日本オペレーションズ・リサーチ学会
2021年 秋季研究発表会

非定常マルコフ過程を用いた人気ダイナミクスの定量的評価

広島大学大学院先進理工系科学研究科　*岩本 和樹 IWAMOTO Kazuki
01307065　広島大学大学院先進理工系科学研究科　土肥 正 DOHI Tadashi
05000041　広島大学大学院先進理工系科学研究科　岡村 寛之 OKAMURA Hiroyuki

1. はじめに

人間の集団行動の理解を通じて，選挙予測，メディア広告の効果予測などを行う研究はトレンド予測と呼ばれ, 社会科学・工学・情報科学のあらゆる分野で適用されている．一方, SNS やウェブ上のコンテンツのアテンション注目度を分析することでトレンド評価などが実施されており, 将来のトレンドを予測する人気ダイナミクスの研究が現在脚光を浴びている．

Shen [1] は，個々のアイテムが人気を獲得する過程を明示的にモデル化するために強化ポアソン過程 (RPP) と呼ばれる確率モデルを提案し, 人気ダイナミクスが 3 つの要素から構成されていると仮定し, RPP モデルの有効性を実データを用いて検証している．

本論文では, 上述の RPP モデルが一般化ポリヤ過程と呼ばれる非定常マルコフ過程 [2] であることを示し, 文献 [1] に含まれる数理的誤りを正すとともに, 時間的緩和関数の異なる形状を仮定し，さらに統計的グループデータに対する尤度関数を用いることによって，より汎用性の高い人気ダイナミクスを記述する確率モデルを提案する．最終的に, 料理レシピ共有サイトの実データを用いた実験を通じて, どのレシピの人気が継続されるかについて緩和関数の観点から考察する．

2. 人気ダイナミクスモデル

時間期間 $(0,T]$ におけるアイテム d の人気ダイナミクスは，注目を受けたときの時間の集合 $\{t_i^d : i = 1,2,\ldots,n_d\}$ によって特徴づけられる．時刻 t までに観測される累積アテンション数 $\{N(t), 0 \le t \le T\}$ は非減少で整数値をとる連続時間計数過程であり, その条件付き強度関数 [1] を

$$\begin{aligned}\zeta(t) &= \lim_{\delta t\to 0}\frac{\mathrm{P}(\text{attentions in } [t,t+\delta t) \mid \Psi_{t-})}{\delta t}\\ &= r_d(N(t-))f_d(t) \qquad (1)\end{aligned}$$

のように定義する．ここで, Ψ_{t-} は確率過程 $N(t)$ の時刻 t 以前までの履歴, $r_d(\cdot)$ は累積アテンション数に依存する関数, $f_d(t)$ は経過時間 t のみに依存する関数である．$r_d(\cdot) = r_d$ が定数のとき, $N(t)$ は強度 $r_d f_d(t)$ の非定常ポアソン過程となり, $f_d(\cdot)$ が定数で $\zeta(t) = r_d(N(t-)) = \lambda_d(t - t_{N(t-)})$ のとき, $N(t)$ は到着時間間隔の故障率が $\lambda_d(\cdot)$ の再生過程となる．

Shen [1] は時刻 t における条件付き強度関数を

$$\zeta(t) = a_d(\alpha + N(t-) - 1)f_d(t;\theta_d) \qquad (2)$$

によって表現し, 対応する $N(t)$ を強化ポアソン過程 (RPP) と呼んだ．ここで, a_d は項目のクオリティを表す定数，$f_d(t:\theta_d)$ は時間的緩和関数であり, アテンションが発生する時間的変化を表わす．θ_d は時間的緩和関数に含まれるパラメータ, $(\alpha + N(t-) - 1)$ は強化機構 (reinforced mechanizm) に相当し, α は非負の整数値パラメータである．

しかしながら、式 (2) を条件付き強度にもつ確率過程はもはやポアソン過程ではなく、一般化ポリヤ過程 [2] と呼ばれる非正常マルコフ過程となる．具体的に, 累積アテンション数 $N(t)$ の確率関数は

$$\begin{aligned}\Pr\{N(t) = n \mid N(0) = 0\} &= \binom{\alpha+n-1}{n} \times \Lambda(t)^{\alpha-1}(1-\Lambda(t))^n,\\ &\quad n = 0,1,\cdots \qquad (3)\\ \Lambda(t) &= e^{-a_d F_d(t;\theta_d)}, \qquad (4)\\ F_d(t;\theta_d) &= \int_0^t f_d(x;\theta_d)dx, \qquad (5)\end{aligned}$$

となり, 平均と分散がそれぞれ

$$\begin{aligned}\mathrm{E}[N(t)] &= (\alpha-1)\left\{e^{a_d F_d(t;\theta_d)} - 1\right\}, \qquad (6)\\ \mathrm{Var}[N(t)] &= (\alpha-1)e^{a_d F_d(t;\theta_d)}\left\{e^{a_d F_d(t;\theta_d)} - 1\right\} \qquad (7)\end{aligned}$$

の負の 2 項分布に従う．

3. 最尤推定

文献 [1] では α を既知のパラメータと仮定しアテンションの発生時刻データから未知パラメータを推定しているが, ここでは α を未知パラメータとし, 人気ダイナミクスのグループデータ $\{(\tau_i, n_i) : i = 1,2,\ldots,m\}$ から推定するものとする．未知パラメータを $\boldsymbol{x} = (a_d, \alpha, \theta_d)$ とすれば, 対数尤度関数は

$$\begin{aligned}&\ln L(\boldsymbol{x})\\ &= \sum_{i=1}^{m}\{\ln(n_i+\alpha-2)! - (n_i-n_{i-1})! - (n_{i-1}+\alpha-2)!\}\\ &- \sum_{i=1}^{m}\Big[\{\alpha-1+n_{i-1}\}\times a_d\cdot\{\Lambda(\tau_i;\theta_d) - \Lambda(\tau_{i-1};\theta_d)\}\Big]\\ &+ \sum_{i=1}^{m}\left\{(n_i-n_{i-1})\times\ln\left[1 - e^{-a_d\{\Lambda(\tau_i;\theta_d)-\Lambda(\tau_{i-1};\theta_d)\}}\right]\right\} \qquad (8)\end{aligned}$$

ここで，1 週間ごとのグループデータを扱うとすれば，τ_i は i 週目の期間を意味し，n_i は時間期間 τ_i までに到着した累積アテンション数である．

これより, 対数尤度関数 $\ln L(\boldsymbol{x})$ を最大にする最尤推定値 $\hat{\boldsymbol{x}} = (\hat{a}_d, \hat{\alpha}, \hat{\theta}_d)$ を求める．

表 1: AIC の比較

	チョコレートケーキ	スポンジケーキ	クッキー	チーズケーキ	スコーン
対数正規分布	**-453014.4**	**-445214.2**	-215137.2	**-381974.4**	-254943.6
切断正規分布	-453013.2	-445212.8	-215136	-381973.4	-254943.4
対数ロジスティック分布	-453013.8	-445214	-215137	-381974.2	-254944
切断ロジスティック分布	-453013.6	-445212.4	-215136	-381973.2	-254943.4
指数分布	-452984.2	-445208.2	-215134.6	-381971.2	-254942
ガンマ分布	-452988.6	-445207.8	-215135.4	-381972.2	-254939
パレート分布	-453010.4	-445212.2	-215135	-381973.2	-254943.2
対数最大値分布	-453011.2	**-445214.2**	**-215137.4**	**-381974.4**	**-254944.2**
切断最大値分布	-453013.8	-445212.4	-215136.2	-381973.2	-254943.2
対数最小値分布	-453014.2	-445214	-215137	-381974.2	-254944
対数最小値分布	-453013.4	-445212.4	-215136.2	-381973.2	-254943.4

4. 数値実験

文献 [1] では, 論文の引用数を分析するために, 緩和関数に対数正規密度関数を仮定し, 既知の α に対して未知パラメータ (a_d, θ_d) を推定している. また, 論文 [1] の検証は式 (6) と式 (7) とは異なる推定値に基づいて行われており, 結果の信憑性に欠くものであった. 本稿ではさらに、時間的緩和関数として対数正規分布に加えて，表 1 に記された計 11 種類のモデルを仮定し, 人気ダイナミクスを説明するための統計分析を実施する.

使用するデータセットは，料理レシピサイト「クックパッド」においてユーザが投稿したレシピに対する「つくれぽ集合」である．したがって，アイテムはレシピであり，アテンションはそれに対する「つくれぽ集合」となる． なお，レシピの選び方として「菓子」「中華料理」「日本食」「洋食」というジャンルの中からつくれぽ数が上位のものを 5 つづつ選んだ． 本稿では，その中から「デザート」レシピの 5 つのデータの分析結果について述べる．モデルの適合性評価尺度として、赤池情報量規準 (AIC)

$$\mathrm{AIC} = -2 \ln L(\hat{x}) + 2\phi \tag{9}$$

を用いる．ここで，ϕ は自由パラメータ数である．AIC が小さいほど適合性の高いモデルとなる．

表 1 は 11 種類の時間的緩和関数を仮定した分析を行った結果であり，最も値の小さい AIC を太字で表示してある．

文献 [1] では緩和関数としてただ一つ対数正規分布を採用していたが，データへの適合性という観点からみると，クックパッドデータのデザートレシピデータにおいては対数正規分布と対数最大値分布が最も適していることがわかる． これは，時間緩和関数を複数仮定することの有効性を示している．

また，図 1 は「デザート」ジャンルのそれぞれのレシピデータの最後のつくレポが観測された時間を 0 として，緩和関数を 1000 週分プロットしたグラフである． これによると，デザートの緩和関数は全体的に下降傾向であり，デザートレシピのつくれぽ数の期待値は今後下降していくことが予測される． 図 2 は「日本食」ジャンルにおける同様のグラフである． デザートと比較すると緩和関数は上昇傾向を示しており，日本食レシピのつくれぽ数は今後上昇していくことが予測される.

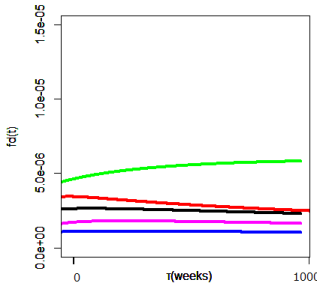

図 1: 緩和関数 (デザート)

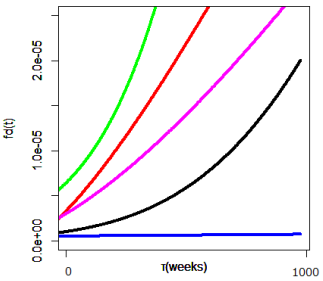

図 2: 緩和関数 (日本食)

5. 結論と今後の課題

本研究では人気ダイナミクスの挙動を一般化ポリヤ過程と呼ばれる非定常マルコフ過程 [2] で記述し, 累積アテンション数の適合性評価を行った．また，緩和関数の観点から人気の出そうなレシピのジャンルについての予測を行った．しかし，予測精度については今回調べていないため，モデルの評価指標として予測精度を検討することも今後の重要な課題である.

謝辞 本研究は，クックパッド株式会社と国立情報学研究所が提供するデータを利用したものである.

参考文献

[1] H. Shen, “Modeling and predicting popularity dynamics via reinforced Poisson processes,” Proceedings of the National Conference on Artificial Intelligence (AAAI-2014), pp. 291297, 2014.

[2] J. H. Cha, “Characterization of the generalized Polya process and its applications,” Journal of Applied Probability, vol. 46, no. 3, pp. 1148–1171, 2014.