The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

メディアサービスにおけるユーザの継続の冪分布に基づくモデル化

Modelling of retention of media service users based on power-law

森下 壮一郎 *1
Soichiro Morishita

*1株式会社サイバーエージェント
CyberAgent, Inc.

In this paper, a model of the visitor retention in media services via the Internet such as video distribution service is shown. Visitor intervals of media services usually follow power-law. Based on the theory, we describe the modeling of visitor intervals and separate the residual into the elements of return and retention. Moreover, the result of the separation according to access log of the Internet TV station “AbemaTV” is shown as an example.

1. はじめに

本稿では，ウェブ上のメディアサービスにおけるアクティブユーザの継続数の統計的モデルについて述べる．

企業の価値を計るための概念として customer equity が提唱されている [1]．これは，将来にわたっての収益性の担保という観点の下で，優良な顧客を抱えている企業は資産を持つ企業と同様に価値があるとみなすものである．

顧客生涯価値（LTV: lifetime value）は customer equity を計算するための考え方の一つで，顧客 1 人あたりに対する売上の将来にわたっての積算と定義される．定義上は顧客の一生涯が対象であるが，実際には目的に応じて視野に入れる期間を設定して，その区間で積算する．具体的には，1 人の顧客が時間経過に応じて一定の確率で離脱すると考えて，対象期間において顧客が継続している確率とユーザ 1 人あたりの平均売上の積とする．

なお顧客数の集計方法は業態に応じて大きく異なる．たとえば動画配信サービスなどのウェブを介したサービスでは，ユーザに対して識別子（ID: identifier）を付与して，期間中にアクセスがあったユーザ（アクティブユーザ）を集計するのが一般的である．ユーザ登録を伴わないサービスの場合はウェブブラウザや専用アプリケーションなどのユーザエージェント（UA: user agent）に対して ID を付与して同様の集計を行う．この ID は自然人と一対一対応しないのであるが，便宜的に ID の単位をユーザとし，このように集計した ID の数をアクティブユーザ数と呼ぶ．しかしながら UA に付与した ID は端末の変更や UA の再インストールなどにより変わるので，素朴な集計では多めの値になることを見積もる必要がある．一方，継続率の計算においては，実際にはユーザが継続して利用していても UA に付与した ID は変わってしまいがちであるので，見かけの継続数は下がってしまう．

以上の理由で継続数の見積もりは困難であるが，一般にユーザを獲得した日と経過日数とのクロス集計による分析が行われており，素朴な集計よりは妥当な結果を得られる．このような分析は，継続数についての性質が同等になるような集団（コホート）を対象に行うので，コホート分析と呼ばれる [2]．コホート分析において，ユーザ登録を伴うサービスで退会の時期が明らかである場合は生存分析が有用である．しかしユーザ登録を伴わないサービスであれば退会処理がそもそも行われない．したがって一定期間の利用が途絶えたユーザを便宜的に離脱ユーザと見なす．離脱の判断を適切に行うためには利用間隔のモデル化が必要である．

一般的な生存分析では生存時間のモデルとして指数分布やワイブル分布を採用し，生存数を経過時間で説明する．一方，利用が途絶えたユーザが利用を再開するきっかけはプロモーションなどの要因によるものであり，このようなイベントは経過時間が異なるユーザに対して同時に発生する．すなわち経過時間を揃えたとき，ユーザによっては異なる時間にイベントが発生する．そのために利用間隔が利用開始日と従属な関係になる．

以上のことから本稿では，メディアサービスにおける継続ユーザ数の統計モデリングを目的として，経過時間とイベント発生の従属性を考慮に入れた上でのサービスの利用間隔のモデル化を行う．具体的には来訪間隔の統計モデリングを行い，さらにその残差について特異値分解によりユーザの継続要素とイベント発生による復帰要素とに分解する手順を示す．

連絡先: 森下 壮一郎，株式会社サイバーエージェント 秋葉原ラボ，morishita_soichiro@cyberagent.co.jp

2. 定式化

サービス開始時点から一定の時間間隔 w おきに集計を行うこととし，間隔 w を単位としたときのサービス開始からの経過時間を t で表す．そして，期間 $[wt, w(t+1))$ の間にユーザからのアクセスがあった場合を時刻 t における来訪とする．

ところでユーザの来訪間隔を確率的に扱う場合，平均到着時間が λ の指数分布に従うものとして表現するのが一般的である．しかしながらウェブ上のメディアサービスにおいてはユーザの来訪間隔が冪分布に従う場合がある [3]．このとき，ユーザの来訪間隔 τ の確率分布は $f(\tau) = P(X = \tau)$ として $f(\tau) \sim \tau^{-\alpha}$ と表現できる．このことから，冪分布に従う値は両対数グラフで一次直線上に分布する．

図 1 にその一例を示す．これは横軸を来訪間隔 τ，縦軸を UU（unique user）数とした両対数グラフである．ただし，縦軸についてはグラフ中の直線の傾きが 1 となるように正規化している（このグラフの元となったデータの詳細は 4 節で述べる）．グラフ中の直線は $\tau > 1$ であるデータを対象にして単回帰モデルで回帰分析を行って得られた回帰直線である（説明変数と被説明変数の両方を対数変換した上での単回帰であるので，実際には非線形回帰である）．なお $\tau = 1$ の場合は本稿のモデル上では継続利用を意味し，来訪間隔が空いていないと見なすべきなので回帰モデルからは除外している．このグラフからユーザの来訪間隔が冪分布に従う様子が確認できる．次節

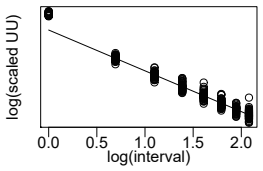

図 1: 冪分布に従う例

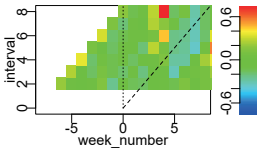

図 2: 残差のヒートマップ

で，この条件における残差を説明するモデルについて述べる．

3. 残差のモデリング

前節までで述べたモデルにおける残差のヒートマップの例を図 2 に示す．横軸は時刻 t，縦軸は来訪間隔 τ である．なお τ が 0 および 1 のときは，冪分布に従うモデルに含まないので除外している．斜めの破線上の要素は時刻 t で来訪したユーザが，その後 $t+\tau$ の各時刻に来訪した数に対応している．垂直な点線上の要素は時刻 $t-\tau$ の各時刻に来訪したユーザが，時刻 t に来訪した数に対応している．例えばヒートマップにおいて右上の $(t,\tau)=(8,8)$ の要素に着目すると，これは時刻 $t-\tau=8$ に来訪したユーザが時刻 $t=8$ に来訪した数に対応している．

ここで時刻 t に来訪したユーザの継続しやすさ（以下，継続要素）$\boldsymbol{p}=\{p_t\}$ と，時刻 t に発生したイベントのユーザの復帰させやすさ（以下，復帰要素）$\boldsymbol{q}=\{q_t\}$ とを考える．そして残差 $r_{t\tau}$ を $\boldsymbol{p}$ と $\boldsymbol{q}$ との線形結合で表現するために，残差を列挙したベクトル $\boldsymbol{r}=\{r_{t\tau}\}$ について次式が成り立つことを仮定する．

$$\boldsymbol{r} = A \begin{pmatrix} \boldsymbol{p} \\ \boldsymbol{q} \end{pmatrix}$$

ここで行列 A は $r_{t\tau}=p_{t-\tau}+q_t$ が成り立つように要素の値を設定した係数行列である．特異値分解で一般化逆行列 A^+ を求めることにより残差 $\boldsymbol{r}$ から $\boldsymbol{p},\boldsymbol{q}$ をそれぞれ求められる．

4. 実証実験

以上で述べた方法の有用性を検証するために，インターネットテレビ局「AbemaTV」*1 へのアクセスログを元にして新規ユーザを含まないコホートを対象に継続の要素と復帰の要素とを分解する実験を行った．なお，このコホートは恣意的に選んだものであるので，以下に示す結果はサービス全体の継続数や離脱数を反映するものではない．

集計期間は 2018 年 1 月 1 日から 2018 年 10 月 1 日である．間隔 w は 1 週間とした．これは週単位の周期性の影響を除くためである．なお図 1 に示したグラフも．離脱判定の閾値 $n=8$ として，集計期間において 8 週目にあたる週（2018 年 2 月 26 日の週）を $t=0$ の週として，提案手法を適用した．結果を図 3 に示す．左上が継続要素 $\boldsymbol{p}$，左下が復帰要素 $\boldsymbol{q}$，右上がこれらを合成したもの，右下が元の値（残差）である．分解結果の合成により元の残差をよく再現できていることが分かる．

次に継続要素 $\boldsymbol{p}$ と復帰要素 $\boldsymbol{q}$ の散布図を図 4 に示す．時刻 $t=9$ の週はいずれも高い．これは 2018 年 4 月 30 日から始まるゴールデンウィークの週であり，特番等の施策が効を奏したのだと考えられる．また，復帰要素が同程度に高い $t=4$ の週は 2018 年 3 月 26 日から始まる週であるが，これは番組改編直後の新番組のインパクトが反映されている．なお，$t=1$ および $t=2$ の週で特に継続要素および復帰要素が低い．これらはいずれも番組改編の直前の時期であり，相対的に継続要素が低くなっている．

*1 https://abema.tv

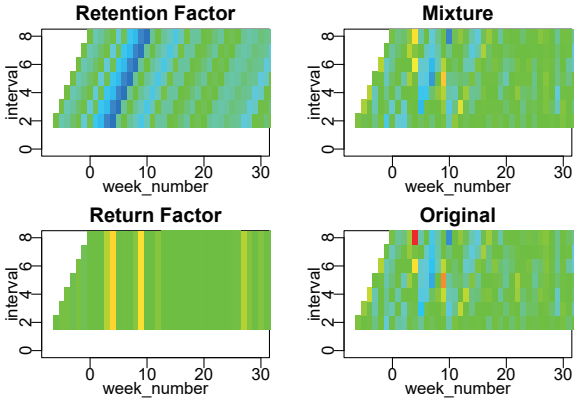

図 3: 分解結果のヒートマップ

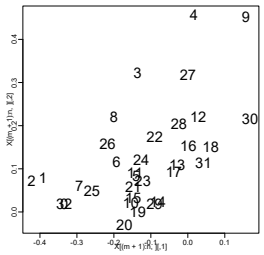

図 4: 分解結果の散布図

5. おわりに

本稿では，ウェブ上のメディアサービスにおける LTV 算出において見積もりが難しいユーザの継続数についての統計モデリングを行った．具体的には，メディアサービスの利用間隔が冪分布に従うことを確認し，さらに統計モデルの残差をユーザの継続しやすさの要素とイベント発生による復帰させやすさの要素との線形和で説明するモデルを提案して，特異値分解によりそれぞれの要素に分解できることを示した．そしてモデルの妥当性を検証するために実環境下で得られたデータに適用して結果を示した．

今後の課題として，周期的な要素の導入や，このモデルに対応したコホートの決定手法の確立などが考えられる．

参考文献

[1] Rust, Roland T., Lemon, Katherine N., Zeithaml, Valarie A.: Return on Marketing: Using Customer Equity to Focus Marketing Strategy, Journal of Marketing, vol.68, no.1, pp.109-127, 2004.

[2] 川口 真一，下村 剛士: プラットフォーム型ビジネスを支えるシステム運営，UNISYS TECHNOLOGY REVIEW 第 136 号，p.37–47, 2018.

[3] 武内 慎，人の音楽鑑賞行動に見られる冪分布に対する現象論的モデルの検証，日本物理学会 2018 年秋季大会，9aM203-3，2018.