ARG WI2 No.11, 2017
WI2-2017-37

アテンションダイナミクスに基づいた オンラインアイテム群の協調構造の抽出

松谷 貫司[†,a]　熊野 雅仁[††,a]　木村 昌弘[††,c]
斉藤 和巳[†††,d]　大原 剛三[††††,e]　元田 浩[†††††,f]

† 龍谷大学大学院理工学研究科電子情報学専攻　†† 龍谷大学理工学部電子情報学科
††† 静岡県立大学経営情報学部　†††† 青山学院大学理工学部情報テクノロジー学科
††††† 大阪大学産業科学研究所

a) *t16m024@mail.ryukoku.ac.jp*　b) *kumano@rins.ryukoku.ac.jp*　c) *kimura@rins.ryukoku.ac.jp*
d) *k-saito@u-shizuoka-ken.ac.jp*　e) *ohara@it.aoyama.ac.jp*　f) *motoda@ar.sanken.osaka-u.ac.jp*

概要　ソーシャルメディアサイトに投稿されたアイデアやニュース，オピニオンなどのオンラインアイテムは，多くの人々に高く評価され共有されていくことによって，そのポピュラリティを獲得していく．このような現象は，人々の日常生活や社会のトレンドにも大きな影響を及ぼす場合があることから，オンラインアイテムのアテンションダイナミクスのモデル化が注目されている．本論文では，対象とするオンラインアイテム群全体の協調構造を抽出することを目的として，ディリクレ過程と Hawkes 過程を融合した新たな確率過程モデルである CHP モデルを提案する．人工データおよび料理レシピ共有サイトの実データを用いた実験により，CHP モデルが将来ポピュラリティの予測において有効であることを示す．さらに，CHP モデルに基づいて共有イベント時系列の発生過程の観点から，料理レシピ共有サイトにおける料理レシピ群の協調構造を明らかにする．

キーワード　協調構造，ソーシャルメディアマイニング，確率過程モデル

1　はじめに

Facebook，Twitter，YouTube，@cosme，Cookpad など，ソーシャルメディアサイトが Web 空間における人々の重要なコミュニケーションの場として発展し続けている．人々はソーシャルメディアを利用して，アイデアやニュース，オピニオンなどの多種多様な情報を容易に，世界に向けて広く発信したり評価したりすることができるようになってきた．ソーシャルメディアサイトに投稿されたそのようなオンラインアイテムは，多くの人々に高く評価され共有されていくことによって，そのポピュラリティを獲得していく．このような現象は，人々の日常生活や社会のトレンドにも大きな影響を及ぼす場合があるので，オンラインアイテムが共有されポピュラリティを獲得していく過程のモデル化が，近年ソーシャルメディアマイニングの分野で注目されている [1, 2, 3, 4, 5]．

ソーシャルメディアサイトに投稿されたオンラインアイテムに対し，それがいつ共有されたかについては一般に観測可能であるが，それを誰が共有したかについてはプライバシーやセキュリティの関係上，必ずしも公開されているとは限らない．したがって少なくとも，オンラインアイテムの共有イベント発生に関する時系列データについては，一般にそれを容易に獲得できると言え，本論文で我々は，ソーシャルメディアサイトに投稿されたオンラインアイテム群に対して，それらの共有イベント時系列の発生過程を同時にモデル化するという問題に取り組む．まず，個々のアイテムはそれ独自の魅力を有していると考えられ，またアイテムの共有イベントについては，その過去の共有イベントがそれ自身の将来の共有イベントの発生を誘発するという，自己エキサイテーション性を有していると考えられる．さらに，こうしたエキサイテーションの性質は異なるアイテム間にも存在し，アイテムの共有イベントは相互エキサイテーション性をもつと仮定するのが自然である．すなわち，あるアイテムの共有イベントは，別のアイテムの将来の共有イベントの発生をも誘発しうると考えられる．このようなアイテム間の相互エキサイテーション性を特徴づけるために，アイテム α' からアイテム α への影響度 $\tilde{w}_{\alpha,\alpha'}$ を考える．このとき，すべてのアイテムペア (α, α') に対して影響度 $\tilde{w}_{\alpha,\alpha'}$ が異なりうると仮定することは現実的でない．実際，アイテムは多数あるが，それに比べて相対的に少量の共有イベント観測データしか得られないことが多く，すべてのアイテム間の影響度を推定することは一般に困難と考えられるからである．

本論文では，共有イベント発生の時系列に基づいて，オンラインアイテム群における関係性を抽出するための確率モデルを提案し，各アイテムの将来ポピュラリティを精度よく予測することを目指す．そのためにまず，我々はアイテム群における**協調構造** $\mathcal{A}_1, \ldots, \mathcal{A}_K$ を導入する．ここに，任意の協調グループ $\mathcal{A}_k$ に対して，それに属する各アイテム α' は，共有イベントの相互エキサ

イテーション性に関し，任意のアイテム α に等しく影響 $w_{\alpha,k}$ を及ぼし，アイテム α' からアイテム α への影響度 $\tilde{w}_{\alpha,\alpha'}$ は $\tilde{w}_{\alpha,\alpha'} = w_{\alpha,k}$ で与えられる．例えば料理レシピ共有サイトでは，そうした協調グループはレシピのジャンルに対応している可能性がある．提案モデルは，計数過程 [6] の一種であり，“rich-get-richer” 現象を捉えるためによく用いられる Hawkes 過程 [7] を土台としている．我々は，ディリクレ過程 [8] を新たなやり方で融合することで Hawkes 過程に協調構造を組み込み，指定されたアイテム群すべてに対して，それらの共有イベント時系列の発生過程を同時にモデル化する．我々の提案モデルを ***CHP* モデル** *(cooperative Hawkes process model)* と呼ぶ．

我々は，共有イベント時系列の観測データから CHP モデルを効率よく推定する手法を開発し，人工データおよび料理レシピ共有サイトの実データを用いた実験により，CHP モデルが既存の Hawkes 過程モデルよりもポピュラリティ予測において有効であることを示す．さらに，CHP モデルに基づいて，共有イベント時系列の発生過程の観点から，料理レシピ共有サイトにおける料理レシピ群の協調構造を明らかにする．

本論文の構成は以下のとおりである．2 章では，Hawkes 過程について述べ，CHP モデルとその推定法を提案する．3 章では，人工データおよび料理レシピ共有サイトの実データに対する実験結果を報告する．最後に，4 章はまとめである．

2 モデル

対象とするオンラインアイテムの集合 $\mathcal{A}$ を固定し，期間 $[0, T)$ でのそれらに対する共有イベント系列の発生過程を同時にモデル化すること考える．ここに，T はそれほど大きくない正数（例えば，2,3 カ月）である．任意のアイテム $\alpha \in \mathcal{A}$ に対して，時刻 t までのその共有イベントを計数過程 $N_\alpha(t)$ [6] としてモデル化することを考える．ここに，$N_\alpha(t)$ は期間 $[0, t)$ 内での α の共有イベント数を表す．$N(t)$ を期間 $[0, t)$ 内での $\mathcal{A}$ に対する共有イベントの総数，すなわち，$N(t) = \sum_{\alpha \in \mathcal{A}} N_\alpha(t)$ とする．各 $n = 1, \ldots, N(t)$ に対して，第 n 共有イベントを組 (t_n, α_n) で表す．これは，アイテム α_n が時刻 t_n に共有されたことを意味する．時刻 t までの $\mathcal{A}$ に対する共有イベント系列を，$\mathcal{T}(t) = \{(t_n, \alpha_n);\, n = 1, \ldots, N(t)\}$ とする．また，任意のアイテム $\alpha \in \mathcal{A}$ に対して，時刻 t までのその共有イベント系列を，$\mathcal{T}_\alpha(t) = \{(t_n, \alpha);\, n = 1, \ldots, N_\alpha(t)\}$ とする．各 $\alpha \in \mathcal{A}$ に対して，$\lambda_\alpha(t)$ を α に対する強度関数とする．ここに，$\lambda_\alpha(t)$ は，時刻 t までの共有イベントの観測系列 $\mathcal{T}(t)$ が与えられたとき，微小な時間窓 $[t, t+dt)$ 内で α の共有イベントが発生する条件付き確率，

$$\lambda_\alpha(t)\, dt \;=\; \mathbb{E}[dN_\alpha(t) \mid \mathcal{T}(t)]$$

を表している（[9] 参照）．本章では，強度関数 $\lambda_\alpha(t)$ のモデル化について考える．

2.1 Hawkes 過程

まず，基本的な Hawkes 過程について述べる．

2.1.1 一様ポアソン過程

各アイテム $\alpha \in \mathcal{A}$ は，それ固有の魅力 $\mu_\alpha > 0$ をもつと考えられる．最も単純な設定では，$\lambda_\alpha(t)$ が $\mathcal{T}(t)$ と独立でありかつ定数であると仮定することである．すなわち，

$$\lambda_\alpha(t) = \mu_\alpha$$

である．これは一様ポアソン過程と呼ばれる．

2.1.2 アイテム間に相互作用のない Hawkes 過程 (HP)

アイテムの共有イベントは，自己エキサイテーション性をもち，“rich-get-richer” 現象を示すと考えられる．これは，アイテム間に相互作用のない Hawkes 過程，

$$\lambda_\alpha(t) = \mu_\alpha + \tilde{w}_{\alpha,\alpha} \sum_{(t_n,\alpha) \in \mathcal{T}_\alpha(t)} \exp\{-\tilde{\gamma}_\alpha (t - t_n)\} \quad (1)$$

として捉えられる．ここに，$\tilde{w}_{\alpha,\alpha} > 0$ は共有イベントの自己エキサイテーションにおけるアイテム α からそれ自身への影響度を表し，$\tilde{\gamma}_\alpha$ は α からの影響度の時間減衰率を表している．この場合，$\lambda_\alpha(t)\, dt \;=\; \mathbb{E}[dN_\alpha(t) \mid \mathcal{T}_\alpha(t)]$ となることに注意しておく．

2.2 多変量 Hawkes 過程 (MHP)

アイテムの共有イベントは相互エキサイテーション性をも有すると考えられるので，一般に，多変量 Hawkes 過程，

$$\lambda_\alpha(t) = \mu_\alpha + \sum_{(t_n,\alpha_n) \in \mathcal{T}(t)} \tilde{w}_{\alpha,\alpha_n} \exp\{-\tilde{\gamma}_{\alpha_n} (t - t_n)\} \quad (2)$$

がモデル化において必要となる．ここに，$\tilde{w}_{\alpha,\alpha_n} > 0$ は，共有イベントの相互エキサイテーションにおけるアイテム α_n からアイテム α への影響度を表す．しかしながら 1 章でも述べたように，$|\mathcal{A}|$ と比べて $|\mathcal{T}(T)|$ が十分に大きくないような現実問題においては，すべてのアイテムペア $(\alpha, \alpha') \in \mathcal{A} \times \mathcal{A}$ に対して，$\tilde{w}_{\alpha,\alpha'}$ が異なる値をもつと仮定することは現実的でない．したがって我々は，$\mathcal{A}$ 内の関係構造を考慮に入れ，各アイテム $\alpha \in \mathcal{A}$ に対する将来の共有イベントを正確に予測することを目指す．

2.3 協調構造

式 (2) で定義される多変量 Hawkes 過程に対して，我々は $\mathcal{A}$ 内に**協調構造** $Z = \{z(\alpha);\, \alpha \in \mathcal{A}\}$ を次のように導入する．$\mathcal{A} = \bigcup_{k=1}^{K} \mathcal{A}_k$ (disjoint union) であり，$z(\alpha') = k$

であるのは，$\alpha' \in \mathcal{A}_k$ であるときに限る．そして，強度関数 $\lambda_\alpha(t)$ は，

$$\lambda_\alpha(t \mid Z) = \mu_\alpha + \sum_{(t_n,\alpha_n)\in\mathcal{T}(t)} w_{\alpha,z(\alpha_n)} \exp\{-\gamma_{z(\alpha_n)}(t-t_n)\} \quad (3)$$

となる．ここに，$w_{\alpha,1},\dots,w_{\alpha,K} > 0$ が存在して，$\forall\alpha' \in \mathcal{A}$ に対し，$\tilde{w}_{\alpha,\alpha'} = w_{\alpha,z(\alpha')}$ であり，また，$\gamma_1,\dots,\gamma_K > 0$ が存在して，$\forall\alpha' \in \mathcal{A}$ に対し，$\tilde{\gamma}_{\alpha'} = \gamma_{z(\alpha')}$ である．協調グループ $\mathcal{A}_k$ に属する各アイテムは，任意のアイテム α に $w_{\alpha,k}$ という形で等しく影響を及ぼし，さらに同一の時間減衰率 γ_k をもつということに注意しておく．

2.4 提案モデルとその推定法

一般に，協調グループの数 K を事前に決定することは困難であるため，観測データから K の値を決定できることが望ましい．したがって我々は，ディリクレ過程 [8] を用いて，$\mathcal{A}$ 内の協調構造を多変量 Hawkes 過程にノンパラメトリックベイズ形式で組み込むことにより，$\mathcal{A}$ に対する共有イベント系列を生成する確率モデルを定義する．我々の提案モデルを *CHP (cooperative Hawkes process)* **モデル**と呼ぶ．ここに，その強度関数 $\lambda_\alpha(t)$ は式 (3) で与えられる．

CHP モデルのパラメータは，$\boldsymbol{\mu} = (\mu_\alpha)_{\alpha\in\mathcal{A}}$，$Z = \{z(\alpha); \alpha \in \mathcal{A}\}$，$\boldsymbol{\gamma} = (\gamma_k)_{k=1}^K$ および $W = (\boldsymbol{w}_k)_{k=1}^K$ である．ただし，$\boldsymbol{w}_k = (w_{\alpha,k})_{\alpha\in\mathcal{A}}$ である．これらのパラメータは以下の手順で生成される．まず，無限次元離散確率分布 $\boldsymbol{\pi} = (\pi_k)_{k=1}^\infty$ が，Stick-Breaking 過程から $k = 1,2,3,\dots$ に対して，

$$v_k \mid \beta \sim \mathrm{Beta}(1,\beta), \quad \pi_k = v_k \prod_{\ell=1}^{k-1}(1-v_\ell)$$

と生成される．ここに，$\mathrm{Beta}(1,\beta)$ はパラメータが 1 と $\beta > 0$ のベータ分布である．次に，$k = 1,2,3,\dots$ に対して，$\phi_k = (\gamma_k, \boldsymbol{w}_k)$ は事前確率分布 H から，

$$\phi_k \mid H \sim H$$

と生成される．ランダム測度 G を

$$G = \sum_{k=1}^\infty \pi_k \delta_{\phi_k}$$

と定義する．ここに，δ_ϕ は位置 ϕ におけるアトム，すなわち，ϕ に集中した確率測度である．G は基底分布 H と集中度パラメータ β のディリクレ過程 $\mathrm{DP}(\beta,H)$ に従って分布していることに注意しておく．Z は，G から各 $\alpha \in \mathcal{A}$ に対して，

$$z(\alpha) \mid G \sim G$$

と生成される．さらに $\boldsymbol{\mu}$ は，パラメータ $\boldsymbol{\eta} = (\eta_0,\eta_1)$ のガンマ分布から各 $\alpha \in \mathcal{A}$ に対して，

$$\mu_\alpha \mid \boldsymbol{\eta} \sim \mathrm{Gamma}(\boldsymbol{\eta}) \quad (4)$$

と生成される．ここに，$\eta_0,\eta_1 > 0$ である．

次に，観測系列 $\mathcal{T}(T)$ からパラメータ $Z, K, \boldsymbol{\mu}, \boldsymbol{\gamma}, W$ を推定する手法について述べる．ディリクレ過程は Chinese restaurant 過程 (CRP) [8] と等価であることから，我々は CRP に基づく近似推論アプローチをとる．事前分布 H は，パラメータ $\boldsymbol{\sigma} = (\sigma_0,\sigma_1)$ のガンマ分布とパラメータ $\boldsymbol{\nu} = (\nu_0,\nu_1)$ のガンマ分布の積とし，$\boldsymbol{\gamma}$ と W は，$k = 1,\dots,K$ と $\alpha \in \mathcal{A}$ に対して独立に，

$$\gamma_k \mid \boldsymbol{\sigma} \sim \mathrm{Gamma}(\boldsymbol{\sigma}) \quad (5)$$

$$w_{\alpha,k} \mid \boldsymbol{\nu} \sim \mathrm{Gamma}(\boldsymbol{\nu}) \quad (6)$$

と生成されるとする．ここに，$\sigma_0,\sigma_1,\nu_0,\nu_1 > 0$ である．

提案推定法では，計数過程に関する重ね合わせの原理を用いて学習アルゴリズムを単純化することを考える．そのために，第 n 共有イベント (t_n,α_n) が第 x_n 共有イベント (t_{x_n},α_{x_n}) によって引き起こされたことを表す潜在変数の集合 $X = \{x_n; n = 1,\dots,N(T)\}$ を導入する．ここに，$x_n = 0,1,\dots,n-1$ であり，$x_n = 0$ は第 n 共有イベント (t_n,α_n) がアイテム α_n の固有の魅力によって引き起こされたことを意味する．具体的には，第 n 共有イベント (t_n,α_n) に対する強度関数 $\lambda_{\alpha_n}(t_n,x_n \mid Z)$ を，

$$\lambda_{\alpha_n}(t_n,x_n \mid Z) = \begin{cases} \mu_{\alpha_n} & \text{if } x_n = 0 \\ w_{\alpha_n,z(\alpha_n)} \exp\left\{-\gamma_{z(x_n)}(t_n - t_{x_n})\right\} & \text{if } 1 \le x_n < n \end{cases}$$

と定義する．ここに，$x_n \in \{0,1,\dots,n-1\}$ である．このとき，独立なポアソン過程に対する重ね合わせの原理より，強度関数 $\lambda_{\alpha_n}(t_n \mid Z)$ (式 (3) 参照) は，

$$\lambda_{\alpha_n}(t_n \mid Z) = \sum_{x_n=0}^{n-1} \lambda_{\alpha_n}(t_n,x_n \mid Z) \quad (7)$$

となる．これはよく知られた性質であり ([7, 10] 参照)，計数過程のベイズ推定に対してもよく用いられている ([11, 12] 参照)．式 (7) の分解は我々の推定法においても重要な役割を果たし，$Z, \boldsymbol{\mu}, \boldsymbol{\gamma}, W$ が与えられたときの $\mathcal{T}(T)$ と X の結合尤度を，扱いやすい積の形で計算できる．これにより，パラメータの推定値 $K^*, \boldsymbol{\mu}^*, W^*, \boldsymbol{\gamma}^*$ は次の 4 ステップ，

1) Z の Gibbs サンプリング

2) X の Gibbs サンプリング

3) $\boldsymbol{\gamma}$ の Metropolis-Hastings サンプリング

4) $\boldsymbol{\mu}$ と W のサンプリングおよび $\boldsymbol{\eta}$，$\boldsymbol{\nu}$，$\boldsymbol{\sigma}$ の更新

を反復することにより得られる．このとき，任意の $\alpha \in \mathcal{A}$ と $k \in \{1, \ldots, K^*\}$ に対して事後確率，

$$\theta_{\alpha,k} = P(z(\alpha) = k | \mathcal{T}(T), \boldsymbol{\mu}^*, \boldsymbol{\gamma}^*, W^*, \beta)$$

が推定でき，よって，各 $\alpha \in \mathcal{A}$ に対して K^* 次元離散確率分布 $\boldsymbol{\theta}_\alpha = (\theta_{\alpha,1}, \ldots, \theta_{\alpha,K^*})$ が得られる．$\Theta = \{\boldsymbol{\theta}_\alpha;\ \alpha \in \mathcal{A}\}$ とする．提案推定法では上記のパラメータだけでなく，X の推定値も得られること注意する．

3 評価実験

本章では，人工データおよび料理レシピ共有サイト Cookpad の実データを用いて，CHP モデルの評価を行う．まず，ポピュラリティ予測性能に関して，CHP モデルを従来の Hawkes 過程モデルである HP モデルと MHP モデルと比較する．次に，CHP モデルに基づいて，ポピュラリティダイナミクスの観点から Cookpad データにおける協調構造の分析を試みる．

3.1 評価指標

予測期間 $[T, T_1)$ において予測するべき共有イベント系列を $\mathcal{T}([T, T_1)) = \mathcal{T}(T_1) \setminus \mathcal{T}(T)$ とする．Hawkes 過程モデルの予測性能の評価によく用いられる方法に従い，予測したい共有イベントの発生時刻までのすべての共有イベントを与えて予測性能を評価する．実験では，以下の一般的な 3 つの評価指標を用いた．

- **PL (Prediction log-likelihood) 指標:** PL 指標では，将来の共有イベント群 $\mathcal{T}([T, T_1))$ の尤度，すなわち，予測したい共有イベントにおいて実際に発生した時刻とアイテムが選ばれる尤度を測定する [13, 14, 11, 9]．任意のポアソン過程モデルの強度関数 $\lambda_\alpha(t)$ に対して，PL 指標は，

$$PL = \sum_{(t_n, \alpha_n) \in \mathcal{T}([T, T_1))} \ln \lambda_{\alpha_n}(t_n) - \int_T^{T_1} \sum_{\alpha \in \mathcal{A}} \lambda_\alpha(t) \mathrm{d}t$$

と定義される．これはポアソン過程における $\mathcal{T}([T, T_1))$ の対数尤度と同じ形であることに注意しておく．

- **LI (Log-likelihood of items) 指標:** PL 指標と異なり LI 指標では，任意の将来の共有イベント $(t_n, \alpha_n) \in \mathcal{T}([T, T_1))$ に対して，その時刻 t_n が与えられたときにアイテム α_n が選ばれる確率，すなわち，将来の共有イベントの時刻が与えられた際にどのアイテムが共有されるかを予測する性能を測定する [11]．LI 指標は強度関数 $\lambda_\alpha(t)$ を用いて，

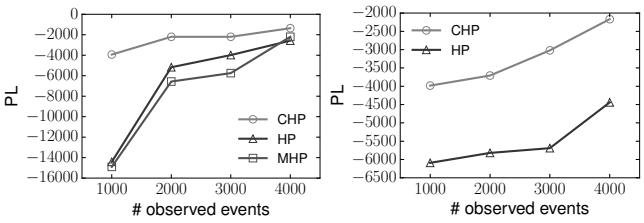
$$LI = \sum_{(t_n, \alpha_n) \in \mathcal{T}([T, T_1))} \ln \frac{\lambda_{\alpha_n}(t_n)}{\sum_{\alpha' \in \mathcal{A}} \lambda_{\alpha'}(t_n)}$$

と定義される．ここに，対数の真数はそれらの総和が 1 となるように正規化されている．

- **AR (Average rank) 指標:** 将来の共有イベントの時刻が与えられたとき，LI 指標がその時刻に実際に発生した真のアイテムの生起確率を測定するのに対し，AR 指標ではより厳格に，その時刻で予測モデルが真のアイテムを選ぶ順位を測定する [9]．AR 指標は，

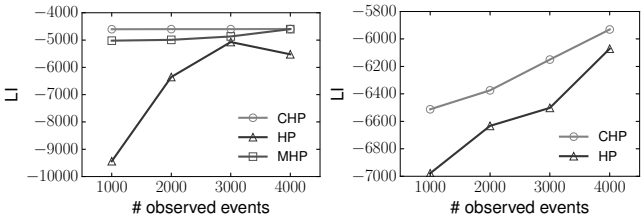
$$AR = \frac{1}{|\mathcal{T}([T, T_1))|} \sum_{(t_n, \alpha_n) \in \mathcal{T}([T, T_1))} rank(\alpha_n; t_n)$$

と定義される．ここに，$rank(\alpha; t)$ は強度関数 $\lambda_\alpha(t)$ に従って時刻 t における α の順位を返す関数である．



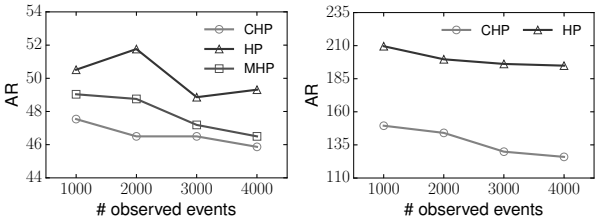
(a) 人工データの結果　(b) Cookpad データの結果

図 1: PL 指標による予測性能の評価



(a) 人工データの結果　(b) Cookpad データの結果

図 2: LI 指標による予測性能の評価



(a) 人工データの結果　(b) Cookpad データの結果

図 3: AR 指標による予測性能の評価

3.2 人工データによる予測性能の評価

CHP モデルを導入する意義とその学習法の有効性を検証するために，まず，CHP モデルから生成したアイテム群の共有イベント系列の人工データを用いて，CHP モデルを 2.1 節および 2.2 節で述べた既存の Hawkes 過程モデルである HP モデルおよび MHP モデルと予測性能の観点から比較した．

アイテム数を $|\mathcal{A}| = 100$，協調グループ数を $K = 10$ と設定し，CHP モデルに基づいてアイテム群の共有イベント系列の人工データを生成し，4 つのデータセットを構築した．ここに，観測データ $\mathcal{T}(T)$ は $|\mathcal{T}(T)| = 1000$, 2000, 3000, 4000 であり，それぞれに対し予測データ $\mathcal{T}([T, T_1))$ は $|\mathcal{T}([T, T_1))| = 1000$ である．HP モデルおよび MHP モデルの学習は，CHP モデルの学習と同様な手法を用いた．また，これら 3 つのモデルともパラメータ推定におけるサンプリングでは，200 回の burn-in を含む 1,000 回の反復を行った．

PL 指標，LI 指標，AR 指標により，CHP モデル，HP モデルおよび MHP モデルの予測性能を比較した．5 回の試行における結果を図 1(a), 2(a), 3(a) に示す．既存モデルは観測イベント数が少ない場合に性能が劣化したのに対し，提案モデルは常に最も性能が高かった．このことは，提案する CHP モデルは既存の Hawkes 過程モデルでは捉えるのが困難な新たな性質を有していること，および開発した CHP モデルの学習法は有効であることを示している．ところで，観測イベント数 4,000 のデータセットに対しては CHP モデルと MHP モデルの性能差は小さくなったが，これは，アイテム数に対して十分に多数の共有イベントが観測されたならば，原理的には CHP モデルを包含している MHP モデルが，観測データから CHP モデルを同定できる可能性があることを示唆している．しかしながら，そのような大量データを獲得することは，一般には困難であることに注意しておく．

3.3 実データによる予測性能の評価

次に，協調構造によりアイテム間の影響関係を組み込むことの有効性を示すために，実データを用いて，提案モデルである CHP モデルと既存モデルである HP モデルを予測性能 (PL, LI, AR) の観点から比較した．実験では，料理レシピ共有サイト Cookpad [1]の実データを用いた．Coopad においてユーザは創作した料理レシピを投稿でき，また，別のユーザはそれらのレシピに賛意を表すメッセージ「つくれぽ」を投稿できる．ここでは，料理レシピとそれに対する賛意メッセージをそれぞれアイテムとその共有イベントとみなした．

共有イベント発生回数の日々の変動に関する定常性を考慮して，Cookpad の 2007 年データを対象とした．2007 年を期間 1 (1 月 1 日から 6 月 30 日) と期間 2 (7 月 1 日から 12 月 31 日) の 2 つの期間に分割し，以下の手順でデータセットを構築した．まず，各期間において，最初の 1 カ月間 (期間 1 では 1 月 1 日から 31 日，期間 2 では 7 月 1 日から 31 日) に 5 回以上共有されたアイテムを対象とした．期間 1 データセットおよび期間 2 データセットのアイテム数は，それぞれ 620 および 1,095 であった．次に，各期間 $j = 1, 2$ に対して，最初から順に対象アイテムの共有イベントを調べていくことにより，4 つのデータセット $\mathcal{D}_1^j, \mathcal{D}_2^j, \mathcal{D}_3^j, \mathcal{D}_4^j$ を構築した．ここに，人工データによる実験と同様，観測データ $\mathcal{T}(T)$ は $|\mathcal{T}(T)| = 1000, 2000, 3000, 4000$ であり，それぞれに対し予測データ $\mathcal{T}([T, T_1))$ は $|\mathcal{T}([T, T_1))| = 1000$ である．例えば $\mathcal{D}_4^2$ は，期間 2 での 1,095 アイテムに対するイベント数 4,000 の観測データ $\mathcal{T}(T)$ とイベント数 1,000 の予測データ $\mathcal{T}([T, T_1))$ から構成されている．各モデルの学習については，人工データによる実験の場合と同じ設定で行った．

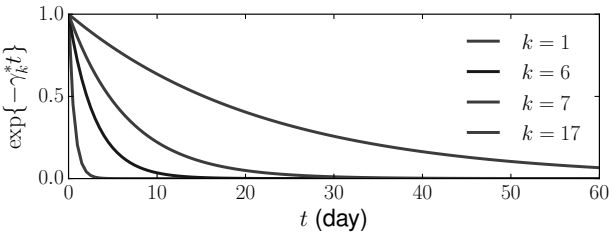


図 4: 抽出された協調グループの指数減衰関数

表 1: 抽出された協調グループの代表レシピ

グループ	タイトル
$\mathcal{A}_1$	ナンプラーそうめん@タイ風 そうめん☆彡 夏バテ知らずの☆そうめん
$\mathcal{A}_6$	レンジで簡単ピクルス とうもろこしは皮ごと簡単レンジでチン！ ノンオイル☆ノンフライ☆ポテトチップス
$\mathcal{A}_7$	揚げない傷めない合せ調味料なし☆麻婆なす おろし玉ねぎでしょうが焼き 豚バラに、甘酢ネギ胡麻だれ。
$\mathcal{A}_{17}$	超気持ちイイ レタスの芯のとり方♪ オーブンいらずのふわふわパン 無駄のない グレープフルーツの剥き方♪

期間 1 データセットと期間 2 データセットの結果は同様であったので，ここでは期間 2 データセットの結果のみを述べる．図 1(b), 2(b), 3(b) に 5 回試行の結果を示す．CHP モデルは，PL, LI, AR のすべての評価指標で HP モデルよりも常に高性能であった．これらの結果は，Cookpad における共有イベント発生過程が相互エキサイテーション性を有し，CHP モデルがそうした性質を過学習することなく捉えられること，そして，協調構造を組み込むことに意義があったことを示唆している．したがって，アイテム間の相互作用の構造を適切に組み込んだ CHP モデルの有効性が実証された．

3.4 実データにおける協調構造の分析

CHP モデルに基づいて，ポピュラリティダイナミクスの観点から Cookpad における協調構造を分析した．紙

[1] https://cookpad.com/

数の関係上，ここでは Cookpad の期間 2 データセットに対する結果のみを報告する．

協調グループ数の推定値 K^* は 22 であった．分析の容易化のために，協調グループのインデックス $k = 1, \ldots, 22$ を γ_k^* が大きい順に並べ替えた．γ_k^* の値が小さくなるほど，指数減衰関数における減衰は緩やかになることに注意しておく．抽出された協調グループ群において，4 つの代表的な指数減衰関数 ($k = 1, 6, 7, 17$ における指数減衰関数) を図 4 に示す．また，これらの協調グループ $\mathcal{A}_1, \mathcal{A}_6, \mathcal{A}_7, \mathcal{A}_{17}$ に対して，$\theta_{\alpha,k}$ の値に基づくランキングの上位 3 位までの代表的レシピを表 1 に示す．まず，$\mathcal{A}_1$ に属するレシピは日本の夏の人気メニューである素麺に関連したものであり，それらの影響は最も急速に減衰していた．$\mathcal{A}_6$ に属するレシピは電子レンジで簡単かつ短時間で調理されるレシピに関連しており，それらの影響は急速に減衰するものの，$\mathcal{A}_1$ よりは緩やかであった．$\mathcal{A}_7$ に属するレシピは日本の家庭料理の惣菜に関連したものであり，それらの影響は $\mathcal{A}_6$ のよりもさらに緩やかに減衰していた．$\mathcal{A}_{17}$ に属するレシピは効率よく料理をするための新規技法と関連しものであり，それらは長期的に影響を与え続けていた．

4 まとめ

本論文では，ソーシャルメディアサイトを対象とし，共有イベント系列に基づいてオンラインアイテム群の協調構造を抽出するために，新たな確率過程モデルである CHP モデルを提案し，各アイテムの将来ポピュラリティを予測することを試みた．CHP モデルは，Hawkes 過程にディリクレ過程を新たなやり方で組み込むことにより構築され，アイテムに対する共有イベントの時系列を生成する．我々は，共有イベントの観測系列から CHP モデルを推定する効率的なベイズ学習法を開発し，さらに，多変量 Hawkes 過程に対する Ogata のアルゴリズムを拡張することにより，CHP モデルの下で将来の共有イベントを予測する有効な枠組みを与えた．

人工データおよび Cookpad データを用いた実験において，PL，LI，AR の 3 つの指標の観点から，CHP モデルとアイテム間に相互作用のない Hawkes 過程 (HP) モデルおよび多変量 Hawkes 過程 (MHP) モデルとの予測性能を比較した．そして，CHP モデルは HP モデルと MHP モデルよりも予測性能が高いこと，特に，観測された共有イベント数が少ない場合には MHP モデルとの性能差がより顕著になることを実証し，協調構造を考慮する CHP モデルの有効性を示した．また，Cookpad データを用いた実験では，CHP モデルに基づいて，ポピュラリティダイナミクスの観点から，Cookpad における料理レシピ間の協調構造を同定し，料理レシピの協調グループに関するいくつかの特徴的な性質を明らかにした．

謝辞

本研究は JSPS 科研費 JP17K00433 の助成を受けたものである．クックパッド株式会社と国立情報学研究所が提供するクックパッドデータを利用した．

参考文献

[1] Szabo, G. and Huberman, B., “Predicting the popularity of online content,” *Communications of the ACM*, vol. 53, no. 8, pp. 80–88, 2010.

[2] Wang, D., Song, C., and Barabási, A.-L., “Quantifying long-term scientific impact,” *Science*, vol. 342, no. 6154, pp. 127–132, 2013.

[3] Shen, H., Wang, D., Song, C. *et al.*, “Modeling and predicting popularity dynamics via reinforced poisson processes,” in *Proceedings of AAAI’14*, 2014, pp. 291–297.

[4] Gao, S., Ma, J., and Chen, Z., “Modeling and predicting retweeting dynamics on microblogging platforms,” in *Proceedings of WSDM’15*, 2015, pp. 107–116.

[5] Zhao, Q., Erdogdu, M., He, H. *et al.*, “Seismic: A self-exciting point process model for predicting tweet popularity,” in *Proceedings of KDD’15*, 2015, pp. 1513–1522.

[6] Aalen, O., Borgan, O., and Gjessing, H., *Survival and event histrory analysis: A process point of view*. Springer, 2008.

[7] Hawkes, A., “Spectra of some self-exciting and mutually exiting point process,” *Biometrika*, vol. 58, no. 1, pp. 83–90, 1971.

[8] Neal, R. M., “Markov chain sampling methods for dirichlet process mixture models,” *Journal of Computational and Graphical Statistics*, vol. 9, no. 2, pp. 249–265, 2000.

[9] Farajtabar, M., Wang, Y., Gomez-Rodriguez, M. *et al.*, “Coevolve: A joint point process model for information diffusion and network evolution,” *Journal of Machine Learning Research*, vol. 18, no. 41, pp. 1–49, 2017.

[10] Farajtabar, M., Du, N., Gomez-Rodriguez, M. *et al.*, “Shaping social activity by incentivizing users,” in *Proceedings of NIPS’14*, 2014, pp. 2474–2482.

[11] Iwata, T., Shah, A., and Ghahramani, Z., “Discovering latent influence in online social activities via shared cascade poisson processes,” in *Proceedings of KDD’13*, 2013, pp. 266–274.

[12] Linderman, S. and Adams, R., “Discovering latent network structure in point process data,” in *Proceedings of ICML’14*, 2014, pp. 1413–1421.

[13] Zhou, K., Zha, H., and Song, L., “Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes,” in *Proceedings of AISTATS’13*, 2013, pp. 641–649.

[14] Zhou, K., Zha, H., and Song, L., “Learning triggering kernels for multi-dimensional hawkes processes,” in *Proceedings of ICML’14*, 2013, pp. 1301–1309.