ブログサービスにおける初期の閲覧行動を考慮した新規読者の定着に関する研究

経営システム工学専攻　西條　直哉

1　研究背景と目的

近年，スマートフォンやタブレット型端末の普及に伴い，様々なソーシャルメディアが利用されている．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 [1]．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．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 [2, 3] 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，新規ユーザを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　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いただいたデータを用いて分析を行う．対象データ期間は 2018/01/01~2018/06/01 であり，具体的には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．

また，対象ユーザを会員登録日，閲覧デバイス，閲覧日数，pv 数，流入ジャンルの 5 つを条件として選定した．その結果，対象ユーザ数は 72848 人となった．

3　分析の流れ

はじめに，新規読者が定着する要因について以下の 2 つの仮説を立てた．

- 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．

4　新規ユーザのセグメンテーション

「興味ジャンル」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．

4.1　興味ジャンルによるユーザの類型化

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，ジャンルを確率的潜在意味解析を用いて類型化する．

ユーザを $X = \{x_1, x_2, ..., x_I\}$，ジャンルを $Y = \{y_1, y_2, ..., y_J\}$，潜在クラスを $Z = \{z_1, z_2, ..., z_K\}$ 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，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．

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

このモデルに対し，以下の対数尤度 L が最大となるような $P(z_k)$，$P(x_i|z_k)$，$P(y_j|z_k)$ を EM アルゴリズムを用いて推定する．

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

本研究では BIC 規準によりクラス数を 6 とした．各ユーザの流入ジャンルを会員登録日の pv 数が最大のジャンルとしたとき，各ユーザは流入ジャンルの所属確率が最大のクラスに所属するとする．また，これらのクラスを流入クラスと定義する．各流入クラスの所属人数を表 1 に示す．

表 1: 各流入クラスの所属ユーザ数

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	6670
2	美容・韓国	5947
3	有名人	31098
4	趣味	9080
5	育児	6475
6	追求	13578

4.2　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 72848 人うち，およそ 20%に当たる 14822 人が初期行動でブログ記事を執筆していた．これらのユーザは，ブログ記事を執筆することを 1 つの目的として会員登録をしているため，ブログ執筆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．

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

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

ブログ閲覧に対する興味の広さについては，閲覧ジャンル数を指標として用いる．各指標は各中央値を閾値として 2 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．各初期行動クラスの所属人数を表 2 に示す．

表 2: 各初期行動クラスの所属ユーザ数

初期行動クラス		初期行動クラス名	所属ユーザ数
読者	関心低い/興味の幅狭い	low_narrow	22150
	関心低い/興味の幅広い	low_wide	6870
	関心高い/興味の幅狭い	high_narrow	10792
	関心高い/興味の幅広い	high_wide	18214
ブログ執筆		post_article	14822

上記で類型化した，流入クラスと初期行動クラスを掛け合わせた 30 セグメントを新規ユーザのセグメントとする．

5 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，以下の式を満たす場合，その新規ユーザは定着していると定義する．

$$\frac{pv_2}{pv_1} \geq 0.5 \tag{5}$$

本研究では，ロジスティック回帰モデルに L_1 ノルムを加えた Lasso を用いる．$\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)^{\mathrm{T}}$，$\boldsymbol{x} = (x_1, x_2, ..., x_p)^{\mathrm{T}}$ とする．ロジスティック回帰モデルでは，パラメータ $\boldsymbol{\beta}$，β_0 は以下の対数尤度関数を最大化する最尤法によって推定する．

$$\begin{aligned} l(\beta_0, \boldsymbol{\beta}) = &\sum_{i=1}^{n} y_i(\beta_0 + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x_i}) \\ &- \sum_{i=1}^{n} \log\{1 + \exp\{\beta_0 + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}\}\} \end{aligned} \tag{6}$$

Lasso では，負の対数尤度関数に罰則項を付与した正則化法を用いてモデルを推定する．

$$\hat{\beta}_0, \hat{\boldsymbol{\beta}} = \underset{\beta_0, \boldsymbol{\beta}}{\mathrm{argmin}}\{-l(\beta_0, \boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j|\} \tag{7}$$

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．

本研究ではサンプルサイズが大きくないため，ブートストラップ法を用いてパラメータを推定する．その際，各変数の推定されたパラメータが 0 でない回数を変数重要度として扱う．また，サンプリング回数を 500 回とする．

説明変数には，閲覧ブログ数やいいね数などの行動変数，閲覧ブログの規模を表すブログ人気度変数，ジャンルの閲覧フラグを表すジャンル変数及びブログの閲覧フラグを表すトップブログ変数の 4 つを用いる．

6 ブログの推薦

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から変数重要度が大きいブログを推薦する． 各セグメントにおいて閲覧人数が多いブログを推薦した場合を比較手法として，このときの推薦が定着に寄与しているかをテストデータを用いて検証する．

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 $Evaluation$ を以下の式で定義する．

$$Evaluation = \frac{R_{view}}{R} - \frac{A_{view}}{A} \tag{8}$$

7 結果と考察

7.1 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から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も，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．

以上により，ジャンルによる定着の差異があるといえる．

7.2 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 3 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 3: 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	ブログ記事対するコメントの投稿 自分と同様な一般ブログの閲覧

まず，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を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で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．

7.3 推薦の評価

提案手法と比較手法でブログを推薦した際の評価指標の平均値を表 4 に示す．

表 4: ブログを推薦した際の評価指標の平均値

	提案手法	比較手法
評価指標	0.128	0.097

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るといえる．

8 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．

今後の課題としては，記事単位まで落とした閲覧行動の分析やブログサービス内の回遊行動の考慮などが考えられる．

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/linkdata/h30_03_houkoku.pdf (2018).

[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] 上門 雄也, 大和田 勇人, 金盛 克俊, 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).