

ブログサービスにおける初期の閲覧行動による新規読者定着評価

西條 直哉^{†1} 大竹 恒平^{†2} 生田目 崇^{†3}

概要: 本研究では、ブログサービスの顧客行動について、特に新規登録顧客がその後どのようにブログを利用するかに着目した。そのため、新規読者の定着評価を行うための分析モデルを提示し、実データによる評価を行った。特に本論文では、ブログサービス利用のきっかけとなった流入クラスと初期の登録行動をクラス分けし、さらに読者と登録や「いいね」の付与などの行動を説明変数とした定着の判別分析を行っている。また、分析の結果から新たなブログの推薦を行いその評価も行った。

キーワード: ブログサービス, 初期閲覧行動, ロジスティック回帰分析, 確率的潜在意味解析

1. はじめに

スマートフォンやタブレット型端末の普及はインターネット上のコミュニケーションを加速させ、特にソーシャルメディアの利用が伸びている。総務省の調査によると、いずれかのソーシャルメディアを利用している人は 80.6% であり、多くの人に利用されていることがわかる [1]。また、ソーシャルメディアの利用状況として、「ほとんど情報発信や発言をせず、他人の書き込みや発言等の閲覧しか行わない」と回答する利用者の割合が、書き込みなどを行う利用者よりも多いという傾向がある [2]。したがって、ソーシャルメディアを運営する企業にとって、情報発信者だけでなく、閲覧者を増加させること、すなわち新規の閲覧ユーザの獲得及びその後の定着化はサービスを発展させるために必要不可欠である。

また、企業にとってユーザの新規獲得コストは大きいため、新規獲得したユーザの離脱を防ぐことは非常に重要である [3]。近年、様々なサービスにおいて、ユーザの継続・離脱を予測する研究 [4-9] が行われているが、新規ユーザに焦点を当てた研究は十分に行われていない。新規顧客については行動データが十分に蓄積されていないことが一因と考えられる。

一方、ユーザの離脱を防ぐための施策の一つに推薦システムがある。推薦システムとは、利用者にとって有用と思われる情報などを選び出し、それを利用者の目的に合わせた形で提示するシステムのことで、協調フィルタリングが最も一般的なアルゴリズムとして挙げられる。しかし、協調フィルタリングには、コールドスタート問題と言われる、新規ユーザには行動履歴がないため推薦できないという問題がある。

本研究では、ブログサービスを対象に、初期行動を考慮した新規ユーザの継続・離脱を予測することを試みる。初

期行動には、新規ユーザが「なぜ会員登録したか」、「何に興味を持っているのか」といった会員登録をしたきっかけが反映されていることが期待できる。会員登録をしたきっかけが異なれば、その後の行動、さらには読者として定着する要因が異なると考えられる。したがって、新規ユーザに対して初期行動を用いて類型化し、セグメント毎に新規ユーザの継続・離脱を予測する。各セグメントのモデルを比較することにより、新規ユーザが読者として定着する要因の違いを明らかにすることを目指す。同時に、新規ユーザが読者として定着しやすくなるようなブログの推薦を行う。

2. 関連研究

本節では、特にソーシャルメディアでの利用者の行動分析やサービスの利用継続に関して、本論文に関連した先行研究について紹介し、本論文の位置づけについてまとめる。

根来ら [4] は、SNS のユーザ行動のログデータを用いて、ユーザのアクティブ度にどのような特性・要因が作用しているかについて統計的に検証し、コミュニケーションの活発性の観点から、SNS における友人数はある程度の最適な人数規模が存在していることを示唆した。古川ら [5] は、ブログネットワークにおいて、ユーザがどのようなページを頻繁に閲覧しているのかについてブログ間の関係性に着目して分析している。特に、ブログネットワークでは、ユーザがブックマークを中心にブログネットワーク内を巡回していることを明らかにした。

和田ら [6] は、インターネットテレビを対象にユーザの視聴行動からサービスの継続・離脱分析を週単位で行うと同時に、ユーザのアクティブ度を定量化する方法を提案している。分析では、翌週アクティブ日数を目的変数として、週単位での継続・離脱を予測し、ユーザの離脱につながるような行動を示唆した。また、[7] では、スマートフォン

^{†1} (株) ブレインパッド

^{†2} 東海大学

^{†3} 中央大学 (連絡先: nama@indsys.chuo-u.ac.jp)

投稿日: 2020 年 12 月 25 日

採録日: 2021 年 3 月 12 日

ームを対象としたユーザの離脱・継続行動分析も行っている。上門ら [8] は、転職サイトを対象に職務経歴書のデータから特徴量を抽出し、会員の離脱要因の抽出及び離脱予測を決定木分析により行っている。

また、宮崎ら [9] は、RNN (Recurrent Neural Network) を用いたユーザの離脱予測モデルを提案し、モデル有用性を示唆した。近年ディープラーニングに代表される複雑なモデルが注目されているが、本研究においては、読者として定着するユーザの特徴抽出を試みることを目的としているため、こうしたモデルによるアプローチは有用ではないと考える。

これらの研究に対して、本研究ではブログサービスを対象に、新規ユーザの継続・離脱を予測することを試みる。既存研究においては、コールドスタートに近い、新規ユーザに焦点を当てた研究は十分に行われていない。しかし、獲得した新規顧客の継続顧客化は重要な経営課題であり、本研究では、新規ユーザの限られた行動データから初期行動に着目する。先に述べた通り、初期行動には、新規ユーザが「なぜ会員登録したか」、「何に興味を持っているのか」といった会員登録をしたきっかけが反映されていることが期待できる。会員登録をしたきっかけが異なれば、その後の行動、さらには読者として定着する要因が異なると考えられる。そこで、新規ユーザを初期行動から類型化し、セグメント毎に新規ユーザの継続・離脱を予測する。また、各セグメントについて、継続利用者として定着する要因の違いを明らかにすることを目指す。

さらに、分析結果から、新規ユーザが読者として定着しやすくなるような施策の提案を目指す。具体的には、ブログ推薦のための方法論を分析の枠組みから提案し、その効果を確かめる。上述のコールドスタートに近い新規ユーザに対して適切な推薦ができれば、顧客維持に寄与できるものとする。

3. 対象データ

本研究では、大手ブログサービスを運営する企業から提供されたデータを用いて分析を行う。ブログは代表的なソーシャルメディアの一つであり、特徴として、SNS などと比較して1つの投稿で発信できる情報量が多いこと、過去に発信された投稿がカテゴリや更新年月などによって整理された形でサイト上に蓄積されるため、過去に発信された投稿を探しやすいということが挙げられる。また、ブログ毎に投稿されるトピックが定まっているものが多く、様々なトピックのブログが存在するため、ユーザがどのブログを閲覧するかは個人の嗜好が反映されやすい。したがって、ブログの閲覧行動を分析することで各ユーザのトピックに対する興味の度合いや変化を明らかにすることができると考えられる。

対象データ期間は2018年1月1日から6月1日であり、

ユーザがいつどのブログを見たかという「ブログ閲覧データ」、各ブログがいつどのくらい閲覧されているかという「ブログ被閲覧データ」、ユーザのデモグラフィック属性やブログの所属ジャンルといった「会員情報データ」の3つである。データ項目の概要を表1にまとめる。

表1 データ概要

データ名	主なカラム
ブログ閲覧データ	日付, デバイス, ユーザ ID, ブロガーID, 記事閲覧数, コメント数, リブログ数, 「いいね」数, 読者登録申請数
ブログ被閲覧データ	日付, デバイス, ユーザ ID, ブロガーID, 被記事閲覧数, 被コメント数, 被リブログ数, 被「いいね」数, 被読者登録申請数
会員情報データ	日付, ユーザ ID, 性別, 年齢区分, ブロガーカテゴリ, 所属ジャンル, 会員登録

また、データ全体のユーザ数は131,863人であった。本研究では、会員登録日の条件(5月4日まで)、pv(ページビュー)の条件(会員登録日以外の閲覧あり)、流入ジャンル(データに含まれるジャンルが大変広範囲であるため、流入ユーザ数が200名以上という条件を設け、この条件を満たさないジャンルから流入したユーザは除く)の条件、を満たす新規ユーザを分析対象として選定した。その結果、対象ユーザ数は72,848人となった。

流入ジャンルについては、本ブログサイトにおいては、様々なジャンルのブログが混在している。そのため、読者の興味のあるジャンルによって閲覧するブログは大きく異なると考えられる。新規会員登録した読者が、会員登録時にどのようなジャンルに興味を持って会員登録したかを流入ジャンルとして定義する。そして、本研究においては、流入時点での興味が明確と考えられるユーザを対象としたいため、流入ジャンルが複数考えられるユーザを除いて分析を行う。この結果、流入ジャンルは「ママ・パパ」、「インテリア・DIY」、「スポーツ」、「仕事術」など77カテゴリとなる。

以上により抽出されたデータを5節以降の分析で用いる。

4. 分析のフレームワーク

分析にあたり、新規読者が定着する要因について以下の2つの視点でその差異があると考えこれを実証する。

- ジャンルによる定着の差異

● 初期行動による定着の差異

一つは、どのようなジャンルのブログに興味を持って会員登録をしたかにより定着する要因が異なるという仮説である。もう一つは、会員登録初期にどのようなブログサービス内行動をしているかにより定着する要因が異なるという仮説である。本研究では、上記の2つの仮説に基づいて分析を行う。まず、新規ユーザのセグメンテーションを行う。上記の仮説に基づき、会員登録時の興味ジャンル及び会員登録初期の行動の二軸により新規ユーザをセグメント分けする。次に、セグメント毎に新規ユーザの読者としての定着を予測するモデルを構築する。構築した各モデルより、各セグメントにおける定着要因の抽出及び、新規読者の定着に寄与するブログの推薦を行う。

5. 新規ユーザのセグメンテーション

ブログの新規ユーザの閲覧行動を評価するために、「流入ジャンル」と「初期行動」の二軸で新規ユーザのセグメンテーションを行う。

5.1 ブログジャンルの類型化

本研究では、新規ユーザを会員登録時の興味ジャンルで類型化するために、興味ジャンルを確率的潜在意味解析 (pLSA, probabilistic latent semantic analysis) を用いて類型化する。

ユーザを $X = \{x_1, x_2, \dots, x_I\}$, ジャンルを $Y = \{y_1, y_2, \dots, y_J\}$, 潜在クラスを $Z = \{z_1, z_2, \dots, z_K\}$ とする。潜在クラス z_k に属すると仮定したときの x_i が起こる条件付き確率を $P(x_i|z_k)$ とし、潜在クラス z_k に属すると仮定したときの y_j が起こる条件付き確率を $P(y_j|z_k)$ とする。潜在クラス z_k の出現確率 $P(z_k)$ を重みとして加算することにより x_i と y_j の共起確率 $P(x_i, y_j)$ は以下のようになる。

$$P(x_i, y_j) = \sum_{k \in K} P(z_k) P(x_i|z_k) P(y_j|z_k) \quad (1)$$

このような潜在クラスを持つモデルにおいては、以下の対数尤度 LL を最大化するように EM アルゴリズムにより $P(z_k)$, $P(x_i|z_k)$, $P(y_j|z_k)$ を推定する。

$$LL = \sum_{x \in X} \sum_{y \in Y} n(x_i, y_j) \log P(x_i, y_j) \quad (2)$$

既存のユーザの閲覧行動によりジャンルをクラスターリングする。そのために、既存顧客 29 万人のユーザを対象として、2018年4月1日～4月30日の各ジャンルの pv 数を用いてジャンルのクラスターリングを行う。本研究では BIC 規準によりクラス数は 6 とした。

そして、各新規ユーザは流入ジャンルによって各クラスに所属するとする。このクラスを流入クラスと定義する。各流入クラスの名称と新規ユーザの人数を表 2 に示す。

表 2 各流入クラスの所属新規ユーザ数

流入クラス	流入クラス名	所属ユーザ数
1	ライフスタイル	6,670
2	美容・韓国	5,947
3	有名人	31,098
4	趣味	9,080
5	育児	6,475
6	追及	13,578

5.2 初期行動によるユーザの類型化

本研究において、会員登録から 3 日間をユーザの初期行動と定義した。初期行動により、新規ユーザがどのような目的及び関心を持って会員登録をしたかを類型化することができると考えた。はじめに、初期行動でブログ記事を執筆しているか否かで分類する。分析対象ユーザ 72,848 人うち、およそ 20% に当たる 14,822 人が初期行動でブログ記事を投稿していた。これらのユーザは、自身のブログ記事を投稿することを 1 つの目的として会員登録をしていると考えられるため、ブログ投稿クラスとして分類する。

次に、初期行動でブログを執筆していないユーザ、すなわちブログを読むことを主な目的として会員登録したユーザの分類を考える。読者として会員登録したユーザについては、初期行動でのブログの閲覧を以下の二軸で考える。

- ブログ閲覧に対する関心の高さ
- ブログ閲覧に対する興味の広さ

ブログ閲覧に対する関心の高さについては、ブログ記事の閲覧数と「いいね」数、コメント数、読者登録数を統合した指標で定義する。初期行動におけるユーザ i の pv 数を Pv_i 、「いいね」の数を $Like_i$ 、コメント数を $Comment_i$ 、読者登録数を $Check_i$ とする。また、記事の閲覧をしているユーザ数を A_{pv} 、対象ユーザのうち「いいね」をしているユーザ数を A_{like} 、コメントをしているユーザ数を $A_{comment}$ 、読者登録をしているユーザ数を A_{check} とする。このとき、ユーザ i のブログ閲覧に対する関心の高さ $Interest_i$ を以下の式で表す。

$$Interest_i = Pv_i + w_{like} \times Like_i + w_{comment} \times Comment_i + w_{check} \times Check_i \quad (3)$$

ただし、それぞれのウェイトは、以下の式のように、pv 数をそれぞれの指標の値で割った、指標 1 カウント当たりの pv 数としている。

$$w_{\{like, comment, check\}} = \frac{A_{pv}}{A_{\{like, comment, check\}}} \quad (4)$$

この結果 $w_{like} = 3.76$, $w_{comment} = 12.57$, $w_{check} = 1.59$ となった。

ブログ閲覧に対する興味の幅広さについては、対象の77ジャンルのうち実際に閲覧したジャンル数を指標として用いる。初期行動でのブログの執筆の有無で分け、ブログ投稿がないクラスについては、上記で説明した関心の高さと興味の幅広さについてそれぞれの中央値 (32.7 と 3) を閾値として2つずつに分類する。これらの二軸でのクラスを掛け合わせることで、読者として登録したユーザを4つに類型化する。

これらのクラスを初期行動クラスとする (図2)。各初期行動クラスの所属人数を表3に示す。

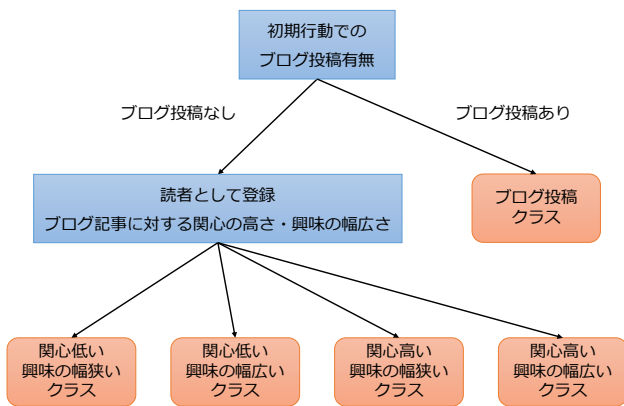


図2 初期行動クラスの分割

表3 各流入クラスの所属ユーザ数

初期行動クラス		初期行動クラス	所属ユーザ数
読者	関心低い/興味の幅狭い	low_narrow	22,150
	関心低い/興味の幅広い	low_wide	6,870
	関心高い/興味の幅狭い	high_narrow	10,792
	関心高い/興味の幅広い	high_wide	18,214
ブログ投稿		post_article	14,822

そして、上記で類型化した流入クラス (表2) と、初期行動クラス (表3) を掛け合わせた30セグメントを新規ユーザのセグメントとする。

6. 読者の定着を予測するモデル

新規ユーザのセグメント毎に新規ユーザが読者としての定着を予測するモデルを構築する。

はじめに、読者の定着について定義する。新規ユーザの会員登録日から14日目までのpv数の合計を pv_1 、15日

目から28日目までのpv数の合計を pv_2 とする。このとき、以下の式を満たしている場合、その新規ユーザは定着していると定義する。

$$\frac{pv_2}{pv_1} \geq 0.5 \quad (5)$$

閾値の0.5は、データ提供元との議論で閲覧行動が時間と共に落ち着くという事実から設定している。流入クラス (6クラス) × 初期行動クラス (読者4 + ブログ投稿1) = 30セグメントにおいて式(5)を満たして定着と判断されるユーザの割合は最小で35.8%、最大で59.5%であった。

6.1 ロジスティック回帰分析による定着の分析

本研究では、各ユーザが定着するか否かに関する要因分析を行うために、ロジスティック回帰モデルに L_1 ノルムによるペナルティを考慮したLassoを用いた分析を行う。 $\beta = (\beta_1, \beta_2, \dots, \beta_p)^T$, $x = (x_1, x_2, \dots, x_p)^T$ とする。教師データが0か1の二値であるような場合、以下の対数尤度関数を最大化する最尤法によりパラメータを推定する。

$$LL(\beta_0, \beta) = \sum_{i=1}^n (\beta_0 + \beta^T x_i) - \sum_{i=1}^n \log\{1 + \exp\{\beta_0 + \beta^T x_i\}\} \quad (6)$$

そして、Lassoでは負の対数尤度関数にパラメータ値の絶対値に関する罰則項を付与した正則化を行い、パラメータを推定する。

$$\hat{\beta}_0, \hat{\beta} = \underset{\beta_0, \beta}{\operatorname{argmin}} \left\{ -LL(\beta_0, \beta) + \lambda \sum_{j=1}^p |\beta_j| \right\} \quad (7)$$

ここで、 λ はパラメータの縮小量を制御するための複雑度変数である。

分析にあたっては、図2に示すように、4月15日までの会員登録ユーザを学習データ、その後のデータをテストデータとして分割している。また、サンプルサイズが大きいため、学習にあたってはクラスの割合をアンダーサンプリングで合わせたうえで、ブートストラップ法を用いてパラメータを推定する。なお、サンプリング回数は500回で5-foldクロスバリデーションによる検証を行っている。Accuracyが最大となるモデルを最適モデルとして、テストデータに当てはめて精度を評価する。モデルの作成に用いた説明変数候補を表4にまとめる。

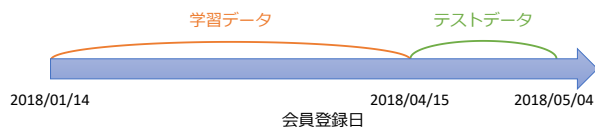


図2 データ期間の分割

表4 説明変数のデータセット

属性		変数名	概要	データ型
目的変数		target	定着の有無（定着=1）	0-1
説明変数	行動	blog_qty	閲覧ブログ数	int
		post_comment	コメント数	int
		count_like	いいね数	int
		count_checklist	読者登録申請数	int
		multi_category_flg	複数カテゴリ（official, general）閲覧フラグ	boolean
		post_article_flg	ブログ執筆フラグ（ブログ執筆クラス以外）	boolean
	ブログ人気度	—	2018年4月の被pv数によるブログの人気度 その人気度のブログの閲覧フラグ ・人気度5～人気度1	5 level boolean
	ジャンル	—	既存ユーザの閲覧が多いジャンルの閲覧フラグ 2018/01/07-2018/01/14の1週間における既存ユーザの閲覧率25%以上 40ジャンル ・オフィシャルジャンル 11ジャンル ・ジェネラルジャンル 29ジャンル	40 level boolean
	トップブログ	—	既存ユーザの閲覧が多いブログの閲覧フラグ 2018/01/07-2018/01/14の1週間における既存ユーザの閲覧率20%以上 38ブログ ・オフィシャルブログ 32ブログ ・ジェネラルブログ 6ブログ	38 level boolean

ブログ人気度については、pv数から5段階（1万未満、30万未満、300万未満、1億未満、それ以上）としてラベリングしている。

また、ジャンルは既存ユーザの閲覧が多いジャンルの閲覧フラグである。ある1週間における既存ユーザの閲覧率25%以上の40ジャンルを変数として採用した。同様に、トップブログは既存ユーザの閲覧が多いブログの閲覧フラグである。ある1週間における既存ユーザの閲覧率20%以上の40ジャンルを変数として採用した。多くのユーザから支持されているブログやジャンルは読者の定着に影響を与えていると考え、これらをモデルの変数とした。

以上の条件で、4節で類型化した新規ユーザのセグメント毎に、新規ユーザの読者としての定着を予測するモデル

を計30モデルを構築する。

6.2 予測の精度と考察

以上のモデルセッティングにおいて、30モデルについての予測結果を混同行列から評価する。30モデルの平均評価値は、精度である Accuracy が 0.625、適合率である Precision が 0.544、再現率である Recall が 0.648、Precision と Recall の調和平均である F 値が 0.590 となった。Precision が低い傾向にあるが、これは、本モデルに使用している説明変数が上側に裾の広いデータであることが一因として考えられる。また、流入クラス別、および、初期行動クラス別に平均を求めた各指標値を表5,6にまとめる。

これらの結果から、流入クラスによる Accuracy の差は

あまりない。しかし、初期行動クラスによる Accuracy の差があった。初期行動で関心が高いユーザは比較的予測しやすいが、初期行動で関心が低く、興味の幅が狭いユーザは、予測がしづらいことがわかる。

表5 流入クラス別の評価値

集計単位	Accuracy	Precision	Recall	F-measure
1 ライフスタイル	0.611	0.524	0.635	0.572
2 韓国・美容	0.627	0.542	0.654	0.590
3 有名人	0.639	0.553	0.666	0.603
4 趣味	0.621	0.531	0.647	0.581
5 育児	0.627	0.558	0.647	0.597
6 自己研磨	0.622	0.558	0.640	0.595

表6 初期行動クラス別の評価値

集計単位		Accuracy	Precision	Recall	F-measure
読者	関心低い/興味の幅狭い	0.582	0.472	0.608	0.530
	関心低い/興味の幅広い	0.615	0.520	0.641	0.572
	関心高い/興味の幅狭い	0.650	0.572	0.677	0.619
	関心高い/興味の幅広い	0.646	0.612	0.656	0.633
ブログ執筆		0.631	0.546	0.658	0.596

6.3 読者定着要因の分析

本研究では、500回のブートストラップ法を用いてパラメータを推定しているため、データセットのサンプリングによって出現する変数が異なる。本研究では500回のうち0でないパラメータ値が推定された回数を変数重要度として扱う。変数重要度が高いものほど定着に影響を与えている変数である。また、各パラメータの推定値は、0と推定されたパラメータを除いた平均値とする。

行動変数のパラメータの推定値を表7に示す。なお、初期行動クラス間で比較しやすいよう、初期行動クラス毎にまとめた。パラメータの推定値については、変数重要度が300以上のもので、その中でも変数重要度が400以上の変数のパラメータの推定値は下線で表す。

表7より、多くのセグメントにおいて **blog_qty** や **count_like** が定着に正の影響を与えていることがわかる。初期行動クラス間で比較するといくつかのことがわかる。

まず、初期行動クラスが **high_narrow**、すなわち関心が高く、興味の幅が狭いクラスにおいては、他の初期行動クラ

スと比較して **blog_qty** が定着に強く正の影響を与えている。初期のブログに対する関心が高いため、多くのブログを閲覧することで定着しやすくなると推察できる。どのようなブログを閲覧すべきかについては、ジャンル変数やトップブログ変数と合わせて見る必要がある。

次に、初期行動クラスが **low_wide**、すなわち関心が低く、興味の幅が広いクラスにおいては、多くの流入クラスで **count_checklist** が定着に正の影響を与えている。このクラスの新規ユーザは、初期行動において、幅広いジャンルのブログを少しずつ閲覧しているユーザである。このようなユーザは会員登録初期において、読みたいブログが特に定まっておらず、さまざまなジャンルを閲覧することで、自分の興味があるブログを探している可能性がある。したがって、興味を持つブログを見つけ読者登録することで、定着的に閲覧するブログを作ることが定着につながると推察できる。

そして、初期行動クラスが **post_article**、すなわちブログを投稿しているクラスにおいては、多くの流入クラスで **post_comment** が定着に正の影響を与えている。このクラスの新規ユーザは、ブログを閲覧するだけでなく、自らブログ記事を投稿するために会員登録したユーザである。自らもブロガーである新規ユーザは、他のブロガーの記事にコメントを投稿することで、他ブロガーとのつながることが読者として定着しやすくなると推察できる。

最後に、初期行動クラスが **high_wide**、すなわち関心が高く、興味の幅が広いクラス及び、初期行動クラスが **low_narrow**、すなわち関心が低く、興味の幅が狭いクラスにおいては、多くの流入クラスで **count_like** が定着に正の影響を与えている。これらのクラスのユーザは、共に自分が良いと思うブログを見つられた場合において定着しやすくなっている。したがって、初期行動クラス **high_wide** では、初期行動時にとにかく様々なブログを閲覧したが、良いと思うブログがみつけれなかったといったユーザが存在すると推察できる。また、初期行動クラス **low_narrow** についても、初期行動時にすぐに良いと思うブログがみつけれなかったため、閲覧をあまりせずに離脱してしまっているユーザが存在すると推察できる。

表7 初期行動クラス毎にまとめた行動変数のパラメータの推定値

初期行動クラス	流入クラス	行動変数					
		blog_qty	post_comment	count_like	count_checklist	multi_category_flg	post_article_flg
low_narrow	1	0.114		<u>0.107</u>			
	2	<u>0.190</u>		<u>0.138</u>			
	3	<u>0.229</u>	<u>0.049</u>	<u>0.215</u>			
	4			<u>0.226</u>	<u>0.110</u>		
	5						
	6			0.129	<u>0.167</u>		
low_wide	1	0.17			<u>0.156</u>	<u>0.689</u>	
	2				<u>0.201</u>		
	3	0.157	<u>-0.095</u>	<u>0.247</u>	0.062		
	4		0.186				
	5	<u>0.314</u>	0.192	0.167	<u>0.275</u>		
	6	<u>0.268</u>	0.28	<u>0.335</u>	<u>0.109</u>		
high_narrow	1	<u>0.411</u>					-1.272
	2	<u>0.399</u>		<u>0.294</u>			
	3	<u>0.261</u>	<u>0.346</u>	<u>0.297</u>	-0.049		-0.222
	4	<u>0.682</u>	0.078	<u>0.140</u>	<u>0.212</u>		
	5	<u>0.516</u>		0.215	0.163		
	6	<u>0.423</u>					
high_wide	1		0.053				
	2		0.092	0.196			0.285
	3	<u>0.194</u>	0.024	0.266	-0.113	<u>0.266</u>	<u>-0.301</u>
	4	<u>0.16</u>		0.139			
	5	<u>0.193</u>	0.125	0.169	0.224		0.231
	6	<u>0.169</u>		0.236		<u>0.321</u>	
post_article	1	<u>0.323</u>	<u>0.134</u>				
	2						
	3		<u>0.093</u>	<u>0.209</u>	-0.083		
	4	<u>0.223</u>		0.101	<u>0.176</u>	0.161	
	5	<u>0.262</u>	0.079		<u>0.117</u>		
	6		<u>0.133</u>	<u>0.133</u>	0.077		

次に、ブログ人気度変数のパラメータの推定値を表8に示す。ここで、数値は人気の高低を表しており、5は人気が高く、逆に1は人气が低い。なお、初期行動クラス間で比較しやすいよう、初期行動クラス毎にまとめた。パラメータの推定値の表記は上記と同様である。

表8より、多くのセグメントにおいて人気度が高いブログの閲覧は定着に正の影響を与え、人気度が低いブログは定着に負の影響を与えていることがわかる。また、初期行動クラスが post_article、すなわちブログを執筆しているクラスにおいては、人気度1や人気度2でも定着に正の影響を与えていることがわかる。自分と同じような一般ブロガーの記事を閲覧することによって、他のブロガーとつなが

ったり、自身のブログを執筆する際に参考にしたりすることが定着しやすくなる要因の一つであると推察できる。

そして、ジャンル変数については、流入クラスによる依存が強いため、流入クラス毎に各初期行動クラスの5モデルのうちいくつかのモデルが定着に影響を与えているのかを集計した。変数重要度が300以上でパラメータの推定値が正であるものを定着に影響を与えている変数とする。表9に流入クラス毎のジャンル変数が定着に影響を与えているモデル数を示す。なお、定着に影響を与えているモデル数の合計が多い上位15変数を示し、各ジャンル変数のジャンルが含まれる流入クラスの値については下線で表す。

表8 ブログ人気度変数のパラメータの推定値

流入 クラス	初期 行動 クラス	ブログ人気度変数				
		人気度5	人気度4	人気度3	人気度2	人気度1
low_narrow	1				<u>0.296</u>	<u>-0.253</u>
	2			0.219		—
	3	0.081	<u>0.187</u>		<u>-0.209</u>	<u>-0.123</u>
	4			<u>0.262</u>		<u>-0.206</u>
	5					-0.228
	6	<u>0.238</u>	<u>0.195</u>	0.096		<u>-0.154</u>
low_wide	1					-0.285
	2					
	3					<u>-0.152</u>
	4					
	5	0.391				-0.316
	6			<u>0.290</u>		-0.199
high_narrow	1					
	2		<u>0.724</u>			
	3	<u>0.311</u>	<u>0.329</u>	-0.057	<u>-0.254</u>	<u>-0.249</u>
	4		<u>0.339</u>	0.239		<u>-0.314</u>
	5				0.286	-0.358
	6		0.183	<u>0.344</u>		<u>-0.344</u>
high_wide	1					<u>-0.226</u>
	2	<u>0.338</u>		0.416		
	3	-0.140	0.211	0.055	<u>-0.214</u>	<u>-0.212</u>
	4	<u>0.190</u>				
	5			<u>1.082</u>		<u>-0.207</u>
	6	0.164	<u>0.347</u>	0.245	-0.179	<u>-0.293</u>
post_article	1	0.242		<u>0.475</u>		
	2				0.24	
	3		<u>0.259</u>			
	4		0.235	0.244	<u>0.270</u>	0.128
	5	0.183				
	6			<u>0.209</u>	<u>0.427</u>	-0.134

表8より、プロフェッショナルが多くの子グメントにおいて、定着に影響を与えていることがわかる。また、アラフィフやインテリア・DIYのようにジャンルが含まれる流入クラスにおいては、全く定着に影響を与えないが、他の流入クラスにおいては、定着に影響を与えるジャンルもあることがわかる。

表9 流入クラス毎のジャンル変数の影響 (上位15変数)

ジャンル変数	流入クラス						合計
	1	2	3	4	5	6	
プロフェッショナル	2	3	5	1	3	5	19
アラフィフ	2	3	4	0	2	5	16
アラフォー	1	2	4	3	1	3	14
30代・ファッション	1	2	4	1	2	3	13
ディズニーレポ	2	1	3	1	2	3	12
インテリア・DIY	0	0	3	1	3	4	11
整理整頓・ミニマルライフ	2	2	3	2	0	2	11
クリエイター	3	1	1	1	1	4	11
子育て(小学生以上)	3	1	3	0	2	1	10
スポーツ	1	0	3	2	1	2	9
役者	2	0	2	4	0	1	9
アラサー	0	1	1	3	1	2	8
タレント	1	1	2	3	1	0	8
芸人	2	0	4	0	0	2	8
音楽・ミュージシャン	2	0	3	1	0	2	8

トップブログ変数についても、ジャンル変数と同様、流入クラスによる依存が強いいため、流入クラス毎に各初期行動クラスの5モデルのうちいくつかのモデルが定着に影響を与えているのかを集計した。変数重要度が300以上でパラメータの推定値が正であるものを定着に影響を与えている変数とする。表10に流入クラス毎のトップブログ変数が定着に影響を与えているモデル数を示す。なお、定着に影響を与えているモデル数の合計が多い上位13変数を示し、各トップブログ変数のブログが所属するジャンルが含まれる流入クラスの値については下線で表す。

表10 流入クラス毎のトップブログ変数の影響 (上位)

トップブログ変数	流入クラス						合計
	1	2	3	4	5	6	
インテリア・DIY_01	2	2	3	2	1	1	11
役者_01	1	2	3	1	1	3	11
タレント_06	0	0	5	1	1	2	9
タレント_05	0	2	1	1	1	3	8
プロフェッショナル_01	2	0	2	2	0	1	7
ファッション_04	1	0	2	3	0	1	7
ママ・パパ_07	1	1	3	0	1	1	7
アイドル_02	2	0	2	1	1	0	6
タレント_01	1	1	1	1	2	0	6
プロフェッショナル_02	1	0	1	0	2	2	6
ファッション_01	1	1	3	0	0	1	6
ママ・パパ_02	1	0	3	0	1	1	6
ママ・パパ_15	1	0	2	1	2	0	6

表10より、インテリア・DIY_01や役者_01が比較的多くのセグメントにおいて、定着に影響を与えていることがわかる。しかし、ジャンルほど多くのセグメントで定着に影響を与えている変数はない。したがって、ジャンルよりもトップブログのほうが、セグメント間での定着に影響を与えている要因が異なることがわかる。

7. ブログの推薦

前節の結果より、多くのセグメントにおいて、新規読者の定着する要因の一つとして、閲覧ブログ数が多いということが挙げられた。また、各セグメントにおいて定着に寄与しやすいブログやジャンルがあることが明らかになった。そこで、本研究では、構築した各モデルを用いて、新規ユーザが読者として定着することを目指したブログの推薦を行う。本研究では、推薦数を対象のブログサービスのレイアウトなどを考慮して、トップブログ3つ、ジャンル3つの計6つとする。各セグメントのモデルにおいて、パラメータの推定値が正のものを変数重要度が大きい順にトップブログ3つ、ジャンル3つを推薦する。また、このときの推薦が定着に寄与しているかについて、テストデータを用いて検証する。なお、セグメントにより定着しているユーザ数の比率が異なるため、テストデータをアンダーサンプリングして検証を行う。

ただし、本研究では実際にレコメンドをすることができないため、レコメンドリストにあるジャンルを実際にテスト期間において閲覧しているかどうかを測定した。

テストデータにおいて、各セグメントの定着ユーザ数を R 、定着ユーザのうち推薦されたブログのいずれかを閲覧しているユーザ数を R_{view} 、非定着ユーザ数を A 、非定着ユーザのうち推薦されたブログのいずれかを閲覧しているユーザ数を A_{view} とする。このときの推薦が定着に寄与しているかを表す評価値を式(8)で定義する。

$$Evaluation = \frac{R_{view}}{R} - \frac{A_{view}}{A} \quad (8)$$

評価値 $Evaluation$ は、推薦したブログを閲覧しているユーザ比率が定着したユーザと定着しなかったユーザでどの程度差があるかを表す指標である。評価値が大きいほど、定着したユーザは閲覧しているが、定着していないユーザは閲覧していないブログを推薦できていることになる。本研究では、比較対象として、各セグメントにおいて学習データにおける閲覧ユーザ数が多い順にトップブログ3つ、ジャンル3つを推薦し、モデルで推薦した場合と評価値を比較する。表11に各セグメントにモデル及び閲覧人数を用いてブログを推薦した際の評価値をまとめる。数値は式(8)の $Evaluation$ の値であり、大きい方が望ましい。最後の列は、本研究のモデルで推薦した際の評価値が閲覧人数を基に推

薦した際の評価値よりも優れているものに○をつけている。

表11 推薦の評価

流入 クラ ス	初期行動クラス		推薦方法		評 価
			モデル	閲覧人 数	
1	読 者	low_narrow	0.077	0.046	○
		low_wide	0.090	0.141	
		high_narrow	0.103	0.172	
		high_wide	0.106	0.030	○
	post_article		0.205	0.239	
2	読 者	low_narrow	0.033	-0.042	○
		low_wide	-0.143	-0.071	
		high_narrow	0.228	0.158	○
		high_wide	0.104	0.094	○
	post_article		0.320	0.175	○
3	読 者	low_narrow	0.046	0.040	○
		low_wide	0.187	0.023	○
		high_narrow	0.191	0.082	○
		high_wide	0.210	0.036	○
	post_article		0.160	0.083	○
4	読 者	low_narrow	0.005	0.021	
		low_wide	0.208	0.146	○
		high_narrow	0.168	0.129	○
		high_wide	0.150	0.118	○
	post_article		0.123	0.170	
5	読 者	low_narrow	0.202	0.128	○
		low_wide	0.190	0.024	○
		high_narrow	0.000	0.079	
		high_wide	0.092	0.053	○
	post_article		0.129	0.106	○
6	読 者	low_narrow	0.146	0.161	
		low_wide	0.128	0.045	○
		high_narrow	0.059	0.307	
		high_wide	0.108	0.079	○
	post_article		0.207	0.146	○
平均			0.128	0.097	○

8. 結果と考察

8.1 ジャンルによる定着の差異

流入クラスを軸として、各セグメントにおけるモデルのパラメータの推定値を比較すると、行動変数やジャンル変数やトップブログ変数において、定着に影響を与える要因が異なることが明らかになった。大きな特徴としては、流入クラス3とその他の流入クラスとの違いである。

流入クラス3では、ブログ記事に「いいね」をすること

やコメントを投稿することが他の流入クラスと比較して強く定着に影響を与えている。有名人ブロガーによるブログには、一般的に知名度があるため、ブログを閲覧する以前のファンが存在することが考えられる。したがって、「いいね」やコメントを投稿することで、有名人ブロガーを応援しているようなユーザは定着しやすくなっていると推察できる。

また、読者登録をすることが定着に負の影響を与えている。読者登録はブログのブックマーク的機能であることから、知っている有名人のブログなのでとりあえず読者登録したが、その後は閲覧しなくなってしまうユーザが少なからず存在すると推察できる。

以上より、ジャンルによる定着の差異があるといえる。

8.2 初期行動による定着の差異

初期行動クラスを軸として、各セグメントにおけるモデルのパラメータの推定値を比較すると、行動変数やブログ人気度変数において、定着に影響を与える要因が異なることが明らかになった。表12に初期行動クラス毎の定着に影響を与える主な要因を示す。

表12 初期行動クラス毎の定着に影響を与える主な要因

初期行動クラス	定着に影響を与える主な要因
low_narrow	ブログ記事に対する「いいね」
low_wide	読者登録
high_narrow	閲覧ブログ数の増加
	閲覧ジャンルの拡大
high_wide	ブログ記事に対する「いいね」
	閲覧ブログ数の増加
post_article	ブログ記事に対するコメントの投稿
	自分と同様な一般ブログの閲覧

表12より、読者として登録し、ブログの閲覧に対して関心が低く、興味の幅が広いクラスでは、読者登録をすることが定着する要因の一つとなっている。このクラスのユーザは、会員登録の時点でさほど関心が高くなく、読みたいブログが定まっていないため、多様なジャンルのブログを探索することによって、読みたいブログを探しているユーザが多いと考えられる。したがって、探索したブログの中から読者登録して読みたいと思えるブログを発見することが定着につながると推察できる。

また、ブログを執筆しているクラスでは、ブログ記事にコメントを投稿することや自分と同様な一般ブログの閲覧が定着する要因の一つとなっている。自らもブログを書きたいと思い会員登録をしたユーザであるため、他の一般ブロガーの記事を参考に閲覧したり、ブログ記事にコメントを投稿することによって、他のブロガーとつながること

読者としても定着しやすくなると推察できる。

以上より、初期行動による定着の差異があるといえる。

8.3 推薦の評価

表11より、一部のセグメントにおいて、モデルで推薦した際の評価値が閲覧人数で推薦した際の評価値を上回らないケースもあったが、全体の平均評価値はモデルを用いて推薦した際の方が高くなっている。したがって、モデルを用いて推薦することで定着に寄与するようなブログの推薦ができていているといえる。

9. まとめと今後の課題

本研究では、ブログサービスを対象として、新規ユーザが読者として定着する要因を明らかにすることを試みた。

「興味ジャンル」及び「初期行動」の二軸でユーザを類型化し、セグメント毎に新規ユーザの定着を予測する判別モデルを構築することで定着要因が異なることを明らかにした。また、モデルを利用することで、新規ユーザの定着に寄与するブログの推薦をすることができた。

本研究では、ユーザの閲覧行動をブログ単位で集計しているため、ユーザの閲覧行動がどのブログを閲覧したかまでとなっており、どの記事を読んだかまでは判別できていない。同一のブロガーの記事であっても、記事によりトピックや記事の長さなどが異なると考えられる。ブログ記事をテキストマイニングするなどにより、どのような記事を読んでいるかを明らかにすることで、さらなる定着につながる要因を明らかにすることが期待できる。定着率の閾値についても本研究では経験上の知見から設定したが、さらに合理的な値の設定については議論の余地が残されている。

また、本研究では、ユーザの閲覧行動を日単位で集計しているため、ユーザがどのような順序でブログサービス内を遷移し、閲覧したかを考慮できていない。ブログサービス内の回遊行動は、ユーザの興味の強さや移り変わりを反映していると考えられる。したがって、このようなブログ記事の閲覧順序についても検討する必要がある。ブログのレコメンドについて、本研究ではテスト期間の閲覧データを正解としたが、レコメンド効果を正確に検証するためには、実際にレコメンドの実験が必要である。今回は様々な制約でできなかったが今後機会があればレコメンド実験を実施したい。

さらに、本研究のモデルにおける予測精度は決して高いものとは言い切れない。こうした問題においては、変数の吟味もさることながら、分析手法においても機械学習やニューラルネットワークなど他の手法を利用することも考えられる。これらは今後の課題としたい。

謝辞

本論文は、本研究は中央大学特定課題研究の研究成果の一部であり、同研究費による助成を受けたものである。

参考文献

- [1] 総務省情報流通行政局: 「ICT によるインクルージョンの実現に関する調査研究報告書」, <http://www.soumu.go.jp/johotsusintokei/> (最終閲覧日: 2020 年 12 月 23 日)
- [2] 和田計也, 福田一郎: “インターネットテレビにおけるユーザの視聴行動分析,” 日本マーケティング学会カンファレンス・プロシーディングス, Vol. 5, pp. 61-65 (2016).
- [3] Verbeke, W., Dejaeger, K., Martens, D., Hur, J. and Baesens, B.: “New insights into churn prediction in the telecommunication sector: A profit driven data mining approach,” *European Journal of Operational Research*, Vol. 218, No. 1, pp.211-229 (2012).
- [4] 根来龍之, 大寄昌子, 木村俊也: “SNS におけるアクティブユーザーの特性 mixi のユーザー行動データによる研究,” 2012 年経営情報学会春季全国研究発表大会要旨集, pp. 9-12, (2012).
- [5] 古川忠延, 松澤智史, 松尾豊, 内山幸樹, 武田正之: “Weblog におけるユーザのつながりと閲覧行動の分析,” 電子情報通信学会論文誌, Vol. J88B, No. 7, pp. 1258-1266 (2012).
- [6] 和田計也, 福田一郎: “インターネットテレビにおけるユーザの視聴行動分析,” 日本マーケティング学会カンファレンス・プロシーディングス, Vol. 5, pp. 61-65 (2016).
- [7] 和田計也, 高野雅典, 福田一郎: “スマートフォンゲームにおけるユーザの離脱・継続行動分析,” エンタテインメントコンビューティングシンポジウム 2014 論文集, pp. 304-306 (2014).
- [8] 上門雄也, 大和田勇人, 金盛克俊: “テキストマイニングを用いた転職サイトの会員離脱予測,” 第 31 回人工知能学会全国大会論文集, Vol. 31, pp. 1-4 (2017).
- [9] 宮崎邦洋, 村山菜月, 山本裕樹, 牛山史明, 大澤昇平, 松尾豊: “深層学習を用いたユーザ離脱予測,” 第 32 回人工知能学会全国大会論文集, Vol. 32, pp. 1-3 (2018).

Evaluation of retention using new customer browsing behavior of blog service

Naoya Saijo^{†1} Kohei Otake^{†2} Takashi Namatame^{†3}

Abstract: In this study, we focused on the customer behavior on blog services, especially how newly registered customers use blogs afterwards. Therefore, we showed an analytical model for evaluating the retention of new customers and evaluated it using an actual data. Especially, in this paper, we classified some classes that triggered the use of the blog service and the initial browsing behavior, and further performed logistic regressions analyses to discriminate whether customer continue to use or not using behavior variables such as registered genre and “like”. Then, we also recommended new blogs based on the results of the analysis and evaluated them.

Keywords: Blog service, initial browsing behavior, logistic regression, probabilistic latent semantic analysis

^{†1} BrainPad Inc.

^{†2} Tokai University

^{†3} Chuo University (Correspondence Author: nama@indsys.chuo-u.ac.jp)