる．そこで，このウィンドウ固有の α とウィンドウの長さとの関係を分析により明らかにする．これにより，各ユーザが各投稿において，どのように新規タグの生成を行っているのかを明らかにすることにつながる．Yule–Simon 過程において考慮されるのは，各ウィンドウを無視した，投稿されたすべてのタグからなるマクロなタグのシーケンスである．タグの同時利用を考慮した場合には，マクロなタグのシーケンスに加えて，ミクロなタグのシーケンスを考えることが可能である．ここでいうミクロなタグのシーケンスとは，同一ウィンドウ内で一人のユーザによって生成されるタグのシーケンスである．

タグ同時利用における新たな種類のタグの生み出され方の分析は，以下の 4 つの部分からなる．まずはじめに，ウィンドウ固有の α とウィンドウサイズとの関係の分析方法について説明を行う．次に，実データに見られるウィンドウサイズと新規タグの生成確率間の関係性についての分析結果を示し，その結果をユーザのタグを付ける動機と合わせて議論する．最

後に結論についてを述べる．

4.1 分析手法

ここでは，ウィンドウサイズ (ω) ごとの新規タグの生成確率 (α) の分析方法の説明と，各サービスにおけるユーザのタグを付ける動機の分析方法の説明を行なう．

4.1.1 ウィンドウサイズごとの，新規タグの生成確率

まずはじめに，ウィンドウサイズ ω ごとの新規タグの生成確率 α の分析方法の説明を行なう．これにより，各ユーザが各投稿において，どのように新規タグの生成を行っているのが明らかになる．

ウィンドウサイズ (ω) と新規タグの生成確率 (α) の関係は以下のように分析を行った．まずはじめに，マクロな投稿のシーケンスを 3ヶ月間ごとに分割する．これは，投稿数やタグの種類数の増加とともに α と ω の相関が徐々に安定すると考えられたためである．サービスの初期段階では十分な語彙がないため，新たなタグの生成確率を適切に考慮できないことが考えられる．実際に最初期の時刻では，投稿に対して付与されるタグのほとんどすべてが新規タグとしてみなされてしまう．マクロなタグのシーケンスは 3ヶ月ごとに分割をおこなうものの，あるタグが過去のすべての期間で利用されたタグも含め，初めて利用された場合に新規タグとして扱うという点は変わらない．また，各投稿はウィンドウサイズごとに各期間で同サイズの 20 個のログスケールのビンにより分割を行った．その後，各ビンに含まれる投稿ごとに新規タグを含む割合を計算する．また，ビンに含まれる投稿数が 100 を超えないビンに関しては計算から除いた．これは，ビンに含まれる投稿数が少ない場合には，その瓶での新規タグを含む割合には意味が無いと考えられるためである．

4.1.2 ユーザがタグを付ける動機

ウィンドウサイズごとの新規タグの生成確率を考えた場合には，ユーザのタグを付ける動機がウィンドウの統計と新規タグの生成確率に対して何らかの影響を与えることが考えられる．各サービスごとに支配的なタグ付けの動機が異なる場合には，ω と α の統計もそれに従って，各サービスごとに異なる相関を示すことが考えられる．例えば，タグはさまざまな目的で利用され，あるユーザは自身の情報管理のためにタグを付与し，あるユーザは自身の投稿をほかのユーザに広く認知させるためにタグを付ける．後者のためのタグ付けである場合には，ある程度一般的な既存のタグを付けることで，他のユーザから自身の投稿を検索されやすくしている事が考えられる．このような場合には，同時に付けられるタグの数が増えるにつれて新規タグの生成確率は減少する．

これまでの研究から，ユーザのタグを付ける動機は大きく 2 種類に分けることが可能であると言われている．1 つは，投稿を他のユーザと共有することを目的としたタグ付けである．

もう一つは，投稿の個人的な再利用を意図したタグ付けである [61, 29, 31]．ここでは，他者への情報共有を目的としたタグ付け行動を，オープンなタグ付けと呼ぶ．反対に，個人的な投稿の管理を目的としたタグ付けをプライベートなタグ付けと呼ぶ．ここでは以下の 2 つの指標から，各サービスと各投稿の背後にあるタグを付ける動機の特徴づけを行なう．それぞれ，ウィンドウサイズ (ω) と，Stromaier らにより提案された，タグ付けの動機によるタグ選択の偏りを捉えた指標である [56].

ウィンドウサイズとユーザのタグを付ける動機について述べた研究に，Heckner らのものがある．Heckner らは，単一の投稿に対して同時に付与されるタグの数は，タグ付けの動機の影響を受けることを明らかにした [32]．ユーザが他のユーザと自身の投稿を共有する目的でタグを使用する場合には，ウィンドウサイズが大きくなる．これはオーバータギングと呼ばれる現象であり，投稿に対して同時に付与されるタグの数が増えると，タグ検索によって他のユーザからその投稿にアクセスされる可能性も高まると考えるのが自然である．つまり，オープンなタグ付けを行うユーザが多いサービスである場合にはウィンドウサイズの平均や中央値は大きな値を示す．反対に，プライベートなタグ付けを行うユーザが多いサービスである場合には，ウィンドウサイズの平均や中央値は小さな値を示す．そこで，ウィンドウサイズの平均と中央値を計算することで，各サービスと各投稿の背後にあるタグを付ける動機の特徴づけをおこなう．

ウィンドウのサイズとユーザのタグを付ける動機に関する上記の解釈は直感的にも理解できる．一方で，この値は実際にユーザが実際に選択するタグの違いについては何も述べていない．このため，ユーザのタグを付ける動機による，選択されるタグの違いという視点からもユーザのタグを付ける動機を特徴づけることが必要である．

Stromaier らはソーシャルタギングにおいて典型的な 2 つのタグを付ける動機と，つけられるタグの違いに着目して両者を特徴づけるための指標を提案した．2 つのタグを付ける動機はそれぞれ，記述者と分類者と呼ばれる [56, 36]．分類者は，個人的な情報検索を目的として，自身が行った異なる投稿に対して同じタグを繰り返し利用する傾向にあるユーザとして定義される．それとは対照的に，記述者は個々の投稿を正確に記述するために，分類者のタグ付けと比較して各投稿に対して付与されるタグの選択が偏っているユーザと定義される．Stromaier らは，説明者が使用するタグは，分類者が使用するタグと比較して，ユーザにとってより一般的な語彙を選択していることを明らかにしている．このことから，説明者がオープンな動機でタグ付けを行い，分類者がプライベートな動機でタグ付けを行うといえる．Stromaier らはこれらのタグを付ける動機を捉える指標として，タグ選択の偏りを捉えた M_0 と M_1 という 2 つの指標を提案している．それぞれの指標は分類者，あるいは記述者的なタグ選択の偏りに対応する．両者の平均をとった M により，各ユーザのタグを付ける動機を特徴付けることが可能である．ここで，M は各ユーザに対して計算される値である．

記述者に関連した指標である，M_0 は

$$M_0 = \frac{|\{i : |R(i)| \leq n\}|}{|T|}, \quad n = \left\lceil \frac{|R(i_{\max})|}{100} \right\rceil \tag{4.1}$$

と定義される．ここでは，$|T|$ はユニークなタグの数，$|R(i)|$ はタグ i が割り当てられた投稿

の数，$|R(i_{\max})|$ は、最も使用されたタグが割り当てられた投稿の総数である．M_0 の取りうる値の範囲は $[0:1]$ であり，ユーザがさまざまな種類のタグを投稿に対して付与し，各タグの総使用回数が少ない場合に大きな値をとる．

分類者に関連した指標である，M_1 は

$$M_1 = \frac{H(R|T) - H_{\text{opt}}(R|T)}{H_{\text{opt}}(R|T)} \tag{4.2}$$

と定義される．T はタグのセットであり，R はリソースのセットである．$H(R|T)$ は条件付きエントロピーであり，各タグが付けられた投稿数の偏りを捉えた値である．$H_{\text{opt}}(R|T)$ は，分類者として理想的な振る舞いを示した場合の理想的な $H(R|T)$ の値である．これは，正規化するための係数としての役割を果たす．M_1 の取りうる値の範囲は $[0:\infty]$ である．しかしながら，ほとんどの場合において 1 未満の値であることが確かめられている．M_1 が 0 に近い値を示す場合には，そのユーザのタグの付け方は分類者的であることを意味する．

これら 2 つの指標の平均として，M は

$$M = \frac{M_0 + M_1}{2} \tag{4.3}$$

と定義される．これは実際に各ユーザのタグ付けを説明者，あるいは分類者として識別する上で用いられる値である．値が 0 に近い場合には，そのユーザのタグ付けはより分類者に近いものであることを意味する．反対に，1 に近い大きな値をとる場合には，ユーザのタグ付けはより説明者に近いことを意味する．各ユーザの M の値を計算し，各サービスごとの分布の違いをみることで，各サービスにおけるユーザのタグを付ける動機がオープンなタグ付けとプライベートなタグ付け，どちらの傾向が強いユーザが多いのかを明らかにする．

今回利用した 4 つのサービスのタグ付けデータは，タグを付ける動機という観点で見ると，直感的にはそれぞれ以下のように分類することが可能である．Delicious は Web ブックマークを共有するために利用されるサービスである．他の 3 つのサービスは，写真の共有のために利用されるサービスである．Delicious を利用するユーザは主にブックマークの保存のためにサービスを使用し，Instagram は通常ユーザの投稿を他のユーザと共有することを意図して利用される [16, 31]．Flickr は Delicious と Instagram の中間に位置するようなサービスである [16, 31]．一方で，RoomClip はまだ研究者の注目を集めていないため，このサービスの主な使用方法は明確に特定されていない．しかしながら，Instagram に当てはまるように，RoomClip にアップロードされた写真はすべてのユーザが自動的にアクセスできるため，Flickr ではなく Instagram に性質が似ていると考えることが可能である．また，Flickr とは異なり，RoomClip にはアップロードされた写真の保存容量の制限も存在しない．

4.2 分析

実際の分析結果をここでは示す．まずはじめに，新規タグの生成確率 α とウィンドウサイズ ω の間に，各サービスでどのような関係が存在するのかを分析により明らかにする．次に，

ウィンドウサイズの平均や中央値と M の分布を各サービスごとに計算し，各サービスにおけるユーザのタグ付けの動機を特定する．最後に，各サービスにおけるタグをつける動機と α と ω の関係について議論する．

4.2.1 新規タグの生成確率 α とウィンドウサイズ ω の相関

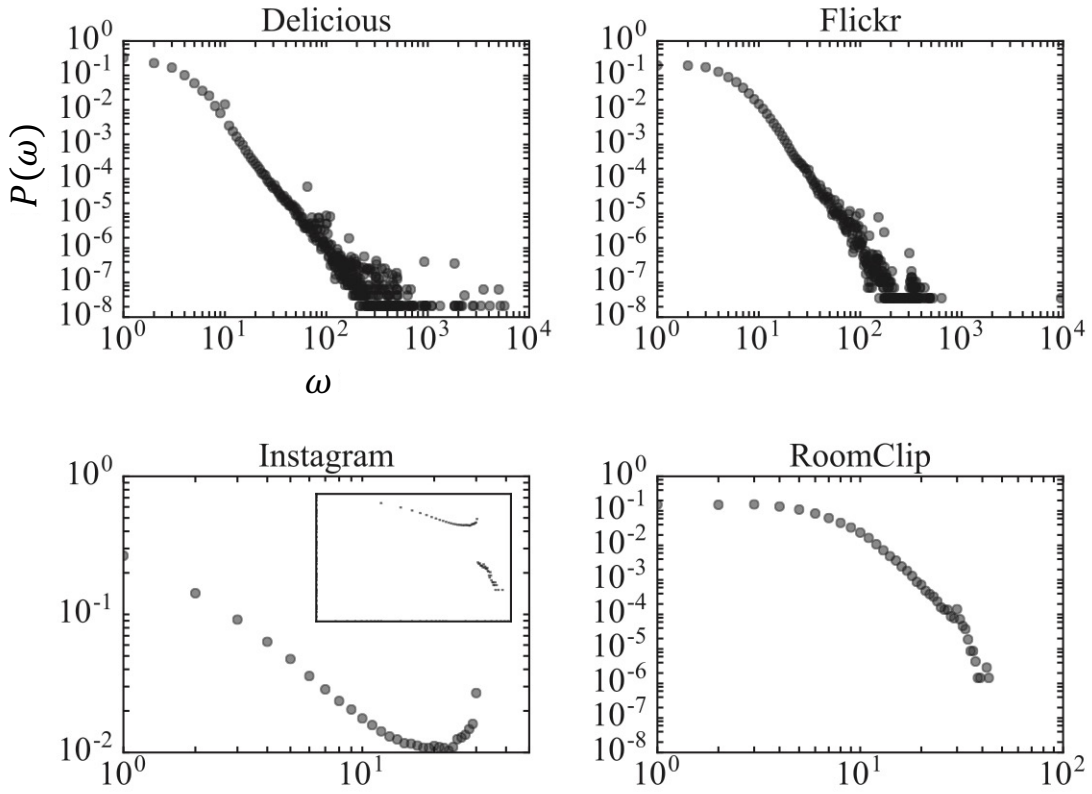


図 4.1: STS を採用する 4 つのサービスにおける，ウィンドウサイズ (ω) の分布．横軸はウィンドウサイズ (ω) であり，縦軸はその割合である．Instagram の結果は ω が 0-30 範囲の場合の結果を切り取ったときのものである．また，Instagram の挿入図は，実際の ω 全体の分布である．

まずはじめに，4 つのデータセットのウィンドウサイズの分布を確認する．図 4.1 に 4 つのデータセットの ω 分布を示す．横軸はウィンドウサイズ (ω) であり，縦軸はその割合である．Instagram の結果は ω が 0-30 範囲の場合の結果を切り取ったときのものである．また，Instagram の挿入図は実際の ω 全体の分布である．図 4.1 を見てわかるように，Delicious と Flickr は同様のべき分布を示すことがわかった．このことは，ω の取りうる範囲が広く，ω の

値が大きくなるにつれて，投稿数が大きく減少することをあらわしている．Instagram および RoomClip の ω の範囲は Delicious，Flickr の範囲にくらべて狭く，べき分布であるとは言えない．一方で，ω が大きい値をとるにつれ投稿数が減少していくという点では Delicious や Flickr と同様である．また，Instagram の場合，$\omega = 30$ に固有のギャップが存在することがわかる．これは，Instagram が独自に持つペナルティ機能により生じるものであることが考えられる．実際に，インスタグラムでは 30 個以上タグが付けられる投稿にはコメントが制限される場合がある．

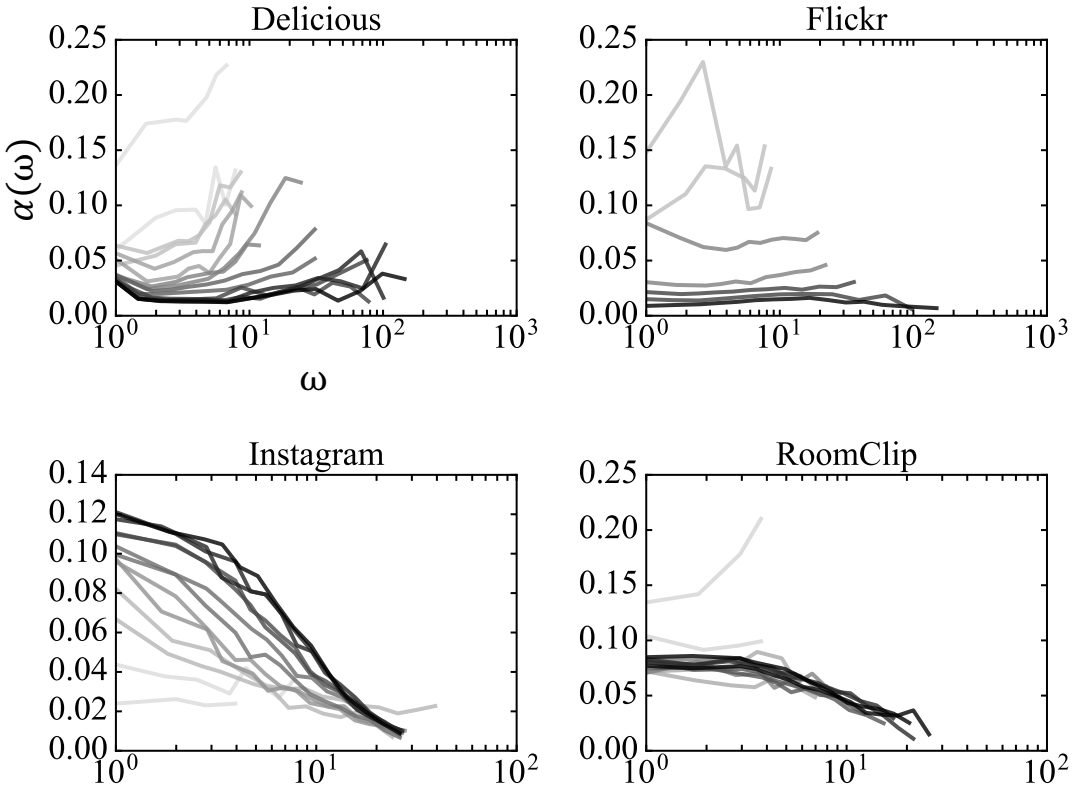


図 4.2: STS を採用する 4 つのサービスにおける，新規タグの生成確率 (α) とウィンドウサイズ (ω) の相関．横軸はウィンドウサイズ (ω) であり，縦軸はそのウィンドウサイズで新規タグを含む割合である．グラデーションは経過時間のインデックスをあらわし，色の濃いものほど投稿された日付の新しいものである．

図 4.2 に，各データセットの α と ω の相関の結果を示す．横軸はウィンドウサイズであり，縦軸がログスケールでビンニングを行った場合の，各ビンの新規タグの生成確率である．また，グラデーションは経過時間のインデックスをあらわし，色の濃いものほど取得された日付の新しいものである．図 4.2 を見てわかるように，ウィンドウサイズと新規タグの生成確率間の関係は，2 つの異なる関係性を示すことがわかった．Delicious と Flickr では，それぞ

れ弱い正の相関，あるいは無相関を示した．Instagram と RoomClip では負の相関を示した．

Yule–Simon 過程では，新規タグはランダムあるいは時間減衰する関数として記述されている．このため，図 4.2 において観測された相関の存在は，Yule–Simon 過程では考慮されていない別の新規タグ生成メカニズムが機能していることを示唆する．各サービスごとの相関の違いは，ユーザのタグを付ける動機によりある程度説明可能であることが考えられる．Delicious および Flickr において観測された弱い正の相関は，新規タグの生成により ω の増加が引き起こされること，あるいは ω の増加により新規タグの生成が引き起こされることを意味する．直感的には，これら 2 つの可能性のうち前者，つまり ω の増加は新規タグの作成によって引き起こされるということは，より理にかなっていることが考えられる．これは，ユーザがプライベートな動機でタグ付けを行った結果である．その目的は，個人使用のために，各投稿の再検索性を高めるためのタグを付けることで自らの投稿を分類することである．このようなタグ付けは，しばしば日付や名前など，他の人から見たら単なる記号でしか無いようなタグ，あるいは新規タグである場合が存在する．それとは対照的に，Instagram および RoomClip データで観測される負の相関は，ω の増加が既存タグの選択によって引き起こされていること，あるいは ω が増加することで既存タグがより選択されやすくなることを意味する．これは，ユーザのオープンなタグ付け行動の結果であり，その目的は他のユーザと自身の行った投稿を共有することにある．タグ付けを，他のユーザから投稿を検索されやすくするという観点で見ると，このような場合には新規タグを使用することは無意味である．

4.2.2 ユーザのタグを付ける動機

ここでは，分析を通して各サービスのユーザがタグを付ける動機を数値的に特徴づける．これにより，プライベートな動機でタグ付けを行なうユーザが多数存在するようなサービスの場合には，新規タグの生成確率とウィンドウサイズの間には正の相関が現れ，オープンな動機でタグ付けを行なうユーザが多数存在するようなサービスの場合には，負の相関が現れるという，上記の解釈が妥当であることを確かめる．各サービスにおけるユーザのタグを付ける動機の分析は，ウィンドウサイズの平均や中央値と M から評価を行う．

表 4.1: STS を採用する 4 つのサービスにおける ω の平均と中央値．

Service	Median of w	Mean of w
Delicious	2	2.96
Flickr	3	4.01
Instagram	3	7.82
RoomClip	4	4.53

まずはじめに，ウィンドウサイズの平均と中央値に着目する．表 4.1 は各サービスの ω の平均と中央値である．Heckner らの研究に従い，ω が大きな値を示すサービスの場合には，サービスの主要なユーザがオープンな動機でタグ付けを行っていることを意味する．ω が小さな

値を示すサービスの場合には，多くのユーザはプライベートな動機でタグ付けを行っていることを意味する．表 4.1 より，ω の平均は Delicious, Flickr, RoomClip, Instagram の順に大きな値となることが明らかとなった．また，中央値に関しても同様の傾向を示すことが確かめられた．このため，Delicious および Flickr のユーザは，Instagram，RoomClip のユーザと比較するとよりプライベートな動機でタグ付けを行っている可能性が高いことが示唆された．一方で，Instagram，RoomClip のユーザは Delicious，Flickr ユーザと比較して，よりオープンな動機でタグ付けを行っていることが示唆された．

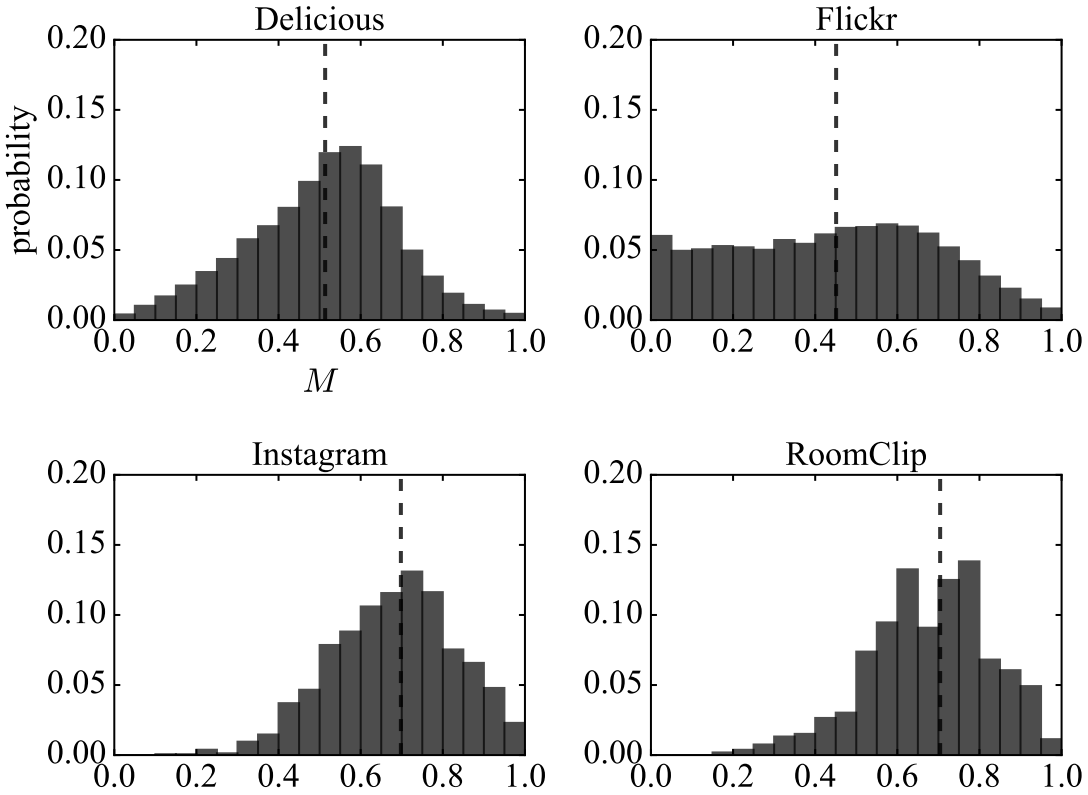


図 4.3: STS を採用する 4 つのサービスにおける，ユーザのタグを付ける動機 (M) の分布．横軸はユーザのタグを付ける動機 (M) であり，縦軸はその値を示すユーザの割合である．$M = 0$ に近いユーザのタグを付ける動機は典型的な分類者である．$M = 1$ に近いユーザのタグを付ける動機は典型的な説明者である．

先ほども述べたように，ω はタグ付けの動機による，実際に付けられるタグの偏りについては考慮していない．そこで，各サービスに参加するすべてのユーザに対して式 (4.3) により定義された M を計算する．これにより，実際に付けられるタグの偏りという観点から，ユーザのタグ付けの動機を測定する．ここでは，各サービスにおいて，200 回以上投稿し $H_{\mathrm{opt}}(R|T)$ の値が 0 ではないユーザのみを対象として分析を行った．

図 4.3 に，各サービスにおいて計算を行った M の分布を示す．横軸は M の値であり，20 個のビンに分割を行い，ヒストグラムの作成を行った．縦軸は，各ビンに含まれるユーザの割合である．前述したように，M の値が小さいユーザのタグ選択は分類者的であることを意味する．反対に，M の値が大きなユーザのタグ選択は説明者的であることを意味する．縦の破線は各サービスにおける M の平均値である．分布の形状を見ることで，各サービスに参加するユーザのタグを付ける動機を特徴付けることができる．左に偏った分布である場合には，そのサービスに参加するユーザのタグを付ける動機は分類者的であることを意味する．反対に，右に偏った分布である場合には，そのサービス参加するユーザのタグを付ける動機は，説明者的であることを意味する．

図 4.3 からわかるように，Instagram と RoomClip は，Delicious や Flickr と比較して，より説明者的な振る舞いでタグを付けるユーザが多数存在することが明らかになった．つまり，Instagram および RoomClip に参加するユーザのタグを付ける動機は，Delicious および Flickr のタグを付ける動機よりもオープンである．Flickr の場合には，分布の形状は比較的フラットである．つまり，他の 3 つのサービスと比較して，Flickr のタグ付けの動機は広く分布している．同時に，Flickr には多くの典型的な説明者が存在することを意味している．

上記の 2 つの分析から，Delicious と Flickr は，プライベートな動機でタグを付けるユーザが多く存在するサービスであるといえる．対照的に，Instagram と RoomClip ではオープンな動機でタグを付けるユーザが多く存在するサービスであるといえる．これらの結果は，4 つのサービスの主な使用法に関する，当初の直感的とも一致する．

4.3 結論

タグ付けをモデル化する上で広く利用される Yule–Simon 過程では，ランダムな確率で新たな種類のタグは生成される．このため，新規タグの生み出されるメカニズムについては，新たな種類のタグは常に生み出され続ける，ということ以上には特に何も述べていない．Yule–Simon 過程において考慮されるのは，各ウィンドウを無視した，投稿されたすべてのタグからなるマクロなタグのシーケンスである．タグの同時利用を考慮した場合には，マクロなタグのシーケンスに加えて，ミクロなタグのシーケンスを考えることが可能である．ここでいうミクロなタグのシーケンスとは，同一ウィンドウ内で一人のユーザによって生成されるタグのシーケンスである．実際のソーシャルタギングにおける新規タグの生成確率は，ウィンドウサイズごと，あるいは投稿するユーザごとに異なることが考えられる．そこで，ある投稿に対して同時に付けられるタグの数と新たな種類のタグの生成確率の間の相関に着目して分析を行った．これにより，新しい種類のタグが同一投稿内でどの生み出されているのかを明らかにする．分析の結果，α と ω の相関は，各サービスごとに異なる傾向を示すことが明らかになった．Delicious や Flickr といったサービスでは弱い正の相関があらわれることが明らかになった．対照的に，Instagram および RoomClip では負の相関があらわれることが明らかとなった．このことは，Yule–Simon 過程では考慮されていないメカニズムが新規タグの生成に対して働いていることを意味する．

各サービスで異なる相関が生じる理由を説明するために，ユーザのタグ付けの動機に着目し分析を行った．ここでは，タグ付けの動機をオープンと，プライベート 2 つの視点から明らかにする．オープンおよびプライベートなタグ付けの動機は，それぞれコミュニケーション指向および個人指向のタグ付け行動と対応する．分析の結果，Delicious や Flickr に参加するユーザの多くがプライベートな動機でタグを付けることが明らかになった．対照的に，Instagram および RoomClip に参加するユーザの多くが，Delicious や Flickr と比較して，オープンな動機でタグを付けることが明らかになった．これらの結果から，多くのユーザがプライベートなタグ付けを行うような，Delicious や Flickr といったサービスでは弱い正の相関が現れることが明らかになった．対照的に，Instagram および RoomClip で観察された負の相関はオープンなタグ付けに関連することを明らかにした．

ここでは、2 つの点について議論を行った．単一の投稿に対する新規タグの生成確率は，投稿に対して同時に付与されるタグ数と相関を持ち，その相関は各サービスごとに異なることを明らかにした．また，各サービスごとに異なる相関は，サービスに参加するユーザのタグを付ける動機によりある程度説明可能であることを明らかにした．これは，サービスの本質的な性質を識別する，新しい指標を設計する上で利用可能であることが考えられる．

第5章　SNSに参加するユーザの早期離脱予測

ウェブ利用者の増加とともに多数のSNSが誕生した．SNSの誕生により人々の情報取得方法は大きく変化し，SNSが情報源としての役割を持ちつつある．SNS内に投稿されるコンテンツは，サービスに参加するユーザ自身の手で作られる．このため，SNSが魅力的であり続ける，あるいは継続的に成長していくためには，新たなユーザの獲得と獲得したユーザの継続利用が重要である．

SNSにおけるユーザのサービス継続利用率は非常に低いことがこれまでに知られている．SNSへの新規参加は手軽である場合が多い．また，サービスから離脱した場合のペナルティも殆どの場合に存在しない．例えば，多くのサービスでは，メールアドレスを登録するだけでサービスを利用することが可能である．またいくつかのサービスでは，既存のSNSのアカウントを利用することで利用を開始することが可能である．このため，良くも悪くも気軽にサービスの利用を開始し，離脱を選択することが可能となっている．実際に，新規ユーザの半数以上がサービスを利用してすぐに離脱することがわかっている [3, 45, 14].

一般的に新規ユーザの獲得には，獲得したユーザを維持するよりも多くのコストが掛かかることが広く知られている [48]．このため，獲得した新規ユーザに対して，サービスの継続利用をいかにして促すかといった点に注目が集まっている．実際にこれまでの研究では，新規ユーザの継続利用を促すような介入方法が多数提案されている [19, 11]．このときに，早期離脱者が予め分類可能となる，あるいは早期離脱の要因が明らかとなることは，早期離脱率の高いユーザに絞った介入や，効果的な新規ユーザのサービス継続を促す介入を考える上で役に立つことから，重要な課題である．

SNSにおける早期離脱者の予測や早期離脱要因の分析には機械学習による手法が広く利用されてきた．例えば，ロジスティック回帰やランダムフォレストのような機械学習手法を用いることで，高い精度で早期離脱者の分類や要因分析が可能であることがわかっている [14, 3]. これらのモデルに対する入力は，サービスに参加してからある期間までの，総投稿数や総フォロワー数などである．つまり，人の手により設計された，ある期間におけるユーザの振る舞いを集計した，静的な特徴量である．このため，これらのモデルをそのまま利用するだけでは，時系列間の相関を考慮することができず，早期離脱者とそうでないユーザの時間的な振る舞いの変化を捉えて予測を行うことは難しい．

ある期間で同一の投稿数のユーザであっても，時間的なふるまいが異なる場合には，それぞれのユーザの早期離脱率は異なることが考えられる．ある期間で同一の投稿数の2人のユーザを考える．あるユーザはサービスに参加した直後だけ投稿を行うユーザである．もう1人のユーザは，観測期間全体でまんべんなく投稿を行うユーザである．前者のユーザはサービ

スに参加した直後にサービスから離脱している可能性が高いことが考えられる，反対に，継続的にサービスの利用を行っている後者のユーザの離脱率は低いことが考えられる．

Yang らはこのような仮設に基づき，新たに SNS に参加したユーザの振る舞いを，時系列として再帰的なニューラルネットワーク (RNN)，あるいは Long short-term memory (LSTM) に対する入力とすることで，既存の機械学習手法よりも高い精度で早期離脱者の分類が可能であることを示した [68]．RNN は深層学習の一種であり，ニューラルネットワークに再帰的な構造を持たせることで，時間的な相関を考慮可能なモデルとなっている．また，LSTM は RNN を拡張したモデルであり，これは長期の時系列の学習が可能なモデルとなっている [33]．

RNN あるいは LSTM を利用し，ユーザの振る舞いの時系列からユーザの早期離脱を予測することで，これまでよりも高い精度で予測を行うことが可能である．しかしながら，RNN，あるいは LSTM では明示的には考慮されていない情報がある．入力される時系列の，各時刻ごとの新規ユーザの早期離脱予測に対する影響度の違いである．これまでの研究から，SNS 内外での人の振る舞いは間欠的な振る舞いを示すことがわかっている [4, 67]．これは，ある時刻において頻繁に活動を行い，それ以外の時刻ではほとんど活動を行わないことを意味する．例えば，SNS に参加するユーザの場合にも，投稿や他者へのいいねなどは，ユーザのログイン時に短期間でまとまって行われていると考えることが自然である．新規ユーザも同様の振る舞いを示すと仮定すると，早期離脱者を予測する上で影響の大きな時刻とそうでない時刻が存在することが考えられる．このような，時間毎の早期離脱への影響度の違いを明示的に考慮することで，より高い精度で早期離脱者の予測が可能となることが考えられる．

入力される時間毎の影響度の違いを考慮する深層学習モデルの構造として，アテンションがある．より広い意味でのアテンションは深層学習モデルにおいて，重要度の違いを考慮する機構を意味する．RNN や LSTM といった時間的な相関を考慮可能な深層学習モデルにおいて，アテンションを導入し単語やフレーム毎の重要度の違いを明示的に考慮することで，文書や動画の分類を高い精度で行うことが可能であることがわかっている [69]．

アテンションにより可能となることは，入力される各時間毎の影響度の違いを考慮した，高精度の予想だけにとどまらない．アテンションにより導入された，各入力ごとの重要度を表す重みを可視化することで，入力のどの部分が予測に対して大きな影響を与えるのかを明らかにすることが可能である．例えば，文書や動画の分類において，単語やフレームのどの部分が大きな影響を及ぼすのかが解釈可能となる．

そこでこの章では，新規ユーザの時間的な振る舞いを入力として，時間毎の影響度の違いを考慮することで早期離脱者の予測を行う新たなモデルを提案する．提案モデルは，LSTM からの各時刻ごとの出力に対してアテンションをかけることで，時間毎の影響度の違いを考慮可能とする．提案モデルの有効性は，一般的な SNS と同様の機能を持つ SNS の新規ユーザを対象に，他のベースライン手法と予測精度の比較を行うことで確かめる．更に，提案モデルを用いて，早期離脱者の分類へ与える影響の違いを時間的，特徴量の面から明らかにする．最後に，特徴量の分析結果を踏まえて，SNS における早期離脱を防ぐ効果的な介入方法を検討する．

この章の具体的な貢献は以下である．

1. 時刻毎の重要度の違いをアテンション機構を用いて明示的に考慮することで，既存手法と比較して高い精度で早期離脱者の予測が可能となることを示した．
2. 新規ユーザの最初期の振る舞いが，早期離脱者の予測に大きな影響を及ぼすことを明らかにした．
3. 新規離脱者の予測では，新規ユーザ自身が行う行為が大きな影響を及ぼすことを明らかにした．一方で，他のユーザからのフィードバックの与える影響は少ないことを明らかにした．
4. 特徴量の分析結果を踏まえた，早期離脱を防ぐ効果的な介入方法の提案を行った．

5.1 時系列間の相関を考慮可能な深層学習モデル

まずはじめに，時系列間の相関を考慮可能な深層学習モデルの説明を行う．ここでは時系列のエンコードをする場合に広く利用される，2 つのモデルの詳細を説明する．

5.1.1 再帰的ニューラルネットワーク (RNN)

Recurrent Neural Network (RNN) は再帰的な構造を持つことで，過去の入力との相関を考慮した学習が可能な深層学習モデルである [39]．RNN は数式的には

$$h(t) = \tanh\left(W'_x x(t) + W'_h h(t-1) + b_h\right) \tag{5.1}$$

のように表現される．ここでは，$x(t)$ は時刻 t における RNN への入力ベクトルであり，$h(t)$ は時刻 t における RNN からの出力ベクトルである．W'_x，W'_h は RNN の入力である $x(t)$，及び $h(t-1)$ に対する重みベクトルであり，これらは学習時に更新される．また，b_h は学習時に更新される，バイアスベクトルである．以降，W に添字がつくものは学習時に更新が行われる重みベクトルをあらわし，b に添字がつくものを学習時に更新が行われるバイアスベクトルとする．

RNN は再帰的な構造を持つことから，時刻 t における出力には $t-1$ の出力を利用して計算される．実際に，$t-1$ の出力である $h(t-1)$ が $h(t)$ を計算する上で利用されている．このような再帰的な構造を持つことで，RNN は時間的な相関を考慮した予測を行うことが可能となっている．

RNN モデルの学習には，一般的に Back propagation through time を用いることで行われる．これは RNN を時間的に展開することで，層の深いニューラルネットワークとみなし，学習を行う手法である．Back propagation through time による学習は，時系列長が増加するに連れて，より深いニューラルネットワークを学習することに相当する．このため，層の深いニューラルネットワークを学習する上で問題となる，勾配消失が入力される時系列が長い場合に発生してしまう．時系列長が長い場合には学習時に勾配が消失してしまうため，RNN では長期の時間的な相関は考慮出来ないことが知られている [33].

5.1.2 Long short-term memory (LSTM)

長期の時系列の学習を行うことを目的として，勾配消失を防ぐための拡張を加えた再帰的ニューラルネットワークがこれまでにいくつか提案されている．代表的なものとして，Long short-term memory (LSTM) と呼ばれるモデルがある．LSTM では，内部状態を保持するように拡張を加えることで，長期間の時間的な相関を学習することが可能なモデルとなっている [33]. 内部状態を持つという点は共通に，LSTM の内部構造には様々なバリエーションが存在する．本研究では，その中でも代表的な拡張であり広く利用されている，Forget gate を加えた LSTM を利用した [20].

Forget gate を加えた LSTM の内部構造は

$$
\begin{aligned}
i(t) &= \sigma_h \left(W_{xi} x(t) + W_{hi} h(t-1) + b_i \right), \\
f(t) &= \sigma_h \left(W_{xf} x(t) + W_{hf} h(t-1) + b_f \right), \\
c(t) &= f(t) c(t-1) + i(t) \tanh \left(W_{xc} x(t) + W_{hc} h(t-1) + b_c \right), \\
o(t) &= \sigma_h \left(W_{xo} x(t) + W_{ho} h(t-1) + b_o \right), \\
h(t) &= o(t) \tanh \left(c(t) \right)
\end{aligned}
\tag{5.2}
$$

と表現される．$i(t)$ は Input gate と呼ばれ，LSTM において保持される内部状態 $c(t)$ に対して，入力される値をどの程度反映させるのかを調整する役割を持つ．$f(t)$ は Forget gate と呼ばれ，過去のセルの状態 $c(t-1)$ をどの程度書き換えるかを決定する役割を持つ．$c(t)$ は LSTM において保持される内部状態である．最終的な LSTM からの出力は $h(t)$ であり，これは Output gate である，$o(t)$ を活性化関数へ入力することでえられる値である．また，σ_h は任意の活性化関数を設定することが可能である．本研究では，特に断りのない場合にはハードシグモイド関数を利用する．

5.2 ユーザの早期離脱予測モデル

ここでは，各時刻毎の早期離脱者の予測への影響度の違いを考慮した提案モデルの詳細を説明する．図 6.3 に提案モデルの概念図を示す．提案モデルは大きく分けて 3 つの層からなる．1. 双方向 LSTM 層，2. アテンション層, 3. クラシフィケーション層の 3 つである．以下にその詳細を示す.

5.2.1 モデルへの入力

本研究ではユーザのイベント時系列を $\mathbf{s} = (s(1), ..., s(\frac{24S}{b}))$ と定義し，続く双方向 LSTM 層に対する入力とした．$s(t)$ は今回の予測に利用したイベントの種類数からなるベクトルである．各要素は等間隔でビンニングを行った，ある期間の各イベントの発生回数である．また，$\frac{24S}{b}$ は観測の最終時刻であり，S は観測期間 (日)，b はビンニング間隔 (時) であり 0 を含まない任意の自然数である．

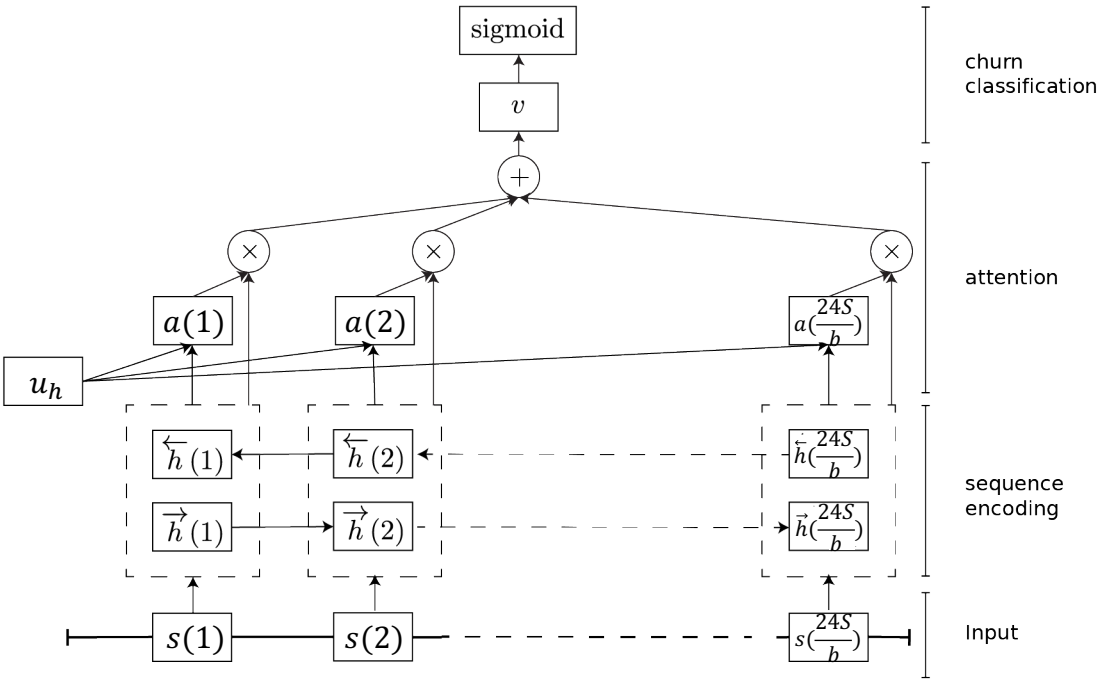


図 5.1: SNS においてユーザの早期離脱を予測する，提案モデルの概念図．提案モデルは大きく分けて 3 つの層からなる．下から順にそれぞれ，1. 双方向 LSTM 層，2. アテンション層，クラシフィケーション層である．ここでは，等間隔でビンニングしたユーザの各イベント数の時系列がモデルへの入力である．$s(t)$ は t 番目の入力であり，これが双方向 LSTM 層への入力である．双方向 LSTM 層からの出力である，イベントの表現ベクトル ($[\overrightarrow{h}(t), \overleftarrow{h}(t)]$) と，グローバルなイベントの表現ベクトル ($u_h$) の内積を Softmax 関数に入力することで正規化を行う．その結果，正規化された時刻 t の影響度 ($a(t)$) が得られる．これを双方向 LSTM 層からの出力と掛け，各時刻の合計をとり Sigmoid 関数へ入力することで，あるユーザが早期離脱する確率が出力される．

5.2.2 双方向 LSTM 層

ここでは，順方向と逆方向のイベント時系列から 2 つの LSTM を学習させ，2 つのモデルの出力を結合した [52]．これは双方向 LSTM と呼ばれ，順方向のみから学習させた場合よりも高次の情報を持つことを期待した．順方向と逆方向の時系列情報からそれぞれ学習を行った時刻 t での各 LSTM からの出力は，それぞれ $\overrightarrow{h}(t)$ と $\overleftarrow{h}(t)$ と定義する．順方向と逆方向，それぞれから学習を行った 2 つの LSTM による出力の組み合わせ方は任意である．ここでは，双方向 LSTM の最終的な出力としてそれぞれの LSTM の出力を結合させ，

$$h(t) = [\overrightarrow{h}(t), \overleftarrow{h}(t)] \tag{5.3}$$

とした．

5.2.3 アテンション層

SNS 内外での人の振る舞いは間欠泉的な振る舞いを示すことがこれまでの研究からわかっている [4, 26]．また最初期の人の振る舞いが，その後の長期的な人の振る舞いに大きな影響を与える可能性が示唆されている．上記の 2 つの理由から，各時刻におけるイベントが早期離脱に対して与える影響は異なることが考えられる．そこで，各時刻の離脱への影響度の違いを考慮するために，各時刻の LSTM 層からの出力に対してアテンションをかけることを考える．これにより，各時刻の予測への影響度の違いを明示的に考慮することが可能となる．

今回の研究では，早期離脱予測の精度を上げるだけでなく，どの時刻の与える影響が大きいのかも明らかにしたい．各時刻ごとの影響度の違いを明らかにすることは，時間的な側面から効果的な介入方法を検討することにつながる．そこで，今回は解釈を容易にするために，Yang らの研究を参考にして以下のような単純なアテンションを導入した [69].

$$
\begin{aligned}
u(t) &= \tanh(W_u h(t) + b_w)^{\mathrm{T}} u_h, \\
a(t) &= \frac{\exp(u(t))}{\sum_t \exp(u(t))}, \\
v &= \sum_t a(t) h(t).
\end{aligned}
\tag{5.4}
$$

$u(t)$ は早期離脱予測に対する，時刻 t の影響度であり，これは正規化する前の値である．$u(t)$ はイベントの表現ベクトルとグローバルなイベントの表現ベクトル (u_h) との内積により計算される値である．u_h はランダムな値で初期化し，学習時に更新される重みである．$u(t)$ は softmax 関数により正規化を行う．その結果が $a(t)$ である．得られた $a(t)$ は，各時刻の双方向 LSTM からの出力に対して重み付けを行う上で実際に用いられる値である．v はアテンション層での最終的な出力であり，$a(t)$ で重み付けされた LSTM 層の出力の全時刻での合計値となっている．

5.2.4 クラシフィケーション層

本研究では，早期離脱者に対して $\hat{y} = 1$ のラベルを割り当て，早期離脱者以外には $\hat{y} = 0$ のラベルを割り当てた．これにより，今回の早期離脱者の予測を 2 値分類問題として扱う．最終的な出力層にはシグモイド関数 ($sigmoid(x) = \frac{1}{1+\exp(-x)}$) を利用した．つまり，新規ユーザが早期離脱者である確率が今回のモデルの最終的な出力となる．あるユーザが早期離脱者かどうかの確率は，アテンション層での最終的な出力 v を用いて以下のように計算され，

$$
y = sigmoid(W_v v + b_v). \tag{5.5}
$$

5.2.5 モデルの学習方法

誤差関数はバイナリクロスエントロピー誤差

$$
loss(y, \hat{y}) = y \log \hat{y} + (1 - y) \log (1 - \hat{y}) \tag{5.6}
$$

を採用した．モデルの学習には Back propagation through time を利用した [49].

5.3 早期離脱の予測に利用した SNS データ

5.3.1 分析に利用した SNS の説明

実際に運用されている商用 SNS のデータを利用しモデルの評価と分析を行った．今回利用したデータは，それぞれのユーザに対してランダムな id を割り当てることで匿名化されており，個人の識別ができない様になっている．分析に利用した SNS では，参加するユーザは写真を投稿し，その写真を介してコミュニケーションを行う．このことから，Instagram に似た種類の SNS ということが可能である．今回の分析に利用した SNS では，SNS 内におけるユーザの振る舞いをタイムスタンプとともに保存している．また，ユーザの登録した時刻もタイムスタンプとともに保存されており，あるユーザがサービスに登録した時刻からの分析が可能である．

SNS に投稿されるコンテンツには，投稿したユーザの手で自由なタグが付与される．付与されるタグは主に写真の分類や，検索に利用される．写真の投稿以外にユーザがとる主な行動は，他のユーザの写真に対する，良いねやコメント，クリップがある．また，他のユーザのフォローも可能となっている．いいねやコメントは Instagram とほとんど同じ機能である．いいねは写真に対して好意を持った場合にとられる機能となっている．コメントは写真を介して，文字によるコミュニケーションを行いたい場合に用いられる機能である．クリップはユーザが気に入った写真をリスト化し保存を可能とする機能である．また，他のユーザのフォローを行うことで，より密なコミュニケーションを取ることが可能となっている．このように，今回分析に利用したサービスは，Twitter や Instagram のような一般的な SNS と同様の機能を有するサービスであるといえる．

新規ユーザ自身が行うアクションと，他のユーザから受け取るアクションをリアクションと呼ぶ．アクションは新規ユーザ自身が行う，写真の投稿，いいね，クリップ，コメント，フォローとした．いいね，クリップ，コメントは他のユーザの投稿した写真に対して行われる行動である．フォローは他のユーザに対して行う行動である．リアクションは新規ユーザが他のユーザから受け取る，いいね，クリップ，コメント，フォローとした．また，アクションとリアクションをまとめてイベントと呼ぶ．

前処理として，前日比で 2 倍以上ユーザの登録数が増加した日に，サービスへ新規登録を行ったユーザは分析から除外した．これは，企業による大規模な広告や，スパムによるノイズの影響を除くための処理である．

5.3.2 ユーザの基本統計

まずはじめに，今回の論文で扱う早期離脱者の定義を説明する．図 5.2 に今回扱った早期離脱予測の概念図をのせる．本研究では，あるユーザがサービスに参加してから S 日間を観測期間とした．S 日間に一度以上アクションを行ったユーザを分類対象とした．それ以外の

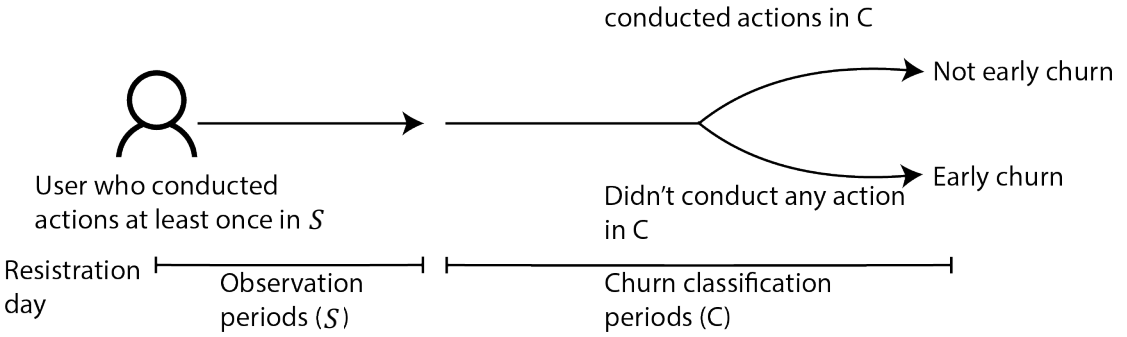


図 5.2: SNS における早期離脱の定義と早期離脱予測の概念図．あるユーザがサービスに参加してから S 日間を観測期間とした．S 日間に一度以上アクションを行ったユーザを分類対象とした．また，観測期間 S 日以降，C 日間一度もアクションを行わなかったユーザを早期離脱ユーザと定義した．S 日間のユーザのイベント時系列を入力として，そのユーザが離脱するかどうかの予測を行った．

ユーザは分析から除いた．観測期間 S 日以降，C 日間一度もアクションを行わなかったユーザを早期離脱ユーザと定義した．S は任意の値を設定することが可能である．また，C に関しても離脱判定として妥当な任意の値を設定することができる．ここでは，Dror らの研究を参考に $S = 7$ とした [14]．また，離脱判定期間は $C = 90$ に設定した．今回分析に利用したデータにおいて，最大アクション間隔が 90 以上のユーザの割合は十分に小さいことを確かめており，妥当なしきい値であると考えた．

新規ユーザの基本統計を表 5.1 に載せる．今回分析対象となる総ユーザ数は 78,479 人である．新規ユーザのうち早期離脱者数は 43,117 人であり，非早期離脱者数は 35,362 人であった．また，早期離脱率は約 55%であり，他の SNS と同様に高い確率で早期離脱が発生していることがわかる．また，早期離脱者とそうでないユーザの各イベント数の平均値を比較すると，早期離脱者のアクション，リアクション数ともに非早期離脱者に比べて小さいことがわかった．このことから，これらの特徴量が早期離脱者の予測に対して影響を及ぼしていることが考えられる．

新規ユーザのイベント発生間隔 (τ_e) とアクション発生間隔 (τ_a) がそれぞれどのような分布となっているのかを確認する．それぞれのイベント発生間隔の分布がべき分布を示す場合には，ユーザの振る舞いが疎な部分と密な部分から構成されることを意味する．図 5.3 に τ_e と τ_a の分布をそれぞれ示す．横軸はイベント発生間隔 (τ_e) またはアクション発生間隔 (τ_a) である．ここでは，両対数グラフで表示するために，それぞれの実際の値に 1 を足した値を示した．縦軸はそれぞれの確率である．黒の点線は補助線である．τ_e に注目すると，べき分布を示すことがわかった．また，τ_a についても τ_e と同様に，べき分布を示すことがわかった．更に，τ_a の分布の傾きは約 20 分を境に変化することがわかった．今回分析に利用した SNS の新規ユーザの振る舞いに関しても，各時刻ごとにイベントやアクションが密な時刻と疎な時刻が存在し，間欠的であることがわかる．τ_a の傾きの変化は，ユーザの振る舞いが変化して

表 5.1: 分析に利用した，SNS における早期離脱ユーザと非早期離脱ユーザの統計値.

	Early churnres (43,117)		Not early churnres (35,362)	
Action	Mean	SD	Mean	SD
Number of likes	1.72	15.8	15.1	70.5
Number of comments	0.10	1.63	1.73	7.96
Number of clips	1.06	4.26	3.38	8.26
Number of follows	1.43	4.07	4.56	21.8
Number of posts	0.12	0.93	1.39	4.86
Reaction				
Number of likes	1.43	12.3	21.8	107
Number of comments	0.09	1.17	1.80	8.77
Number of clips	0.02	0.41	0.45	3.18
Number of follows	0.19	1.18	1.62	10.1

いることを意味し，これは一度ログインしたユーザは複数のアクションを 20 分ほどの間隔で行っていることが考えられる.

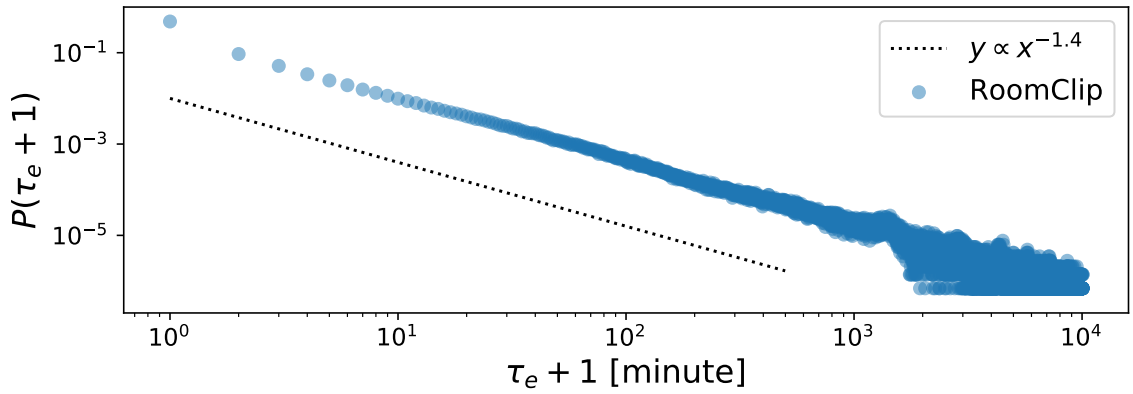
5.4 予測精度の比較実験

提案モデルの有効性を確かめる．ここでは，これまでの研究で用いられた手法と，提案手法の早期離脱ユーザの予測精度の比較を行うことで確かめる.

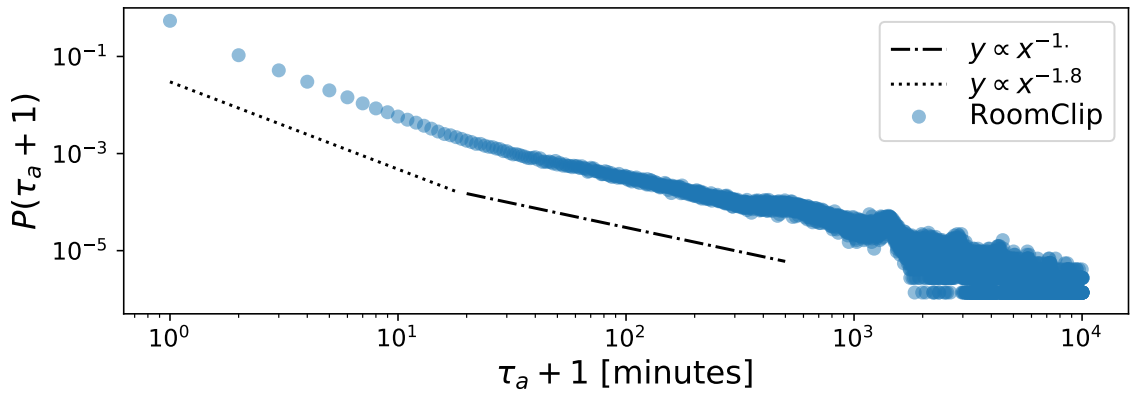
5.4.1 評価指標

Area Under the ROC Curve (AUC) を利用することで，各モデルの評価を行う．AUC は各ラベルに偏りがある場合にもモデルの評価を行うことが可能な指標となっている．ROC は分類しきい値を変更しながら，偽陽性率 (FPR) の関数として真陽性率 (TPR) を描画することにより得られる値である．TPR と FPR はそれぞれ，$TP/(TP+FN)$，$FP/(TN+FP)$ である．TP は早期離脱者を正しく識別出来た人数，TN は早期離脱者を誤って識別した人数，FP は非早期離脱者を正しく識別出来た人数，FN は非早期離脱者を誤って識別した人数である．AUC では完全にランダムに分類するモデルの場合には 0.5 を示し，完璧に各ラベルを分類可能なモデルの場合には 1 の値を示す.

本研究では，未学習のデータに対する予測性能の高さを見るために，10-分割交差検証を行いモデルの評価を行った．10-分割交差検証では，対象とするデータをランダムに 10 分割する．10 個に分割されたもののうち，1 個をテストデータとして性能評価に利用し，残りのデータをモデルの学習に利用する．この処理を 10 個の分割に対して繰り返し行うことで，評価指標の平均をとるものである．また，トレーニングデータのうち，1 割をバリデーションデータ



(a) イベント発生間隔 (τ_e) の分布.



(b) アクション発生間隔 (τ_a) の分布.

図 5.3: 新規ユーザのイベント発生間隔 (τ_e) とアクション発生間隔 (τ_a) の分布．横軸はイベント発生間隔 (τ_e) またはアクション発生間隔 (τ_a) である．ここでは，両対数グラフで表示するためにそれぞれの実際の値に 1 を足した値とした．縦軸がそれぞれの確率である．黒の点線は補助線である.

とした．バリデーションデータは，グリッドサーチによるハイパーパラメータの最適化に利用した．

5.4.2 前処理

今回扱う SNS のデータは，早期離脱したユーザと早期離脱しなかったユーザのラベル数が不均質である．このように各ラベルの数に偏りが生じている場合にはモデルの学習結果が多い方のラベルに依存してしまう．このような問題に対処するために，アンダーサンプリングを利用した．これは，各ラベル数に偏りが存在する場合に用いられる典型的な手法である．アンダーサンプリングでは，多い方のラベルを少ない方のラベル数に合わせてランダムに再サンプリングし，両方のラベル数を均等にする．

各イベントはユーザがサービスに登録してから一定時間ごとに等間隔でビンニングを行い，各イベントの発生回数の集計を行った．今回は $b = 1, 3, 6, 12, 24$ 時間ごとにビンニングを行った．ビンニングの間隔が小さい場合には，より詳細な時間的なふるまいの変化を捉えたデータであることを意味する．反対に，ビンニングの間隔が大きな場合には静的な特徴量からなるデータとなる．また，図 5.3 から示されたように，新規ユーザのイベント時系列は素な部分と密な部分からなる．このため，b が小さい場合には，各時刻ごとの影響度の違いがより顕著になることが考えられる．

5.4.3 比較する手法の説明

本研究では，SNS におけるユーザの早期離脱を予測する上で利用される 2 つの機械学習モデル (ロジスティック回帰，ランダムフォレスト) と，アテンション機構を持たない LSTM モデルである，LSTM w/o Attention をベースラインモデルとして提案手法との比較に利用した．

ロジスティック回帰は 2 値分類問題に対して利用される古典的な回帰モデルである．ランダムフォレストはアンサンブル学習の一種であり，大量の弱分類木の多数決を取ることで，汎化性能を高めたモデルとなっている．これまでの研究では，SNS におけるユーザの離脱をモデル化する際にランダムフォレストが高い予測性能を発揮することがわかっている [14]．

各機械学習モデルの実装には Python の機械学習ライブラリである scikit-learn を利用して行った [44]．ランダムフォレストにおける，弱分類木の数は固定値として 500 とした．木の分割を停止する分割後のデータ数と，各木の学習に利用する特徴量の数はグリッドサーチにより最適化を行った．ロジスティック回帰における特徴量の正規化方法はグリッドーサーチにより最適化を行った．それ以外のパラメータについては分類精度に対する影響は低いと考え，scikit-learn にて予め設定されているデフォルトの値を利用した．

時系列を考慮したモデルである，提案モデル (LSTM w/ Attention) と LSTM w/o Attention，2 つのモデルを構築し評価を行った．LSTM w/o Attention はアテンション層を除き，双方向 LSTM の各出力の平均をシグモイド関数へ渡し予測を行うモデルである．両モデルの実装は python の深層学習 API である Keras を用いて行った [12]．LSTM w/ Attention と，LSTM w/o

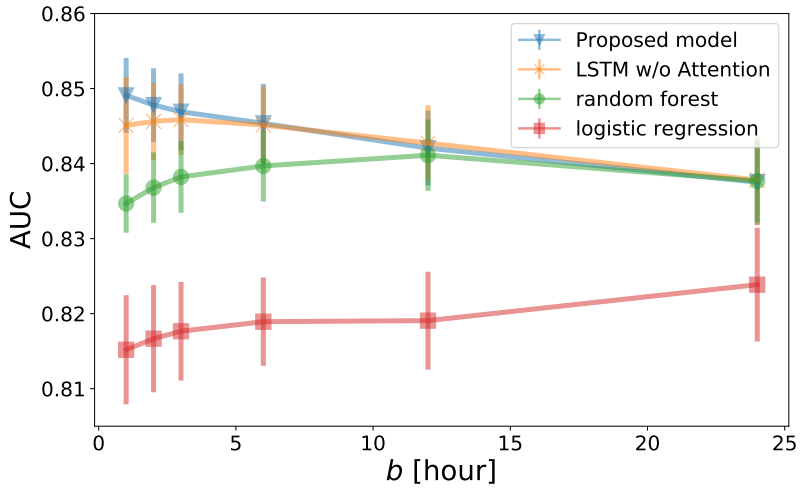


図 5.4: ビンニング幅 (b) を変化させながら，SNS における早期離脱ユーザの AUC 値による分類精度を各モデルで比較した場合の結果．横軸がビンニング幅 (b) であり，縦軸が AUC 値である．各点が 10-分割交差検証を行ったときの平均 AUC 値である．棒線は標準偏差である．提案手法 (LSTM w/ Attentio)，LSTM w/o Attention，ランダムフォレスト，Logistic regression における最も識別性能が高いビンニング幅を選択した場合の AUC 値の平均と分散はそれぞれ ($Mean = 0.849, SD = 0.00499$)，($Mean = 0.845, SD = 0.00469$)，($Mean = 0.841, SD = 0.00390$)，($Mean = 0.823, SD = 0.00757$) である．

Attention のハイパーパラメータは共通とした．入力層と LSTM セルの数は 10 とした．入力は z-score により正規化をおこなった．モデルの学習には Adam を利用し，エポック数は 30，バッチサイズは 64 とした．それ以外の値は Keras のデフォルトの設定を利用した．LSTM ベースの手法は計算時間の観点から，グリッドサーチによるハイパーパラメータのチューニングは行わなかった．

5.4.4 予測精度の結果

図 5.4 にビンニング間隔 (b) を変化させたときの各モデルの AUC の変化を示す．横軸がビンニング幅 (b) であり，縦軸が AUC 値である．各点が 10-分割交差検証を行ったときの平均 AUC 値である．棒線は標準偏差である．提案手法 (LSTM w/ Attentio)，LSTM w/o Attention，ランダムフォレスト，Rogistic regression における最も識別性能が高いビンニング幅を選択した場合の AUC 値の平均と分散はそれぞれ，($Mean = 0.849, SD = 0.00499$)，($Mean = 0.845, SD = 0.00469$)，($Mean = 0.841, SD = 0.00390$)，($Mean = 0.823, SD = 0.00757$) である．$b = 24$ を選択した場合，つまり静的な特徴量から予測を行う場合にはランダムフォレストが高い分類精度を発揮することがわかった．ビンニング幅を短くし，入力となる時系列長が長くなることで，LSTM ベースの手法の性能がランダムフォレストを上回るということがわかった．また，アテンション機構を持たない LSTM では，小さい b を選択し入力する時系列帳が長くなっ

た場合に，ある一定の b を超えたところで予測精度は向上しなくなる．一方提案手法では，b の値を小さくした場合にも予測性能が向上しつづけている．更に，$b = 1$ としたときの提案手法の予測性能が最も高くなることがわかった．

提案モデルが実際に早期離脱者を予測する上で，時刻毎の影響度の違いを考慮して予測を行っていることを確かめる．図 5.5 に提案モデルが早期離脱者と非早期離脱者を正しく予測できたときのイベントと a の関係を示す．左図は非早期離脱者，右図が早期離脱者の例である．上図の縦軸は各時刻の $a(t)$ の値であり，横軸がビンのインデックス (t) である．点線は各時刻の影響度が一定だった場合 ($\frac{1}{168} = 5.95 \times 10^{-2}$) の理論値である．下図はイベントの発生した時刻に黒い縦線を描画したものである．イベントはアクションとリアクションごとに分け，色の濃さが濃いほど，イベントの密度が高いことを表す．図 5.5 が示すように，$a(t)$ が理論値を下回る時刻と，上回る時刻が存在することがわかった．このように，実際に提案モデルは各時刻の入力に対して異なる重要度を考慮して早期離脱者の予測を行っていることがわかる．

各時刻ごとの相対的な重要度の違いを明示的に考慮する提案モデルが，小さなビンニング間隔 (b) を選択した場合に，最もよい予測性能を示すことを確認した．イベントの時系列は b が小さい値をとるようになるにつれて，詳細な時間的変化を持った入力から予測することになる．このときに，b が小さくなることで重要度の低い時刻の数が増えることも考えられる．アテンション機構を持たない LSTM モデルも他の機械学習によるベースラインモデルも適切に影響度の低い時刻に対処することができないため，b が小さい値を示す場合に予測性能が悪化してしまっていることが考えられる．一方で，図 5.5 に見られるように，提案モデルでは b が小さい値を取る場合にも異なる重みを付与することで，適切に処理できていることが考えられる．これらの理由により，小さな b を取る場合に提案手法が最も良い予測性能を示していることが考えられる．

5.5 特徴量の分析と，分析結果を踏まえた介入案の検討

新規ユーザの早期離脱を予測する上で，ユーザのどのような振る舞いが予測に対して与える影響が大きいのかを，特徴量の分析を通して明らかにする．ここでは，以下の 2 つの視点から特徴量の分析を行う．

- 早期離脱者の予測に対して，どの時刻の与える影響が大きいのかを明らかにする．
- 早期離脱者の予測に対して，どのイベントが与える影響が大きいのかを明らかにする．

上記の分析は，予測精度の実験において最も良い予測性能を示した，$b = 1$ としたときの提案手法を利用して行った．モデルの評価と同様に，それぞれの結果は 10-分割交差検証を行い，10 回の平均を求めた．

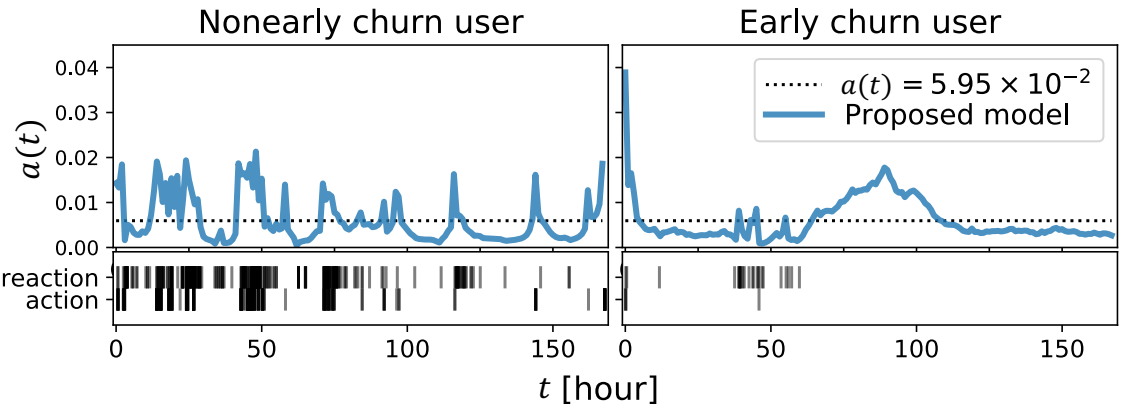


図 5.5: SNS において，提案モデルにより早期離脱ユーザの予測を行った場合のイベントとアテンションの強さ ($a(t)$) の関係の可視化例．左図は非早期離脱者，右図が早期離脱者の例である．上図の縦軸は各時刻の $a(t)$ の値であり，横軸がビンのインデックス (t) である．点線は各時刻の影響度が一定だった場合 ($\frac{1}{168} = 5.95 \times 10^{-2}$) の理論値である．下図はイベントの発生した時刻に黒い縦線を描画したものである．イベントはアクションとリアクションごとに分け，色の濃さが濃いほど，その時刻のイベントの発生数が多いことを表す．

5.5.1 早期離脱予測への時間的な影響

提案モデルにおけるアテンションの可視化を行うことで，早期離脱予測への時間的な影響度の違いを明らかにする．これは，早期離脱者の離脱を防ぐうえで，どの時刻での介入が効果的であるのかを考慮する際に役立つことが考えられる．ここでは，10-分割交差検証を行った場合に，各分割のテストデータにおける $a(t)$ の平均を $a_f(t)$ 定義する．$a_f(t)$ を各時刻ごとに計測することで早期離脱予測に対する各時刻のグローバルな影響度の違いを明らかにする事が可能である．

図 5.6 は 10-分割交差検証を行ったときの $a_f(t)$ の平均である．図 5.5 の上図とほぼ同一の図であるが，縦軸が $a_f(t)$ の平均であるという点で異なる．縦軸は $a_f(t)$ の平均であり，横軸がビンのインデックス (t) である．点線は補助線であり，これは各時刻の影響度が $b = 1$ で，一定だった場合の理論値である (5.95×10^{-2})．また，薄い青は標準偏差である．

図 5.6 を見てわかるように，$a_f(t)$ は一定値ではなく，各時刻ごとに早期離脱予測への影響度は異なることがわかった．$t < 30$ と $t > 155$ の場合には，$a_f(t)$ が一定だった場合の理論値よりも大きな値を示すことが明らかとなった．特に，$t = 0$ の場合に最も大きな値を示すことが明らかとなった．

今回の結果から，ユーザがサービスに参加した最初期の振る舞いが，ユーザの長期的なふるまいである，離脱行動に対して大きな影響を及ぼすことが明らかとなった．最初期の予測に対する影響度が大きい原因として考えられるのは，最初期の早期離脱者と非早期離脱者の振る舞いが異なり，予測に対して重要な情報を含んでいることである．また，直感的には前日に何らかのアクションを行ったユーザは次の日にもアクションを起こす可能性が高いことが

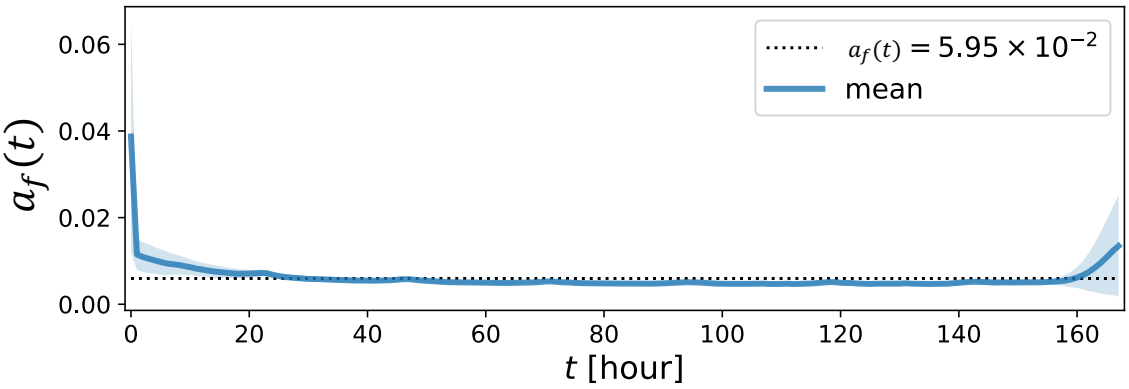


図 5.6: SNS における早期離脱ユーザの予測を行った場合の，アテンションの平均．図 5.5 の上図とほぼ同一の図であるが，縦軸が $a_f(t)$ の平均であるという点で異なる．縦軸は $a_f(t)$ の平均であり，横軸がビンのインデックス (t) である．点線は補助線であり，これは各時刻の影響度が $b = 1$ で，一定だった場合の理論値である (5.95×10^{-2})．また，薄い青は標準偏差である．

考えられる．このことは，図 5.3 において，新規ユーザのアクション間隔がべき分布を示すことからも確認できる．結果として，$S = 7$ に近い時刻でアクションを行ったユーザは，$S = 7$ を超えた後もアクションを行いやすく，これが $S = 7$ に近い時刻で $a_f(t)$ が高くなっている理由として考えられる．

5.5.2 早期離脱予測への各イベントの影響

次に，どのイベント時系列が早期離脱者の予測に対して大きな影響を与えているのかを明らかにする．ここでは，モデルの学習に利用したライク，コメント，クリップ，フォロー，投稿の中から 1 つの時系列のみを入力とする．このときのモデルの予測性能の変化をみることで，各イベント時系列が単体として離脱予測に対して与える影響の大きさを明らかにする．また，各イベントはリアクションとアクションで区別を行った．

図 5.7 に各イベントのみをモデルへの入力として 10-分割交差検証を行ったときの予測性能の変化を示す．縦軸が各イベントをあらわし，横軸がそのイベントだけを学習させ早期離脱予測を行ったときの AUC 値である．また，all はアクションすべて，あるいはリアクション全てから学習を行った場合の結果である．黒い点線は補助線であり，これはすべてのイベントから学習を行ったときの AUC 値である．エラーバーは標準偏差である．上図がアクションのみを学習させた結果であり，下図がリアクションのみを学習させた結果である．アクションすべてを学習させるだけで，リアクションを含めたすべてのイベントを学習させた時と同程度の分類性能を示すことがわかった．また，各アクション単体で見ると，ライクとフォローの分類性能が高く，コメントとポストの分類性能は低いことがわかった．一方で，リアクショ

ンすべてを学習させた場合にも，予測精度は各アクションのみを学習させた場合と同程度であり，低い分類性能であることが明らかとなった．

サービスに参加した直後のユーザの総投稿数は早期離脱者かどうかによらず少なく，他のユーザとの関係性も希薄である．ライクやいいねは投稿に対して行われるものであり，投稿数が少ない段階では他のユーザからのフィードバックである，リアクションを受け取ることは少ない．結果として，早期離脱者の予測に対してリアクションが与える影響が小さくなっていることが考えられる．また，写真の投稿やコメントにかかる時間的あるいは心理的なコストは，ライクやフォローに比べて高い．例えば，写真を投稿するためには写真の撮影や，加工を行う必要があり，コメントを行うためには文章を考える必要がある．一方で，ライクやフォローはボタンを押すだけで完結するアクションである．このため，ライクやフォローといった比較的かんたんな行為に，早期離脱者とそうでないユーザを区別可能な差が生じていることが考えられる．

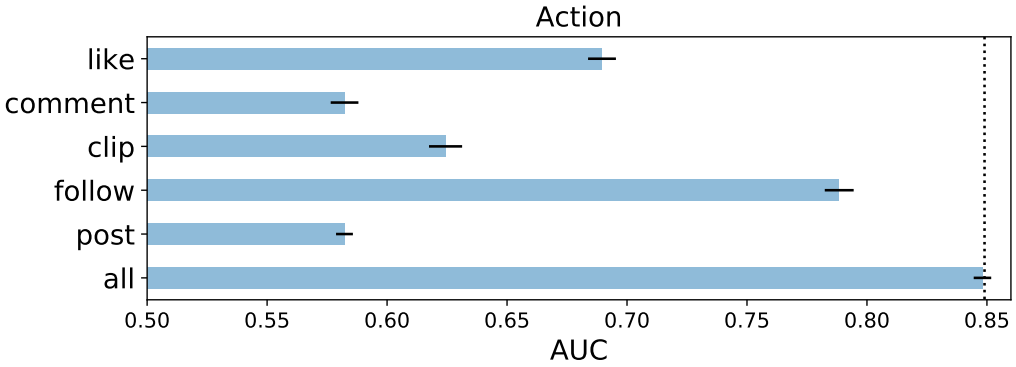
5.5.3 分析結果を踏まえた介入案

モデルに対して入力した各特徴量が，早期離脱者の予測に対して与える影響度の分析を行った．ここでは，イベントの種類と時刻，2 つの観点から行った．その結果，サービスに参加してから最初期の振る舞いが離脱予測に対して大きな影響を及ぼすことを明らかにした．また，新規ユーザ自身が行うフォローやライクといった比較的簡単に行える行為が，早期離脱者の予測に対して大きな影響を与えていることを明らかにした．一方で，新規ユーザは他のユーザとの関係性が希薄であり，他者からのリアクションの与える影響は小さいことを明らかにした．

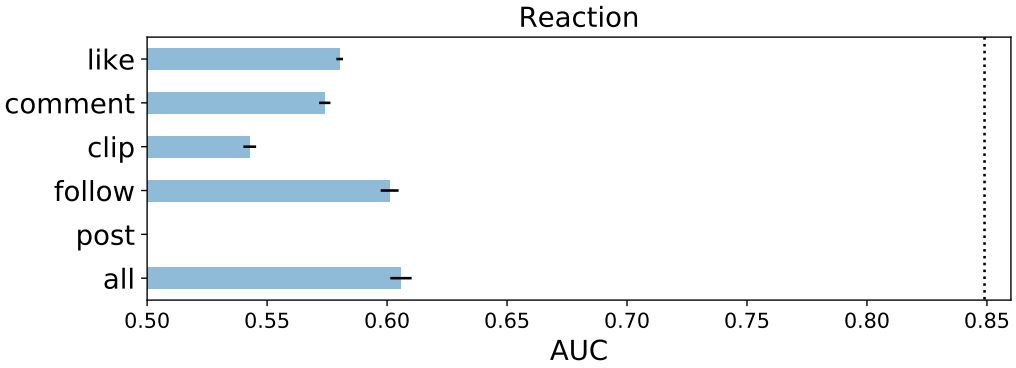
これらの分析結果を踏まえて具体的な介入方法を考える．比較的コストの少ない介入として，サービスに参加した直後に，フォロワー数の多いユーザや，いいね数の多い写真をランダムに掲示することが有効である可能性がある．また，より高度な介入として，対象とするユーザの過去の振る舞いを考慮することで，ユーザの好みを反映したフォロワーの推薦や写真の推薦が有効であることが考えられる．しかしながら，新規ユーザの場合には，サービスに参加した直後であり，推薦を行う上で利用可能なデータが少ない．このような場合には推薦の精度は低くなってしまう．Freyne らがしたように，他のサービスの利用履歴を用いた推薦を行うことで，より効果的な推薦と介入につながることが考えられる [19].

5.6 まとめ

SNS では参加障壁が低く，離脱時のペナルティが存在しないことから新規ユーザの早期離脱率が高いことがこれまでに知られている．SNS において，コンテンツを生み出すのはサービスに参加するユーザである．このため，新たに参加したユーザの維持が SNS の成長と密接な関係を持つ．SNS における早期離脱者の正確な予測や，早期離脱の要因が明らかになることは，効果的な早期離脱防止策の考案につながることから重要な課題である．SNS 内外での



(a) 10-分割の交差検証において，各アクションのみからユーザの早期離脱予測を行った場合の平均 AUC 値の変化.



(b) 10-分割の交差検証において，各リアクションのみからユーザの早期離脱予測を行った場合の平均 AUC 値の変化.

図 5.7: 10-分割の交差検証において，各イベントのみからユーザの早期離脱予測を行った場合の平均 AUC 値の変化．縦軸が各イベントをあらわし，横軸がそのイベントだけを学習させ早期離脱予測を行ったときの AUC 値である．また，all はアクションすべて，あるいはリアクション全てから学習を行った場合の結果である．黒い点線は補助線であり，これはすべてのイベントから学習を行ったときの AUC 値である．エラーバーは標準偏差である．上図がアクションのみを学習させた学習させた結果であり，下図がリアクションのみを学習させた結果である．

人の振る舞いは間欠的な振る舞いを示し，早期離脱者の識別をする上で情報量の多い時刻とそうでない時刻が存在することが考えられる．そこで本研究では，時刻に対するアテンション機構を加えた LSTM を利用することで，明示的に各時刻毎の早期離脱予測への影響度の違いを考慮したモデルを構築した．

Twitter や Instagram と似た機能を持つ SNS を対象に，既存モデルと早期離脱者の予測精度の比較を行い，提案モデルの有効性を確かめた．更に，モデルへの入力に利用した各特徴量

が離脱予測に対して与える影響度の分析をイベントの種類と時刻，2 つの観点から行った．その結果，サービスに参加してから最初期の振る舞いが早期離脱予測に大きな影響を及ぼすことを明らかにした．また，新規ユーザは他のユーザとの関係性が希薄であり．自分自身が行う行為が早期離脱者かどうかを予測する上で大きな影響を及ぼすことを明らかにした．今回の結果は早期離脱可能性の高いユーザに対してコストを掛けた介入方法が検討可能となるだけでなく，具体的な介入方法を検討する上でも役に立つものである．今回の結果は，Twitter, Instagram といった一般の SNS データにも適応可能であり，今回の結果が SNS 発展の一助となることが考えられる．

第6章　従業員の早期離脱予測

テクノロジーの進歩により，世の中の流行がこれまで以上に早く変化するようになった．同業他社でうまくいったサービスやアイディアが手軽に模倣，あるいは改良が可能となり，企業はより競争的な環境にさらされている．企業においてサービスを提供するのは，その企業で働く従業員である．競争的な環境で勝ち抜くためには，各従業員の能力を最大限に発揮させることで，提供するサービスの質を高める必要がある．このため，Human Resource Management (HRM) に高い関心が持たれている．

従業員の離脱が直接，あるいは間接的に企業に与える負の影響は大きい．従業員が離脱した場合には離脱した従業員の穴を埋めるために，新たに従業員の雇用を行う必要がある．また，離脱した従業員の能力に，新たに雇用した従業員の能力を近づけるためには，多くの時間やコストがかかる．実際に，従業員が離脱することで，採用コストの増加や，提供するサービスの質の質が低下してしまうことがこれまでに知られている [51]．このため，従業員の離脱は HRM における最も重要な課題の一つとして，これまでに多数の研究が行われている．

離脱する可能性の高い従業員が予め予測可能となることで，離脱可能性の高い従業員に対して面談やトレーニングといったコストをかけた介入を行うことが可能となる．効果的な介入を行うことで，従業員の離脱率が減少することが考えられる．この時に，従業員の予測に対して影響の高い特徴量が明らかとなることで，効果的な介入方法を検討することが可能となる．また，従業員の離脱を見越して早めに採用活動を行うことが可能となる．これにより，従業員の離脱によるサービスの質の低下を防ぐことが可能である．このため，従業員の離脱を予測するモデルや，モデルを利用した要因分析がこれまでの研究で多数行われている [57].

これまでの研究から，新規従業員の離脱率は特に高いことが知られている [25]．雇用前の限られた情報に基づいた働き方の予測と，実際の業務の間には大きな差があることがその理由として考えられる．従業員の離脱を予測するモデルはこれまでに多数提案されているものの，従業員の早期離脱を予測するモデルはこれまでにほとんど提案されていない．従業員の早期離脱を予測するモデルがほとんど存在しない理由には，新規従業員の早期離脱予測を行う重要性は低いと考えられていたためである．例えば，早期離脱する可能性の高い人材を採用しないことで，従業員の早期離脱を防ぐことが可能である．このため，一旦採用を行った新規従業員の離脱を防ぐのではなく，採用時に候補者のフィルタリングを行うという方法が取られる場合が多い [57].

一方で，人材不足により従業員の確保が課題となっている業界においては，採用時のフィルタリングは有効な手法であるとは言えない．むしろ，雇用した従業員にいかに長く働いてもらうのかを考慮する必要がある．従業員の確保が課題となっている業界に，飲食業がある．

飲食業界全般の離脱率は高く，人手不足が大きな問題となっている [66]．更に悪いことに，多くの先進国では少子高齢化により労働可能人口が減少している．このため，これまで以上に従業員を集めることが難しくなることが考えられる．特に，日本ではその傾向が顕著であり，有効求人倍率は 2013 年 の 0.8 倍から年々増加し，2018 年には 1.50 となっている [18]．そのうち，2019 年 1 月の飲食業における有効求人倍率は 3.28 – 3.36 となっている [58]．そこで本研究では，日本の飲食店チェーン店で雇用された新規従業員に着目し，早期離脱者の予測を行う新しいモデルの構築を行う．

テクノロジーの進歩により，日本の飲食店では紙とペンによる勤怠管理から，システムを利用したデジタルな勤怠管理へと以降しつつある．これにより，より詳細な勤怠時系列を利用することが可能となった．勤怠時系列が持つ情報は勤務日時だけではなく，誰といつ，どこで働いたかといった各従業員の詳細な働き方や，仕事との関わり方が記録されている．このため，勤怠時系列を利用することで，従業員の早期離脱予測をこれまでよりも高い精度で予測可能であることが考えられる．

そこでこの章では，SNS におけるユーザの早期離脱予測に利用した RNN を参考にして，新規従業員の勤怠時系列と属性情報から従業員の早期離脱を予測する新たなモデルの構築を行う．日本の匿名化された飲食チェーン店の従業員データを利用し，提案するモデルの評価を行う．離脱予測で利用される代表的な手法と比較を行い，提案手法の有効性を確かめる．更に，特徴量の分析を行い，各特徴量が早期離脱者の予測に対して与える影響を明らかにする．今回の結果は，飲食チェーン店における，新規従業員の早期離脱の防止策の検討や人材データの解析を行う上で，勤怠管理時に保持すべき情報を検討する上で役に立つものと考えられる．

本研究の具体的な貢献は以下のとおりである．

1. 勤怠時系列と属性情報を利用した新規従業員の早期離脱を予測する RNN ベースのモデルを構築した．

2. 日本の飲食チェーン店のデータを用いてベースライン手法との比較を行い，提案手法の有効性を確かめた．

3. 新規従業員の早期離脱予測に対して，それぞれの特徴量が与える影響を明らかにした．

6.1 従業員の早期離脱予測に利用したデータの説明

6.1.1 日本の飲食チェーン店

本研究では日本の飲食チェーン店で働く従業員の匿名化された勤怠情報と属性情報を用いて，早期離脱者の予測と提案手法の評価を行った．2016 年 1 月 1 日から 2018 年 3 月 21 日までの期間に働いた従業員の勤怠情報と属性情報の提供をうけた．220 店舗を対象として，この期間に各店舗で働いていた従業員数は 1260 人である．このうち，617 人が期間中に新たに雇用された新規従業員である．各店舗ごとに若干の違いはあるものの，各店舗で提供されるサービスはほとんど同一のものである．対象企業ではシフト制の働き方を採用しており，各

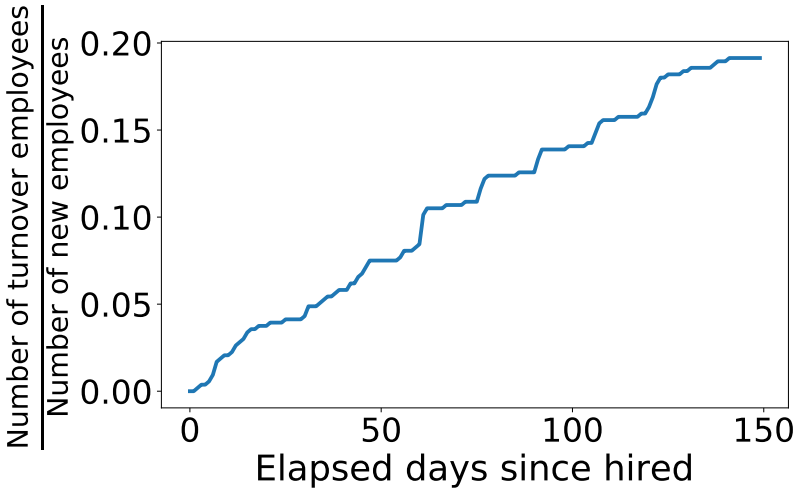


図 6.1: 日本のとある飲食チェーン店において，新たな従業員が雇用されてからの経過日数と離脱率の関係．横軸は新規従業員が初めて出勤した日からの経過日数であり，縦軸は離脱した従業員の割合である．ここでは，最後の勤怠日から 30 日間一度も勤怠がない従業員を離脱した従業員とした．

従業員の出社時間や働き方は異なる．勤怠管理システムを利用した勤怠管理が行われ，従業員は各店舗ごとに設置された端末に出勤時と退勤時に従業員カードを読み取らせることで出勤時刻と退勤時刻が記録される．また，各従業員に対して匿名化されたユニークな id が付与されており，従業員の属性情報と勤怠情報を結びつけることが可能となっている．

はじめに，今回対象とした飲食チェーン店における新規従業員の早期離脱の程度を確かめる．ここでは，最後の勤怠日から 30 日間一度も勤怠がない従業員を離脱者とした．日本では，新たに雇用された従業員がいきなり 30 日以上連続して有給を取得することは稀であり，妥当なしきい値であると考えた．図 6.1 に雇用されてからの経過日数と離脱者の割合の関係を示す．離脱率はほぼ線形に増加しており，90 日以内に約 13%の新規従業員が，150 日以内に約 19%の新規従業員が離脱している．このように，今回対象とした飲食チェーン店においても，早期離脱が大きな問題となっていることがわかる．

6.1.2 早期離脱者の定義

今回扱った早期離脱予測の説明を行う．図 6.2 に今回扱った予測の概念図を示す．従業員として初めて勤怠した日から C 日以内に離脱した従業員を早期離脱者と定義する．C は $C = [90, 150]$ と変化させた．C 日以降，30 日間一度も勤怠がない従業員を早期離脱者とし，1 度以上勤怠がある従業員を非早期離脱者とした．今回の定義で予測を行うため，はじめの勤怠日に $C + 30$ 日を足した日付が観測期間の最終日 (2018 年 3/21 日) を超えていない従業員のみを分析対象とした．初めて勤怠した日から S 日間を予測のための観測期間とし，S 日間の勤怠時系列と

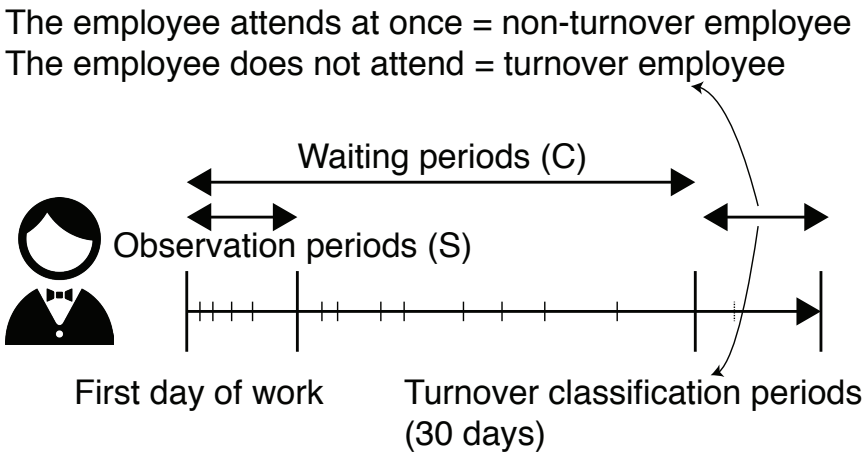


図 6.2: 飲食チェーン店における早期離脱の定義と，早期離脱予測の概念図．初めて働いた日から C 日後に 30 日以上連続して出勤がなかった従業員を早期離脱従業員とし，それ以外を非早期離脱従業員とした．また，初めて働いた日から S 日間の勤怠時系列を利用して，従業員の早期離脱を予測する．

属性情報から早期離脱者の予測を行った．S も同様に $S = [7, 28]$ と変化させた．ここでは，S 日以内に離脱している従業員に関してはそもそも予測する必要がないと考えた．S 日以降，C 日以内に一度以上勤怠情報が存在する新規従業員のみを分析対象とした．表 6.1 に今回の予測対象とした新規従業員数と，早期離脱者数をそれぞれ示す．

予測に利用する特徴量は飲食チェーン店のエリアマネージャとの議論により設計を行った．予測に利用した従業員の属性情報と勤怠時系列を表 6.2 に示す．従業員の属性情報として利用した特徴量は年齢，性別，一時的な従業員として対象企業で働いていた場合にはその雇用形態 (アルバイト，時給制の契約社員，月給制の契約社員，その他)，一時的な従業員として対象企業で働いていた場合の勤怠日数，店舗での役割 (店長，副店長，調理責任者，その他)，採用された月である．年齢と一時的な従業員として対象企業で働いていた場合の勤怠日数は連続変数であり，それ以外は対応するカテゴリのみを 1 としたワンホットのベクトルである．勤怠時系列として利用した特徴量は，勤怠開始時刻，勤怠終了時刻，勤怠時間，勤怠カテゴリ (出勤，非出勤日，欠勤，公休，有給，振替休日，忌引)，曜日である．勤怠開始時刻，勤怠終了時刻，勤怠時間は連続変数であり，それ以外は対応するカテゴリのみを 1 としたワンホットのベクトルである．出勤がない場合には勤怠開始時刻，勤怠終了時刻，勤怠時間はそれぞれ 0 とした．

表 6.1: 飲食店における早期離脱従業員と非早期離脱従業員の人数.

S (in day)	$C = 90$		$C = 150$	
	New employee	Churn employee	New employee	Churn employee
7	545	68	521	100
28	533	56	509	88

表 6.2: 従業員の早期離脱予測に利用した特徴量.

Employee demographic data	Description
Age	Continuous variable (in years)
Gender	One-hot vector
Temporary employment type	One-hot vector
Period of a temporary employee	Continuous variable (in days)
Role on a Store	One-hot vector
Employed month	One-hot vector
Employee working data	**Description**
Beginning time	Continuous variable (in minutes)
Finishing time	Continuous variable (in minutes)
Working hour	Continuous variable (in minutes)
Working type	One-hot vector
Days of the week	One-hot vector

6.2 飲食チェーン店における新規従業員の早期離脱を予測する提案モデル

ここでは，提案モデルの説明を行う．提案モデルでは，新規従業員の勤怠時系列と属性情報を用いて新規従業員の早期離脱の予測を行う．勤怠時系列は RNN によりエンコードし，属性情報と組み合わせることで新規従業員の早期離脱率を出力する．図 6.3 に提案モデルの概念図を乗せる．はじめに，モデルへの入力の説明を行う．次に，従業員の勤怠時系列と属性情報を組み合わせた提案モデルの説明を行う．最後に，提案モデルの学習方法について説明を行う．

6.2.1 入力

新規従業員の勤怠時系列を $\mathbf{s} = (s(1), ..., s(t), ..., s(S))$ と定義する．t は従業員として初めて勤怠を開始してからの経過日数である．$s(t)$ は t 日目の勤怠情報を表すベクトルである．$w(t)$

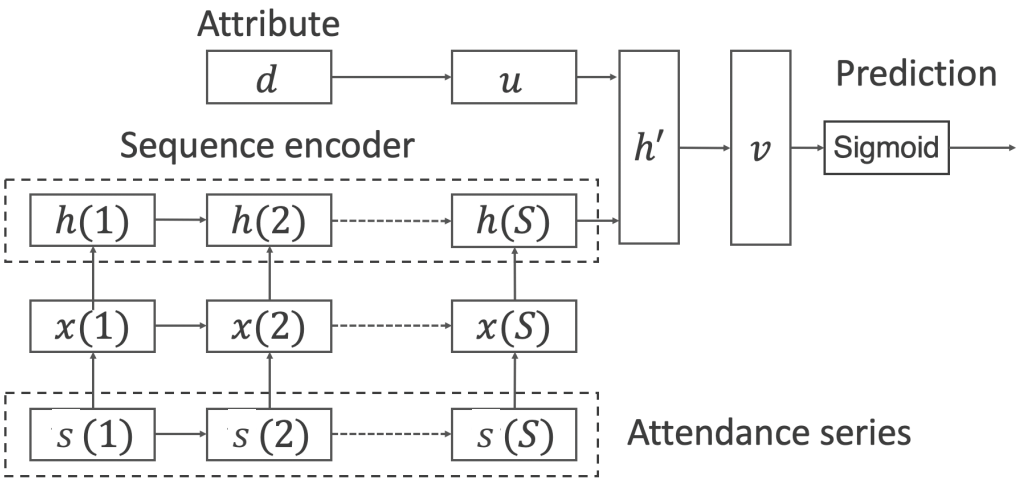


図 6.3: 飲食チェーン店において，新規従業員の勤怠時系列と属性情報から早期離脱を予測する提案モデルの概念図．提案モデルへの入力は勤怠時系列と従業員の属性情報である．勤怠時系列は 1 層のニューラルネットワークに入力された後，RNN へ入力する．RNN からの最終時刻での出力 ($h(S)$) と従業員の属性情報 (d) のベクトル表現を結合させ，h' を得る．h' を 1 層のニューラルネットワークへ入力し，その出力を sigmoid 関数への入力することで従業員の早期離脱率を出力する．

は単層のニューラルネットワークへ入力し，得られた隠れ表現 $x(t)$ をシーケンスエンコーダーへの入力 $x(t) = \mathrm{ReLU}\left(W_s s(t) + b_s\right)$ とした．d は新規従業員の属性情報であり，d についても同様に単相ニューラルネットワークへ入力し，隠れ表現 $u = \mathrm{ReLU}\left(W_d d + b_d\right)$ を得る．

6.2.2 提案モデル

提案モデルでは新規従業員の勤怠時系列を RNN，あるいは LSTM によりエンコードを行う．RNN，あるいは LSTM から得られた入力の最終時刻における出力 $h(S)$ と，従業員の属性情報を組み合わせることで早期離脱者の予測を行う．従業員の属性情報の隠れ表現である u と $h(S)$ を結合させることで，属性情報と勤怠情報を表すベクトル $h' = [h(S), u]$ を得る．h' を更に隠れ層への入力とし，得られる出力 v を出力層への入力とする $v = \mathrm{ReLU}\left(W_{h'} h' + b_{h'}\right)$．

本研究では，早期離脱者に対して $\hat{y} = 1$ のラベルを割り当て，非早期離脱者には $\hat{y} = 0$ のラベルを割り当てた．出力層にはシグモイド関数 ($sigmoid(x) = \frac{1}{1+\exp(-x)}$) を利用し，早期離脱者である確率を出力し，

$$y = sigmoid(W_v v + b_v). \tag{6.1}$$

6.2.3 モデルの学習方法

今回のデータは各ラベルの個数に大きな偏りが生じている．そのため，バイナリクロスエントロピーに対して重みを加えた

$$loss(y, \hat{y}) = (1 - w)y \log \hat{y} + w(1 - y) \log (1 - \hat{y}) \tag{6.2}$$

をロス関数とし，ラベル間の偏りに対処した．ここでは，w は学習データにおける早期離脱者の割合である．モデルの学習には back propagation throught time を利用した [49].

6.3 従業員の早期離脱予測精度の評価実験

従業員の離脱予測において利用される代表的な手法と比較を行うことで，提案手法の有効性を確かめる．以下にその詳細についてを記載する．

6.3.1 評価指標

各モデルの分類性能を評価するために，次の 3 つの指標を用いた． ROC と PR は分類しきい値を変更して計算されるため，しきい値とは無関係な値である．数値的には，0 から 1 の範囲を取り，1 に近い値ほど良好な分類パフォーマンスであることをあらわす．また，F-measure は，実際に分類しきい値を設けて予測を行った場合の評価指標である．

ROC は SNS における離脱予測においても利用した指標であり，分類しきい値を変更しながら，偽陽性率 (FPR) の関数として真陽性率 (TPR) を描画することにより得られる値である．TPR と FPR はそれぞれ，$TP/(TP + FN)$，$FP/(TN + FP)$ である．TP は早期離脱を正しく識別出来た人数，TN は早期離脱者を誤って識別した人数，FP は非早期離脱者を正しく識別出来た人数，FN は非早期離脱者を誤って識別した人数である．PR は ROC とほとんど同じやり方で計算されるが，横軸が Recall (TPR)，縦軸が Precision ($TP/(TP + FP)$ であるという点で異なる．C の値が小さい場合にはラベルに大きな偏りが生じており，各ラベルの偏りが大きい場合には PR による評価が ROC よりも適した評価指標であるといわれている [13]. F-measure は Precision と Recall (TPR) との調和平均である．

未学習のデータに対する当てはまりの良さを評価するために，$k = 10$-分割交差検証を行った．更に，トレーニングデータのうち 2 割をバリデーション用のデータとした．バリデーションデータはハイパーパラメータを決定する際に利用した．各モデルのハイパーパラメータを変化させながらバリデーションデータを用いて各指標を計算し．最大となったときのテストデータの結果を示す.

6.3.2 ベースラインモデル

本研究では，従業員の離脱予測に利用される 3 つの機械学習モデル (Logistic Regression (LR), Support Vector machine (SVM)，Random Forest (RF)) と 3 層のニューラルネットワーク (MN)

をベースラインモデルとした．これらのモデルに対する入力は提案モデルに対する入力と同様のものである．また，従業員の属性情報の隠れ表現を結合させず，従業員の勤怠時系列から予測を行ったものを RNN と LSTM とした．勤怠時系列と属性情報を組み合わせた提案手法を RNN+DF, LSTM+DF 呼ぶ．モデルへの入力のうち，連続値に関しては min-max normalization を行った．

各機械学習モデルの実装には Python の機械学習ライブラリである scikit-learn を利用し実装を行った[1]．各ベースライン手法も class_weight の引数に balanced を与え，ラベルの偏りへ対処した．また，バリデーションデータに対してグリッドサーチを行いハイパーパラメータの決定を行った．

深層学習モデルの実装は python の深層学習 API である keras を用いて行った[2]．モデルの学習には Adam を利用した．バッチサイズは [32,16] とし，バリデーションデータのロスを観測し，減少しなくなったところで学習を終了させた．各層の重みの次元数は [10,20] とした．各層への入力前にはドロップアウト層を設け，ドロップアウト率は [0.2,0.3] とした．これらのハイパーパラメータはグリッドサーチを行い決定した．それ以外のパラメータが与える影響は低いと考え，それ以外のパラメータについては keras のデフォルトの設定を利用した．

6.3.3 結果

10-分割交差検証を行ったときの，各評価指標の平均を表 6.3 に示す．最も予測性能が良かった場合には太字で表現を行った．勤怠時系列と属性情報を組み合わせて学習を行った RNN+DF あるいは LSTM+DF の ROC, PR, F-measure の値が最も高い値を示すことがわかった．また，RNN+DF と LSTM+DF の ROC, PR, F-measure の値の大小は S や C により異なることがわかった．

RNN+DF と LSTM+DF が最も高い性能を示したことから，提案手法の有効性を確かめた．一方で，勤怠時系列を LSTM でエンコードするのか，RNN でエンコードするのかといった，エンコード方法による違いは少ないことがわかった．エンコード方法に違いが現れなかった理由として考えられるのは，今回の研究では入力される時系列長 (S) が最大で 28 と短い時系列であり，時間的な長期の相関を保持する必要がなかったことがある．

6.4 各特徴量が予測に対して与える影響の分析

次に，各特徴量を除きモデルを構築した場合にテストデータのロスがどのように変化するのかを見ることで，各特徴量が早期離脱者の予測に対してどのような影響を与えるのかを明らかにする．ここでは特に，どの特徴量が予測に対する影響度が低いのかを明らかにする．すべての特徴量を学習させた場合のテストデータのロスに対する各特徴量を除いた場合のロスの変化の比率を計算した．以下，すべての特徴量を学習させた場合のテストデータのロスを

[1] https://scikit-learn.org/stable/
[2] https://keras.io/

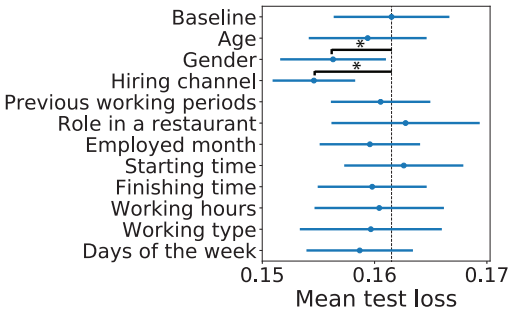
表 6.3: 飲食チェーン店で働く従業員の早期離脱を各モデルで予測した場合の予測性能の結果．最も予測性能が良かった場合には太字で表現を行った．

S (in day)	Model	$C = 90$ (in day)			$C = 150$ (in day)		
		ROC	PR	F-measure	ROC	PR	F-measure
7	LR	0.716	0.291	0.345	0.694	0.357	0.386
	SVM	0.700	0.277	0.000	0.691	0.321	0.000
	RF	0.680	0.259	0.050	0.658	0.296	0.188
	MN	0.691	0.295	0.340	0.688	0.327	0.410
	RNN	0.638	0.319	0.235	0.645	0.337	0.379
	LSTM	0.632	0.278	0.273	0.636	0.354	0.349
	RNN+DF	**0.719**	0.334	**0.352**	**0.706**	**0.375**	0.397
	LSTM+DF	0.691	**0.372**	0.305	0.698	0.371	**0.420**
28	LR	0.739	0.342	0.296	0.614	0.320	0.309
	SVM	0.694	0.290	0.029	0.665	0.336	0.040
	RF	0.633	0.306	0.114	0.647	0.277	0.126
	MN	0.684	0.289	0.300	0.666	0.355	0.354
	RNN	0.587	0.311	0.262	0.661	0.336	0.249
	LSTM	0.611	0.322	0.300	0.582	0.314	0.235
	RNN+DF	0.705	0.265	0.284	**0.722**	0.333	**0.375**
	LSTM+DF	**0.765**	**0.357**	**0.333**	0.706	**0.365**	0.350

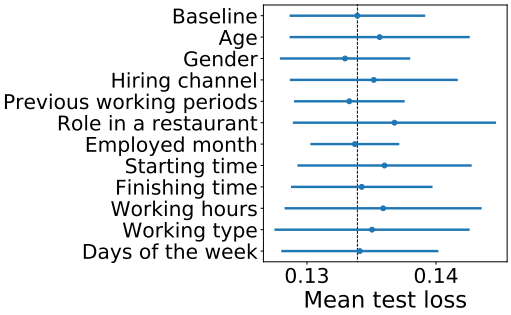
ベースラインロスと呼ぶ．ある特徴量を除いた際にロスの比率が 1 を下回る場合にその特徴量が予測に対して与える影響は低くノイズとなっていることを意味する．反対にロスの比率が 1 を超える場合にはその特徴量が予測に与える影響は大きいことを意味する．提案手法である，LSTM+DF を利用し計算を行い，ハイパーパラメータは上記の実験と同様の方法で決定し，$k = 10$-分割交差検証による平均を計算する．

図 5.7 にテストデータのロス比率の平均と分散を示す．各グラフの縦軸は取り除いた特徴量であり，横軸はその特徴量を除いたときのロスの変化である．それぞれの青い点と水平線は，それぞれ平均と標準偏差である．点線は $x = 1.0$ であり，何も特徴量をのぞかなかった場合の値である．また，$x = 1.0$ から優位な差がある場合を $*$ で表現した．$S = 7$ と $C = 90$ の性別と雇用チャネルに優位な差があることが明らかになった ($p < 0.05$)．また，$S = 7$ と $C = 150$ の性別にも有意な差があることが明らかになった ($p < 0.05$)．その他の特徴量に関しては有意な差は見られなかった．

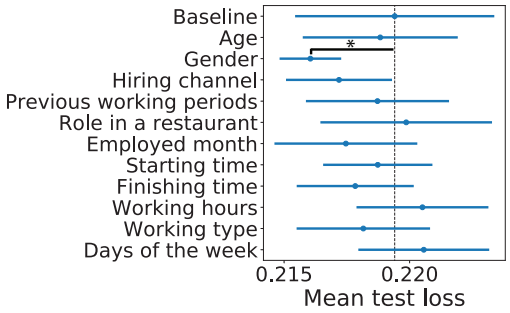
性別と雇用チャネルのテストデータのロスの平均は，有意差がない場合にも S と C のほぼすべての組み合わせで，ベースラインロスよりも低くなっている．このため，性別と雇用チャネルは、S と C の組み合わせによっては有意差こそ見られないものの，新規従業員の早期離



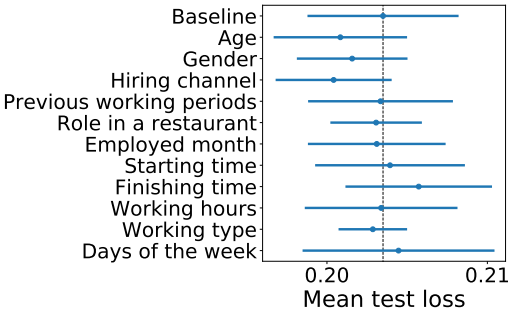
(a) $S = 7, C = 90$ の結果.



(b) $S = 28, C = 90$ の結果.



(c) $S = 7, C = 150$ の結果.



(d) $S = 28, C = 150$ の結果.

図 6.4: 飲食店で働く従業員の早期離脱を予測するモデルにおけるベースラインロスと各特徴量のみを除いた場合のロス比率．各グラフの縦軸は取り除いた特徴量であり，横軸はその特徴量を除いたときのロスの変化である．それぞれの青い点と水平線は，それぞれ平均と標準偏差である．点線は $x = 1.0$ であり，何も特徴量をのぞかなかった場合の値である．また，$x = 1.0$ から優位な差がある場合を $*$ で表現した $(p < 0.05)$.

脱にはほとんど影響を与えず，予測精度をある程度低下させる特徴量となっている．

性別が予測に対して影響を与えない理由として考えられるのは，早期離脱には性別固有の理由 (出産など) は少ないことが考えられる．また，採用チャネルが予測に対して影響を与えない理由として考えられるのは，従業員としての義務と重い責任によって早期離脱が引き起こされることが考えられる．例えば，労働時間の増加やレストランの管理といったことは，一時的な従業員として働くだけではわからない可能性がある．一方で，予測に対して有意に良い影響を与える特徴は見られなかった．

6.5 まとめと今後の課題

従業員の勤怠時系列と属性情報を利用することで新規従業員の早期離脱を予測する新たなモデルの構築を行った．匿名化された日本の飲食チェーン店の従業員データを利用してモデルの評価を行った，離脱予測において利用される代表的な手法と比較を行い，提案手法の有効

性を確かめた．更に，各特徴量が早期離脱者の予測に対してどのような影響を与えるのかを明らかにした．ここでは特に，どの特徴量が予測に対する影響度が低いのかに着目した．その結果，性別と雇用チャンネルは予測に対してほとんど影響を与えず，逆にモデルの性能に対して悪影響を及ぼすことがわかった．

今回は日本の飲食チェーン店を対象に分析を行った．飲食業以外の業種でも提案するモデルが有効であるのかを確かめる必要がある．例えば，医療業界では，飲食業と同様にシフト制の働き方を採用し，高い離脱率であることから，提案手法が有効であることが考えられる．また，勤怠管理と属性情報を用いた早期離脱予測では，最も識別性能の高かった提案手法でも F-measure が $0.35 - 0.4$ と改善の余地がある．このため，アンケート等の情報を組み合わせることで，より精度を向上させることが可能であることが考えられる．

第7章　結論

7.1　まとめ

Web の誕生により多数の情報の比較が容易となった．これにより，個々の選択肢はこれまで以上に競争的な環境にさらされている．タグや SNS，あるいは企業は競争的な環境にさらされている選択肢の例であり，これらの選択肢は様々な選択肢との比較を通して人々から選択される．このような競争的な環境においてこれらの選択肢が選択される，あるいは選択され続けるためには，複数の選択肢からの選択と，一旦選択した選択肢からの離脱という，2 つの側面から人の行為選択の詳細なメカニズムを明らかとする必要がある．

これを明らかにする上で，近年になり大規模に取得可能となった人の行動ログの分析は有益である．行動ログは，個人あるいは集団による人の行為選択の結果を反映したものである．例えば SNS では，多数のユーザがサービスに参加し投稿や他のユーザとのやりとりがサーバ上にタイムスタンプとともに保存されている．取得される行動ログは Web の世界だけに限ったものではなく，企業における人材管理も電子的に行なわれつつある．

そこで本研究では，複数の選択肢からの選択と，一旦選択したコミュニティからの離脱という，2 つの側面から人の行為選択の詳細なメカニズムを明らかとすることを目的に，タグや SNS あるいは企業の選択に関する行動ログを用いたモデル化と予測を行った．これを明らかとすることは競争的な環境においても人から選択されるための，あるいは選択され続けるための効果的な介入方法を検討する上で有益な知見となることが考えられる．今回扱った具体的な研究課題は以下の 2 つである．

- SNS におけるタグ付けデータを利用して，人の同時選択メカニズムをモデル化と分析により明らかにする．
- SNS，あるいは企業に新たに参加した人のそれぞれのコミュニティ内での振る舞いから，コミュニティからの早期離脱を予測するモデルの構築と分析により，人の早期離脱メカニズムを明らかにする．

まずはじめに，複数の選択肢がある場合に人はどの様なメカニズムで選択を行なうのかをタグ付けデータを用いることで明らかにした．タグ選択の分析では，人の選択行動をモデル化する上で用いられる Yule–Simon 過程をベースに，タグの同時利用を考慮可能なように拡張を加えた新たなモデルを提案し分析を行った．Yule–Simon 過程では，2 つのメカニズムによりタグ付けのメカニズムを記述する．新たな種類のタグは常に生み出され続ける，既存のタグの選択は優先的選択性に従うということである．このような 2 つの単純なメカニズムでタ

グの同時選択に見られる共起の偏りや，新規タグの生み出され方をどの程度再現可能であるのかを明らかにした．

共起の偏りに関しては，タグ共起の階層性を含む大部分を Yule–Simon 過程の，新たな種類のタグは生み出され続ける，既存のタグは優先的選択性に従い選択されるというメカニズムにより説明可能であることを明らかにした．この事はタグ共起の偏りを説明する上でも，優先的選択性の果たす役割は大きく，それぞれのタグが持つ意味的な繋がりが共起の偏りに対して与える影響は小さな物であるということを意味する．一方で，タグの同時利用における新規タグの生み出され方に関しては，Yule–Simon 過程の持つランダムに生み出されるという点からは外れることを確かめた．また，その選択メカニズムはユーザのタグを付ける動機により異なり，他者との情報共有を意図したサービスの場合には既存のタグが選択されることで同時に選択されるタグは増加し，自分の情報管理を意図したサービスの場合には無相関，あるいは新たな種類のタグが生み出されることで同時に選択されるタグは増加することを明らかにした．これは，他のユーザの反応に対して鈍感なユーザは独創的な選択行動を取り，他のユーザの反応に対して敏感なユーザは，他のユーザに同調するような選択行動を取りやすいということを意味する．

続く研究課題では，SNS，あるいは企業に新たに参加した人のそれぞれのコミュニティ内での振る舞いを用いて，人のコミュニティからの早期離脱を予測するモデルの構築と分析により，人の早期離脱メカニズムを明らかにした．SNS におけるユーザの早期離脱を予測する上でこれまでに提案されている再帰的ニューラルネットワークをベースに，各時刻間の予測に対する影響度の違いを明示的に考慮したモデルを新たに提案し，その有効性を代表的な機械学習手法との比較により確かめた．更に，特徴量分析を行い，SNS の場合には最初期のフォローやライクといった行為が早期離脱に対して与える影響が特に大きいことを明らかにした．更に，SNS と同様に高い早期離脱率の日本の飲食チェーン店を対象に，再帰的ニューラルネットワークを用いてモデル化を行い，採用チャンネルと性別が早期離脱に対して与える影響は少ないことを明らかにした．

本研究では，タグの選択と SNS あるいは企業からの離脱に関する行動ログを利用して，複数の選択肢からの選択と，一旦選択した選択肢からの離脱という，2 つの側面から人の行為選択のモデル化や分析を行った．今回の 2 つの研究により得られた結果は，効果が高いと考えられるユーザの獲得戦略や，それにより獲得したユーザの維持戦略を考える上で，特に重要な点について示唆を与えるものである．例えば，新規ユーザの獲得戦略として既存の参加者数が他と比べて十分に多いように見せるような工夫や，既存ユーザの維持戦略として選択直後の新規ユーザに対するフォローアップといったことである．これらに対してコストをかけることで，効果的な介入作となることが期待される．

また，本研究では複数の選択肢からの選択と，一旦選択した選択肢からの離脱という，2 種類の行為選択に対してそれぞれ新たなモデルを提案している．今回提案したモデルは，同様の種類の行動ログや，今回の分析では扱っていない行動ログを分析する際にベースラインとして利用可能である．今回提案したモデルは単純なモデルであることから，提案したモデルの拡張は容易である．このため，今回の分析に利用したものと同様の種類の行動ログに対し

てさらなる分析を行い，新たに観測された現象を考慮可能なように，今回提案したモデルをベースにした拡張や，それによる行為選択のさらなる理解が進むことが期待される．更に，提案したモデルを他の行動ログに見られる行為選択に対して適用可能な様に拡張が行われることも期待される．

7.2 今後の課題

本研究では複数の選択肢の中からの選択と一旦選択した選択肢からの離脱，2 種類の行為選択を対象に，行動ログを用いたモデル化と予測を通してそのメカニズムを明らかにした．一方で，いくつかの点で課題が残されている．

本研究では，SNS におけるタグ選択と，SNS や企業といったコミュニティ選択という，2 種類の行為選択に関する行動ログを用いて研究を行った．選択という意味では 2 つの行動ログは共通であるものの，タグ選択は気軽な選択行為であり，コミュニティ選択は気軽に行う事は難しい選択行為であるという点で異なる．このため．今後は異なる行動ログに対して同様の分析を行い，両行動ログ間の差異や類似点を明らかにする必要がある．これには複数の SNS からの選択や，複数の企業からの選択といった，一般的には公開されていない行動ログにアクセスする必要がある．将来的にこれを可能とすることで，今回明らかにしたメカニズムが気軽に行なう行為選択であるかどうかによらず，観測可能であるのかを明らかにすることが可能である．

また，本研究の最終的な目的は人の行為選択のメカニズムを明らかにし介入を行なうことにある．しかしながら，本研究では介入実験は行っておらず，介入によって実際の人の行為選択がどのように変化するのかといった点を実際に確かめる必要がある．介入実験を行う場合に考慮しなければならない点はいくつか存在する．例えば，早期離脱者を対象とした介入実験を行なう場合には，コミュニティの運営者が簡単に予測結果を利用可能なシステムを構築する必要がある．具体的には，運営者にとってわかりやすいインターフェースを準備し，予測結果を自動かつグラフィカルに利用可能である必要がある．このためには，現在の分析用のコードだけでは十分ではないことが考えられる．SNS の場合は更に，新規ユーザの数は膨大であることから，それぞれの予測には大きな計算コストがかかる．このような点で，コミュニティの運営者，ひいてはシステム設計者とのより密なやり取りを通してシステムの設計を行なう必要がある．

また，介入実験では参加者の行為選択を良くも悪くも歪めてしまう．このため，介入の結果起こりうる事象を慎重に検討し，その介入実験に参加するすべての人が介入による利益を享受出来るように実験を設計する必要がある．例えば，企業における従業員の早期離脱予測では，各従業員の早期離脱する可能性が予め分かる．離脱可能性の高い従業員に対して適切な介入を行うことで，より長くその企業ではたらいてもらえるようになることが考えられる．これは，雇用する企業にとっての利点だけでなく，従業員にとっての利点でもある．例えば，従業員が早期離脱してしまった場合には．次の職を探す上で不利な経歴となってしまう．早期離脱の予測を利用した異なる介入案として，早期離脱率に応じて人事評価を決めることが

考えられる．つまり，離脱可能性の低い従業員を早く昇進させ，離脱可能性の高い従業員の昇進を遅らせるといったものである．このような場合には，早期離脱率の高い従業員の離脱率は，介入により更に高まることが考えられ，介入により不利益を被る従業員が存在してしまう．このため，介入のガイドラインを定め，介入実験に参加するすべての従業員がそのメリットを享受出来るかどうかを慎重に検討する必要がある．

こうした点を考慮しながら，人の行為選択を明らかにするための介入実験，あるいは社会実装を今後の研究では行っていくことを検討している．

付録

Windowed Yule–Simon 過程の指数の導出

提案モデルから次数分布の指数の導出を行う．$k_i(t)$ は時刻 t でタグ i と共起したタグの数，$\langle A\rangle$ を写真が含むタグ数の期待値，M をタグの種類とすると，$k_i(t)$ の時間発展式は

$$k_i(t+\Delta t) = k_i(t)+(1-\alpha)\frac{n_i(t)\langle A\rangle}{\sum_{i=1}^{M} n_i(t)}(\langle A\rangle-1)\Delta t \tag{7.1}$$

と定義できる．今回はポアソン分布に従うと仮定しているため $\langle A\rangle$ を定義することが可能である．時刻 t におけるタグの総数は

$$\sum_{i=1}^{M} n_i(t) \quad = \quad \langle A\rangle t \tag{7.2}$$

と定義され，t が十分に大きければ

$$\tag{7.3}$$

$$k_i(t) \quad \approx \quad \langle A\rangle n_i(t) \tag{7.4}$$

と近似が可能である．以上を用いて書き換えることで

$$k_i(t+\Delta t) \quad \approx \quad k_i(t)+(1-\alpha)\frac{k_i(t)}{(\langle A\rangle-1)t}(\langle A\rangle-1)\Delta t, \tag{7.5}$$

$$\begin{aligned}
\frac{dk_i(t)}{dt} &\approx (1-\alpha)\frac{k_i(t)}{t}, \\
\int \frac{dk_i(t)}{k_i(t)} &\approx (1-\alpha)\int \frac{dt}{t}, \\
\ln k_i(t) &\approx (1-\alpha)\ln t + C_1, \\
k_i(t) &\approx C_1 t^p, \quad p = 1-\alpha.
\end{aligned} \tag{7.6}$$

また，$k_i(t_i)$ の初期値を平均値で代表することで

$$\begin{aligned}
k_i(t_i) &\approx C_1 t_i^p = \langle A\rangle - 1, \\
C_1 &\approx (\langle A\rangle - 1)/t_i^p
\end{aligned} \tag{7.7}$$

となる．これにより求めた C_1 を用いることで

$$k_i(t) \approx (\langle A\rangle - 1)(t/t_i)^p. \tag{7.8}$$

次に累積分布を求める．累積分布は

$$\begin{aligned} P(k_i(t) < k) &\approx P((\langle A\rangle - 1)(t/t_i)^p < k), &(7.9)\\ &= P\left(t_i > \left((\langle A\rangle - 1)/k\right)^{1/p}\right), \\ &= 1 - (\langle A\rangle - 1)^{1/p} k^{-1/p}. &(7.10)\end{aligned}$$

累積分布を微分することにより次数分布の指数は

$$p(k) \propto k^{-1/(1-\alpha)-1} \tag{7.11}$$

となる．

謝辞

本研究を円滑に行なうにあたって，多大なご支援を頂いた筑波大学・加藤和彦教授に深く感謝いたします．筑波大学・岡瑞起准教授には研究指導や議論等で 5 年間を通して，大変お世話になりましたこと，深く感謝いたします．筑波大学・阿部洋丈准教授には，本論文を執筆する上で多数の議論やアドバイスを頂きましたこと，深く感謝致します．本論文の副査を心良く引き受けていただき，的確なご指摘を頂いた，筑波大学・北川博之教授，同大学・山中敏正教授，同大学・善甫啓一助教に深く感謝いたします．筑波大学・鈴木健嗣教授，東京大学・池上高志教授，会津大学・橋本康弘上級准教授には研究に関する議論を多数行わせていただきました事，深く感謝いたします．また，他の研究プロジェクトにおいて，産業技術総合研究所・大槻麻衣博士，サイバーダイン株式会社・白石遼一郎博士，筑波大学・佐野祐士氏とともに，5 年間を通して行わせていただいた議論は，本研究をまとめる上で大きな助けとなりましたこと，深く感謝いたします．また，日頃より研究生活における様々な面でご協力頂いた加藤研究室の皆様，岡研究室の皆様，EMP 事務室の皆様に深く感謝いたします．最後に，5 年間あたたかく見守って頂いた，家族の皆様に深く感謝いたします．

参考文献

[1] 2018 年 日本の広告費. https://www.dentsu.co.jp/news/release/2019/0228-009767.html. [Online; accessed 3-January-2020].

[2] EAP サービス市場に関する調査結果 2014. https://www.yano.co.jp/press-release/show/press_id/1235. [Online; accessed 3-January-2020].

[3] Jaime Arguello, Brian S Butler, Elisabeth Joyce, Robert Kraut, Kimberly S. Ling, Carolyn Rosé, and Xiaoqing Wang. Talk to me: foundations for successful individual-group interactions in online communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 959–968, 2006.

[4] Albert-László Barabási. The origin of bursts and heaby tails in human dynamics. *Nature*, 435(7039):207–211, 2005.

[5] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[6] Alain Barrat, Marc Barthélemy, Romualdo Pastor-Satorras, and Alessandro Vespignani. The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(11):3747–3752, 2004.

[7] Nita Cantrell and Mort Sarabakhsh. Correlates of non-institutional food service turnover. *Hospitality Review*, 9(2):52–59, 1991.

[8] Ciro Cattuto, Andrea Baldassarri, Vito D. P. Servedio, and Vittorio Loreto. Vocabulary growth in collaborative tagging systems. *arXiv*, 2007.

[9] Ciro Cattuto, Alain Barrat, Andrea Baldassarri, Gregory Schehr, and Vittorio Loreto. Collective dynamics of social annotation. *Proceedings of the National Academy of Sciences of the United States of America*, 106(26):10511–10515, 2009.

[10] Ciro Cattuto, Vittorio Loreto, and Luciano Pietronero. Semiotic dynamics and collaborative tagging. *Proceedings of the National Academy of Sciences of the United States of America*, 104(5):1461–1464, 2007.

[11] Boreum Choi, Kira Alexander, Robert E. Kraut, and John M. Levine. Socialization tactics in wikipedia and their effects. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, pages 107–116, 2010.

[12] François Chollet et al. Keras. "`https://keras.io`", 2015.

[13] Jesse Davis and Mark Goadrich. The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine learning*, pages 233–240, 2006.

[14] Gideon Dror, Dan Pelleg, Oleg Rokhlenko, and Idan Szpektor. Churn prediction in new users of Yahoo! Answers. In *Companion Proceedings of the 21st International Conference on World Wide Web*, pages 829–834, 2012.

[15] Samira Ebrahimi K., Vincent Michalski, Kishore Konda, Roland Memisevic, and Christopher Pal. Recurrent neural networks for emotion recognition in video. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 467–474, 2015.

[16] Evan Tarver. This Is Why Instagram Is Winning Over Flickr. `http://www.investopedia.com/articles/markets/082015/why-instagram-winning-over-flickr.asp`, 2015. [Online; accessed 17-May-2016].

[17] Emilio Ferrara, Roberto Interdonato, and Andrea Tagarelli. Online popularity and topical interests through the lens of Instagram. In *Proceedings of the 25th ACM Conference on Hypertext and Social Media*, pages 24–34, 2014.

[18] The Japan Institute for Labour Policy and Training. Job openings-to-applicants ratio. https://www.jil.go.jp/english/estatis/eshuyo/e0208.html, 2013. [Online; accessed 12-March-2019].

[19] Jill Freyne, Michal Jacovi, Ido Guy, and Werner Geyer. Increasing engagement through early recommender intervention. In *Proceedings of the Third ACM Conference on Recommender Systems*, pages 85–92, 2009.

[20] Felix A. Gers, Jürgen Schmidhuber, and Fred Cummins. Learning to forget: Continual prediction with LSTM. *Neural Computation*, 12(10):2451–2471, 2000.

[21] Hyo G. Geun, Richard W. Redman, and Marjorie C. McCullagh. Predictors of turnover among asian foreign-educated nurses in their 1st year of US employment. *The Journal of Nursing Administration*, 48(10):519–525, 2018.

[22] Scott A. Golder and Bernardo A. Huberman. Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32(2):198–208, 2006.

[23] Olaf Görlitz, Sergej Sizov, and Steffen Staab. PINTS: peer-to-peer infrastructure for tagging systems. In *Proceedings of the 7th International Conference on Peer-to-peer Systems*, 2008.

[24] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6645–6649, 2013.

[25] Alastair M. Gray and Victoria L. Phillips. Turnover, age and length of service: a comparison of nurses and other staff in the National Health Service. *Journal of Advanced Nursing*, 19(4):819–827, 1994.

[26] Lei Guo, Enhua Tan, Songqing Chen, Xiaodong Zhang, and Yihong E. Zhao. Analyzing patterns of user content generation in online social networks. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 369–378, 2009.

[27] Manish Gupta, Rui Li, Zhijun Yin, and Jiawei Han. An overview of social tagging and applications. In *Social Network Data Analytics*, pages 447–497. 2011.

[28] Harry Halpin, Valentin Robu, and Hana Shepherd. The complex dynamics of collaborative tagging. In *Proceedings of the 16th International Conference on World Wide Web*, pages 211–220, 2007.

[29] Tony Hammond, Timo Hannay, Ben Lund, and Joanna Scott. Social bookmarking tools (I): a general review. *D-Lib Magazine*, 11(4):1082–9873, 2005.

[30] Harold S. Heaps. *Information Retrieval: Computational and Theoretical Aspects*. Academic Press, Inc., 1978.

[31] Markus Heckner, Michael Heilemann, and Christian Wolff. Personal information management vs. resource sharing: towards a model of information behaviour in social tagging systems. In *Proceedings of the Third International AAAI Conference on Weblogs and Social Media*, pages 42–49, 2009.

[32] Markus Heckner, Tanja Neubauer, and Christian Wolff. Tree, funny, to_read, google: what are tags supposed to achieve? a comparative analysis of user keywords for different digital resource types. In *Proceedings of the 2008 ACM Workshop on Search in Social Media*, pages 3–10, 2008.

[33] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[34] Raghav Pavan Karumur, Tien T. Nguyen, and Joseph A. Konstan. Early activity diversity: assessing newcomer retention from first-session activity. In *Proceedings of the 19th ACM*

Conference on Computer-Supported Cooperative Work & Social Computing, pages 595–608, 2016.

[35] Stuart A. Kauffman. Investigations: the nature of autonomous agents and the worlds they mutually create. Santa Fe Institute, 1996.

[36] Christian Körner, Roman Kern, Hans-Peter Grahsl, and Markus Strohmaier. Of categorizers and describers: an evaluation of quantitative measures for tagging motivation. In *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia*, pages 157–166, 2010.

[37] Nan Lin, Daifeng Li, Ying Ding, Bing He, Zheng Qin, Jie Tang, Juanzi Li, and Tianxi Dong. The dynamic features of Delicious, Flickr, and YouTube. *Journal of the American Society for Information Science and Technology*, 63(1):139–162, 2012.

[38] James M. McFillen, Carl D. Riegel, and Cathy A. Enz. Why restaurant managers quit (and how to keep them). *Cornell Hotel and Restaurant Administration Quarterly*, 27(3):37–43, 1986.

[39] Larry Medsker and Lakhmi C. Jain. *Recurrent Neural Networks: design and Applications*. CRC Press, 1999.

[40] Mark E. J. Newman. Clustering and preferential attachment in growing networks. *Physical review E*, 64, 2001.

[41] Mark E. J. Newman. The structure and function of complex networks. *Society for Industrial and Applied Mathematics Review*, 45(2):167–256, 2003.

[42] Aditya Pal, Shuo Chang, and Joseph A. Konstan. Evolution of experts in question answering communities. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, pages 274–281. AAAI, 2012.

[43] Romualdo Pastor-Satorras, Alexei Vázquez, and Alessandro Vespignani. Dynamical and correlation properties of the internet. *Physical Review Letters*, 87(25), 2001.

[44] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[45] Jagat Sastry Pudipeddi, Leman Akoglu, and Hanghang Tong. User churn in focused question answering sites: Characterizations and prediction. In *Companion Proceedings of the 23rd International Conference on World Wide Web*, pages 469–474, 2014.

[46] Erzsébet Ravasz and Albert-László Barabási. Hierarchical organization in complex networks. *Physical Review E*, 67(2), 2003.

[47] Erzsébet Ravasz, Anna L. Somera, Dale A. Mongru, Zoltán N. Oltvai, and Albert-László Barabási. Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551–1555, 2002.

[48] Frederick F. Reichheld and W. Earl Sasser. Zero defections: quality comes to services. *Harvard business Review*, 68(5):105–111, 1990.

[49] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

[50] Matthew J. Salganik. *Bit by Bit: Social Research in the Digital Age*. Princeton University Press, 2017.

[51] V. Vijaya Saradhi and Girish Keshav Palshikar. Employee churn prediction. *Expert Systems with Applications*, 38(3):1999–2006, 2011.

[52] Mike Schuster and Kuldip K. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

[53] Börkur Sigurbjörnsson and Roelof Van Z. Flickr tag recommendation based on collective knowledge. In *Proceedings of the, 17th International Conference on World Wide Web*, pages 327–336, 2008.

[54] Mikhail V. Simkin and Vwani P. Roychowdhury. Re-inventing willis. *Physics Reports*, 502(1):1–35, 2011.

[55] Herbert A. Simon. On a class of skew distribution functions. *Biometrika*, 42(3/4):425–440, 1955.

[56] Markus Strohmaier, Christian Körner, and Roman Kern. Understanding why users tag: a survey of tagging motivation literature and results from an empirical study. *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, 17:1–11, 2012.

[57] Stefan Strohmeier and Franca Piazza. Domain driven data mining in human resource management: a review of current research. *Expert Systems with Applications*, 40(7):2410–2420, 2013.

[58] The Japan Ministry of Health, Labour and Welfare. Job introduction situation according to an occupation. `https://www.mhlw.go.jp/content/11602000/G38-3101.pdf`, 2019. [Online; accessed 24-March-2019].

[59] Peg Thoms, Paula Wolper, Kimberly S. Scott, and Dave Jones. The relationship between immediate turnover and employee theft in the restaurant industry. *Journal of Business and Psychology*, 15(4):561–577, 2001.

[60] Gergely Tibély, Péter Pollner, Tamás Vicsek, and Gergely Palla. Extracting tag hierarchies. *PloS ONE*, 8(12), 2013.

[61] Jennifer Trant. Studying social tagging and folksonomy: a review and framework. *Journal of Digital Information*, 10(1), 2009.

[62] Francesca Tria, Vittorio Loreto, Vito D. P. Servedio, and Steven H. Strogatz. The dynamics of correlated novelties. *Scientific Reports*, 4, 2014.

[63] T. Vander Wal. Explaining and showing broad and narrow folksonomies. http://www.personalinfocloud.com/blog/2005/2/21/explaining-and-showing-broad-and-narrow-folksonomies.html?rq=broad, 2005. [Online; accessed 10-December-2015].

[64] John R. Wilke. Retailing: Supercomputers manage holiday stock. Wall Street Journal, 1992.

[65] John C. Willis. *Age and Area: A study in Geographical Distribution and Origin of Species*. Cambridge University Press, 1922.

[66] Robert H. Woods and James F. Macaulay. R for turnover: Retention programs that work. *Cornell Hotel and Restaurant Administration Quarterly*, 30(1):78–90, 1989.

[67] Qiang Yan, Lianren Wu, and Lan Zheng. Social network based microblog user behavior analysis. *Physica A: Statistical Mechanics and Its Applications*, 392(7):1712–1723, 2013.

[68] Carl Yang, Xiaolin Shi, Luo Jie, and Jiawei Han. I know you'll be back: Interpretable new user clustering and churn prediction on a mobile social application. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 914–922, 2018.

[69] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 14th Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, 2016.

[70] George U. Yule. A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F.R.S. *Philosophical Transactions of the Royal Society B*, pages 21–87, 1925.

[71] George K. Zipf. *The Psycho-biology of Language: An Introduction to Dynamic Philology*. International Library of Psychology, 1935.

公表論文リスト

博士論文と直接関係のある論文

学術雑誌論文

- 佐藤晃矢, 岡瑞起, 橋本康弘, 加藤和彦, Yule–Simon 過程によるタグ共起ダイナミクスのモデル化と分析. 人工知能学会論文誌, 30(5):667–674, 2015.

査読付き国際学術会議論文

- Koya Sato, Mizuki Oka, Yasuhiro Hashimoto, Takashi Ikegami, Kazuhiko Kato. How the nature of web services drives vocabulary creation in social tagging. In Proceedings of the 2nd International Conference on Information Science and System, pages 17–21, 2019.
- Koya Sato, Mizuki Oka, Kazuhiko Kato. Early churn user classification in social networking service using attention-based long short-term memory. In Proceedings of the 14th Pacific Asia Workshop on Intelligence and Security Informatics (PAKDD Workshop), pages 45–56, 2019.
- Koya Sato, Mizuki Oka, Kazuhiko Kato. Early turnover prediction of new restaurant employees from their attendance records and attributes. In Proceedings of the 30th Database and Expert Systems Applications, pages 277–286, 2019.

博士論文と直接関係しない論文

学術雑誌論文

- 佐野祐士, 佐藤晃矢, 白石僚一郎, 大槻麻衣, 球技における視触覚刺激提示がプレイスキルに及ぼす影響. 日本バーチャルリアリティ学会論文誌, 22(4):493–502, 2017.

査読付き国際学術会議論文

口頭発表

- Koya Sato. Design and implementation of the augmented volleyball court. In Companion Proceedings of the 2018 ACM International Conference on Interactive Surfaces and Spaces, pages 19–24, 2018.
- Koya Sato, Yuji Sano, Mai Otsuki, Mizuki Oka, Kazuhiko Kato. Augmented recreational volleyball court. In Proceedings of the 10th Augmented Human International Conference, 2019.
- Yuji Sano, Koya Sato, Ryoichiro Shiraishi, Mai Otsuki, Koichi Mizutan. Visuo-haptic interface to augment player's perception in multiplayer ball game. In Proceedings of the International Conference on Artificial Reality and Telexistence & Eurographics Symposium on Virtual Environments, 2019.

ポスター&デモ

- Ryoichiro Shiraishi, Koya Sato, Yuji Sano, Mai Otsuki. Haptic directional instruction system for sports. In Proceedings of the Asia Haptics 2016, 2016.
- Yuji Sano, Koya Sato, Ryoichiro Shiraishi, Mai Otsuki. Sports support system: augmented ball game for filling gap between player. In Proceedings of the 2016 ACM International Conference on Interactive Surfaces and Spaces, 2016.
- Koya Sato, Jun Izawa. An influence of motor costs in human reinforcement learning. In Proceedings of the Society for NeuroScience 2016 Annual meeting, 2016.
- Mizuki Oka, Koya Sato, Yasuhiro Hashimoto, Takashi Ikegami. Emergence of individuality on social tagging dynamics. In Proceedings of the Complex Networks 2016, 2016.
- Yuji Sano, Koya Sato, Ryoichiro Shiraishi, Mai Otsuki. Player perception augmentation for beginners using visual and haptic feedback in ball game. In Proceedings of the IEEE Conference on Virtual Reality 2019, 2019.