生態系モデルに基づくオンライン活動データの非線形解析※

松原　靖子[†*a)]　　櫻井　保志[†]　　Christos　FALOUTSOS[††]

Ecosystem on the Web: Non-Linear Dynamical Systems for Online Social Activities※

Yasuko MATSUBARA[†*a)], Yasushi SAKURAI[†], and Christos FALOUTSOS[††]

あらまし　本論文では，大規模オンライン活動データのための非線形解析手法である EcoWeb (Ecosystem on the Web) について述べる．本研究では，“Xbox”, “PlayStation”, “Wii” 等のオンライン検索キーワードの出現件数に関する時系列データが与えられたとき，それらのキーワード間の潜在的な関連性や競合性，そして季節性等の重要なパターンを自動抽出することを目的とする．より具体的には，オンラインユーザ活動の推移パターンを，自然界の生態系における種内・種間競争として捉えることで，潜在的なユーザ資源 (ユーザの興味，時間等) を各アクティビティ (Xbox 等のキーワード) がどのように共有，あるいは競合しているかを非線形動的システムとして表現する．実データを用いた実験では，EcoWeb が様々なオンライン活動における長期的な非線形パターンや季節性等の重要なトレンドを発見し，更に，長期的な将来予測を高精度に行うことを確認した．

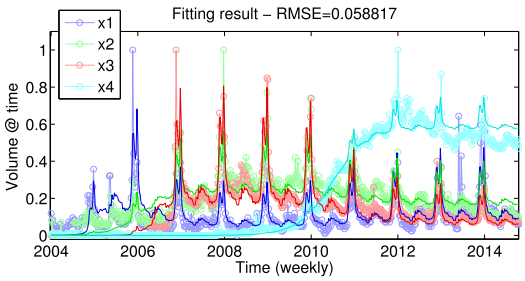
キーワード　生態系モデル，時系列データ，非線形動的システム

1. まえがき

Facebook や Twitter を始めとするオンラインメディアの発展と，Web 上でのユーザ活動の活発化に伴い，社会活動，経済活動等における行動分析の機会が増えつつある [1]〜[4]．例えば，“Xbox”, “PlayStation”, “Wii” 等のオンライン検索キーワードの出現件数に関する時系列データが与えられたとき，本研究の目的は，それらのキーワード間の潜在的なパターンやルールを発見することで，社会学的，行動学的な分析，あるいは，市場調査等の重要なタスクを実現することである．

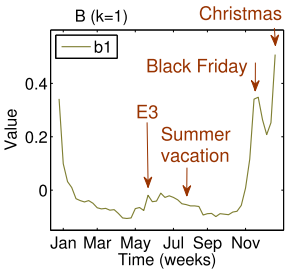
一般に，従来の時系列データ解析手法では，フーリエ変換，ウェーブレット変換，あるいは，自己回帰モデル (AR: autoregressive model)，カルマンフィルタ (KF: Kalman filters) 等の既存技術を用いて特徴分析を行うが，これらの手法では，ドメイン知識に応じた特徴的なパターンをモデル化することができない．しかしながら，多くの場合，時系列データは，ドメインに応じた潜在的なルールに従い自然に生成されるものであり，それらのダイナミックスを理解し表現することで，従来手法にはない，より柔軟な時系列解析を行うことができる．例えば，本研究で扱う Web 上の検索キーワードの時系列データは，ユーザのオンライン活動のパターンに基づき生成されるため，ユーザの潜在的な行動パターンを理解することで，データを新たな側面から解析することができる．より具体的には，本研究では，キーワード（アクティビティ）ごとの活動量（検索数）の推移を，自然界の生態系 (ecosystem) における種 (species) の個体数の推移として捉えることにより，Web 上の活動を非線形動的システムとしてモデル化する．ここで，自然界の種 (species) に対する食料資源 (food resources) は，Web 上でのキーワード（アクティビティ）に対するユーザ資源（ユーザの時間，お金，興味等）に相当し，Web 上において，d 個のキーワード（アクティビティ）がユーザ資源を取り合うことで競合関係を形成し，更に，季節性を伴いながら推移を繰り返すようなシステムであると

† 熊本大学大学院先端科学研究部，熊本市
Faculty of Advanced Science and Technology, Kumamoto University, 2–39–1 Kurokami, Chuo-ku, Kumamoto-shi 860–8555 Japan
†† カーネギーメロン大学，米国
Department of Computer Science, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, United States
* 現在，国立研究開発法人科学技術振興機構，さきがけ
a) E-mail: yasuko@cs.kumamoto-u.ac.jp
※ 本論文は，データ工学研究専門委員会からの推薦論文である．
DOI:10.14923/transinfj.2016DET0002



(a) EcoWeb の学習結果

(b) 競合関係ネットワーク



(c) 季節活動パターン

図 1　EcoWeb を用いた特徴自動抽出と出力結果
Fig. 1　Modeling power of EcoWeb: Our method is *fully automatic*, requiring no user intervention.

仮定する．

本論文では，大規模オンライン活動データのための非線形解析手法である EcoWeb (Ecosystem on the Web) について述べる [5](注1)．より具体的には，以下の問題を扱う．

d 個のキーワード（アクティビティ），n の期間で構成される大規模時系列データ $X = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_d\}$ が与えられたとき，以下の重要なタスクを解決する．

- 潜在的な競合関係の発見（“Xbox” vs. “PlayStation” 等）
- 季節性の発見（クリスマス，夏休み等）
- 将来のユーザの行動予測

具体例. 図 1 は，*Google* (注2) におけるビデオゲーム産業に関連する $d = 4$ 個のアクティビティ：“Xbox” ($\boldsymbol{x}_1$), “PS2, PS3” ($\boldsymbol{x}_2$), “Wii” ($\boldsymbol{x}_3$), “Android” ($\boldsymbol{x}_4$) のクエリ検索数に対する EcoWeb の解析結果を示している．ここでは，2004 年から 2014 年にかけて，週ごとのデータを用いている．EcoWeb は，以下のような重要なパターンを発見することができる．

- 長期的な非線形パターンの学習: 図 1 (a) は，四つのアクティビティ（キーワード）に対するオリジナルデータ（丸印）と提案手法における推定量（実線）を示している．図からも明らかなように，提案モデルは年単位の周期性や，長期的な成長，減衰パターンを柔軟に表現している．例えば，2006 年に発表された “Wii” の成功に伴い，競合関係にある “Xbox” の検索数は一時的に低下している．その後 2011 年にかけて “Wii” の人気が低下傾向にあるが，これは恐らく，従来の Wii のユーザが，モバイル端末 “Android” によるソーシャルゲームへと興味を移したことによる影響であると考えられる．
- 種間競合関係: EcoWeb は，近年のオンラインソーシャルゲーム参入に伴うビデオゲーム産業界における競争関係の複雑化を自動的に抽出することができる．図 1 (b) は，四つのアクティビティ（キーワード）に対する潜在的な関係性を表現した，競合関係ネットワークである．各エッジは二つの異なるキーワード間の競合度を示しており，より太いエッジは，より強い関係性を示す．例えば，“Wii” から “Android” の間の赤いエッジは，ユーザの興味が，“Wii” から “Android” へ移っていく様子を表現している．同様に，“Xbox” は，“PlayStation”，“Wii” と強い競合性（青線）があり，2007 年から 2010 年にかけて一時的な減衰パターンを有していることがわかる．
- 季節性を伴うユーザ活動: 図 1 (c) は，四つのキーワードに共通する周期的な時系列パターンを示している．具体的には，年単位の周期性をもち，11 月の “Black Friday” と 12 月のクリスマスに大きなスパイクをもち．更に，毎年 6 月頃に開催される世界最大のコンピュータゲーム関連の見本市：E3 (Electronic Entertainment Expo) に関連するスパイク，そして夏休みにかけての中期的なピークを伴う．

本論文の貢献. 本研究では，生態系モデルに基づくオンライン活動データの非線形解析手法である EcoWeb を提案する．EcoWeb は次の特長をもつ．

（1） 提案モデルは，潜在的な競合関係や季節活動等の時系列パターンを柔軟に表現し，ユーザの直感に合致した重要な特徴を自動的に抽出する．

(注1)：http://www.cs.kumamoto-u.ac.jp/~yasuko/software.html
(注2)：http://www.google.com/trends/

（2） 計算コストは入力データのサイズに対して線形である．
（3） 既存手法と比較し，高精度でのパラメータ学習を高速に行うと同時に，非線形性を有する時系列パターンの長期的な予測を行うことができる．

2. 関連研究

時系列データの解析に関する研究は多岐にわたる [3], [6]～[9]．自己回帰モデル (AR: autoregressive model)，線形動的システム (LDS: linear dynamical systems)，カルマンフィルタ (KF: Kalman filters) は代表的な技術であり，これらに基づく時系列の解析と予測手法が数多く提案されている [10]～[12]．時系列ビッグデータの研究としては，TriMine [13] は大規模複合時系列イベントデータのための高速な予測手法であり，FUNNEL [14] は大規模疫病テンソルデータのための非線形モデルである．RegimeCast [15] はデータストリームのリアルタイム将来予測に焦点を当てている．文献 [16] では多次元時系列シーケンスのための特徴自動抽出手法を提案した．Rakthanmanon らは文献 [17] において，兆単位 (“trillions”) の時系列シーケンスを対象とした DTW の類似探索問題を扱っており，Yang らは時系列イベントシーケンスのためのモデルを提案した [18]．

ソーシャルメディアとオンラインユーザ活動の分析に関する研究も活発化している [19]～[26]．Gruhl ら [27] はブログ等のオンライン活動と Amazon.com における売り上げの関係性に着目し，Ginsberg ら [1] は，オンライン検索数の推移からインフルエンザの流行をトラッキングし，実際のインフルエンザのウィルスとオンラインのユーザの活動に強い相関があることを示した．文献 [28]～[30] では，キーワードの出現数の推移と消費者の活動の関連性を示している．文献 [31] では，ソーシャルネットワーク上での情報拡散過程をモデル化し，文献 [32], [33] において，それぞれ，コンテンツの再訪パターンと，アクティブユーザ数の推移に関する分析を行っている．Prakash ら [34] は，ネットワーク上において，二つの異なる商品やアイディアがどのように競合するかを議論し，任意のグラフ構造上での理論的なモデル化を行った．

関連研究と本研究の位置づけ. 表 1 は，既存手法と EcoWeb の能力の比較である．

- ロトカ・ボルテラ (LV: Lotka-Volterra) モデル [35]，ロジスティック方程式 (LF: logistic function) [36]，SI (susceptible-infected) モデル [37] や他の非線形方程式 [14], [34], [38], [39] は，ドメイン知識に基づくが，ユーザの活動パターンや周期的なパターンを表現できない．
- ウェーブレット変換やフーリエ変換は単一の時系列シーケンスのための解析手法であり，競合関係のような潜在的な関係性をもつ複数の時系列シーケンスのパターンを表現することができない．
- AR, LDS, SARIMA, TBATS [40]，あるいはその他の関連する予測手法である AWSOM [41], PLiF [11]，TriMine [13] は，全て線形方程式に基づくため，本研究で対象とする非線形性を有する時系列データの表現には適していない [3]．更に，これらの手法はパラメータの設定を要する．
- AutoPlait [16]，SWAB [42]，pHMM [43] は，時系列シーケンスのダイナミックスを表現し，セグメンテーションの能力を有するが，複数の時系列データ集合に対し，長期的な非線形のダイナミックスを表現することができない．

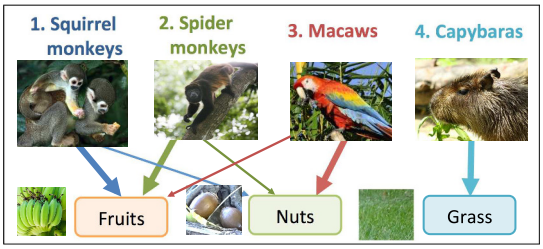
表 1 既存手法との比較
Table 1 Capabilities of approaches.

	LV	DWT	AR++	AutoPlait	**EcoWeb**
ドメイン知識	√				√
多次元シーケンス	√			√	√
周期性		√	√	√	√
自動化				√	√
将来予測			√		√

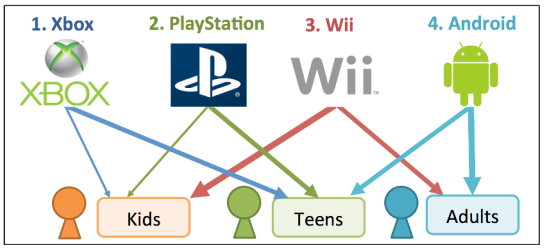
3. 背　　景

ここで，実世界のジャングルにおける生態系を考えてみる．草や果実を食べる草食動物や，その草食動物を狙う肉食動物等，ジャングルには，様々な生き物が生息している．例えば，複数のクモザルとリスザルが，同じテリトリー内においてバナナを食べて集団生活をしている場合，次の時刻（例えば，来月，来年等）には，何頭のクモザルが見込まれるだろうか．これは，個体群生態学における重要な研究対象であり，生物の個体数の推移を数学的にモデル化することは重要な課題である [39], [44]．

生態系における競合関係. 生態系における競合関係を表現するにあたり重要なのは次の二つのメカニズムである．(a) 制約を伴わない上での成長：無限大の食物資源があると仮定した場合の個体の増加率 r．例えば，リスザルの各個体が次の時刻に r 匹の子を産むような状態．(b) 食物資源と種内・種間競争：有限の食物資



(a) ジャングルの生態系



(b) Web 上の生態系

図 2　ジャングル（アマゾン熱帯雨林）と Web 上（ゲーム産業）における生態系の比較
Fig. 2　Illustration of jungle (e.g., Amazon rainforest) vs. Web (e.g., game industry).

源が与えられた上での（つまり，環境収容力 K に対する）個体の最大数 K の制限．ここで，種内競争とは，同種による競合（例えば，2 匹のリスザルが果物を奪い合う状況）であり，種間競争とは，異なる種による競合（リスザル対クモザル）を示す．この競合関係は，種の個体数の爆発的増加を防ぐ効果をもつ．例えば，生態系においてリスザルの個体数が多く，果物の数が少なければ，競争は激化し，リスザルの個体数の成長は制限される．

上記の現象を表現するための最もシンプルな方法として，ロトカ・ボルテラ (LV: Lotka-Volterra) の競争モデルが挙げられる [45]．このモデルは，次の非線形微分方程式を用いて，d 種の競争関係を表現する．

$$\frac{dP_i}{dt} = r_i P_i \left(1 - \frac{\sum_{j=1}^{d} a_{ij} P_j}{K_i}\right), (i = 1, 2, \ldots, d) \tag{1}$$

ここで，

- P_i: i 番目の種の個体数．
- r_i: i 番目の種の成長率．密度調整が生じない場合の繁殖率に相当する ($r_i \geq 0$)．
- K_i: i 番目の種の（競合種が存在しない場合の）環境収容力 ($K_i \geq 0$)．
- a_{ii}: 種内競争．i 番目の種内における食料資源競争 ($a_{ii} = 1$)．
- a_{ij}: 種間競争．異なる 2 種間での競合関係 ($a_{ij} \geq 0$)．

このとき，時刻 t は連続値であり，dP_i/dt は導関数とする．i 番目の種に対し，現在の個体数 P_i に対し成長率 r_i で毎時刻繁殖していく．このモデルは，複数の種間において，共通の資源をめぐり競争関係が生じている状態を表現する．図 2 (a) は，実世界のジャングルにおける野生動物の関係性を示している(注3)．これらの種が，果物や木の実等の食物資源を共有し，特定の地域で共生していると仮定すると，i 番目の種が摂取する食物資源を共有している種の個体総数は，次の式で表現される：$a_{i1}P_1 + \cdots a_{ij}P_j \cdots P_i + \cdots a_{id}P_d = \sum_{j=1}^{d} a_{ij}P_j$．ここで，$a_{ij} (i \neq j)$ は，種間競争係数と呼ばれ，i 番目の種に対する j 番目の種の影響力の強さを表す．

4. 提案モデル

本章では，提案モデルである EcoWeb について述べる．本研究で扱うオンライン活動データ：$X = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_i, \cdots, \boldsymbol{x}_d\}$ は，d 個のキーワード，n の期間で構成され，$\boldsymbol{x}_i$ は，i 番目のキーワードのシーケンス ($\boldsymbol{x}_i = \{x_i(t)\}_{t=1}^{n}$) を表す．本研究の目的は，時系列データ X が与えられたとき，(a) X の非線形ダイナミックスを表現し，(b) 各シーケンス間の潜在的な関係性を発見し，(c) 将来のオンライン活動を予測することである．

ここで，異なる二つのキーワード間の関係性というものはどのように表現することができるだろうか．例えば，Xbox や Wii の間の関係性，あるいは，Facebook と LinkedIn の違いとは一体なんであろうか．これらの Web 上のキーワードは，自然界の野生動物のように競合関係にあると言えるだろうか．

提案モデルの概要. Web 上の生態系とはどのように定義できるだろうか．Web における仮想的なコミュニティの中に，自然界と類似した現象は見られるだろうか．本研究では，Web 上に様々な仮想的な種 (virtual species) が存在し，それらの種が相互作用することにより，時間発展していくと仮定する．

図 2 (b) は，Web における生態系を表している．自

(注3)：Image courtesy of xura, criminalatt, David Castillo Dominici, happykanppy at FreeDigitalPhotos.net.

然界のジャングルにおいて，クモザルやカピバラが生息しているように，Web 上の空間にも Xbox, PlayStation のような仮想的な種（アクティビティ）が存在している．

以下では，提案モデルに対する二つの重要なアナロジーについて述べる．

- **キーワード（アクティビティ）＝ 種 (species)**: Web 上のキーワードは，ユーザからの興味を得ることで存在しており，自然界に生息する生物のように振る舞う．例えば，キーワード（Wii）とユーザ（子供）の関係は，リスザルと果物，あるいは，カピバラと草の関係に似ている．ジャングルにおいても，Web 上においても，資源なしに種は生存できない．
- **ユーザ資源 ＝ 食料資源**: ジャングルの生態系と同様に，Web 上には多くのユーザとユーザ資源が存在する．ここでのユーザ資源とは，例えば，ユーザの興味や注目，時間やお金等に相当する．一般に，ユーザは，同時に複数の目的（アクティビティ）のために時間や興味を消費することができない(注4)．図 2 (b) に示すように，ユーザには様々なグループが存在する．例えば，子供は Xbox, PlayStation, Wii 等のビデオゲームに興味をもつが，大抵の成人は，アンドロイドを好む．

ここで更に，上記のアナロジーに加え，本研究では次の三つの要素が必要となる．

- (**G1**): キーワードの非線形時系列パターン
- (**G2**): 異なるキーワード間の競合関係
- (**G3**): ユーザ活動の季節性

自然界の生態系では，種の個体数は，繁殖率や各個体の寿命に基づき，緩やかに時間発展していく．Web 上においても，各キーワードの注目度（つまり，各ユーザの興味や関心の量）は時間とともに推移していく．もし，ある新商品（アンドロイド等）が魅力的であった場合，多くのユーザがその商品に注目し，友人や他のユーザに推薦する．これにより，この商品の注目度は指数関数的に成長する．本研究では，このような非線形的な時系列パターン (**G1**) を表現するために，非線形差分方程式を用いる．

続いて，異なる二つのキーワード間の潜在的な関係性 (**G2**) について考える．例えば，図 1 (a) において，

(注4)：例えば，N 人のユーザがいる場合，1 日あたり最大で $N \times 24$ 時間のユーザ資源が存在する．各キーワードに対するユーザ資源の量は，対象とするキーワードの種類や，各ユーザの興味の対象により変化する．

表 2　ジャングルと Web における生態系の比較
Table 2　Analogy: Jungle vs. Web.

ジャングル	Web
種 (リスザル，カピバラ)	キーワード・アクティビティ(Wii)
食料資源 (果物，草)	ユーザ資源 (子供，成人)
個体数	注目度
気候・季節	季節のイベント (クリスマス)

表 3　主な記号と定義
Table 3　Symbols and definitions.

記号	定義
d	キーワード（アクティビティ）の総数
n	時系列の長さ
X	d 次元の時系列シーケンス集合 ($X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d\}$)
$\boldsymbol{x}_i$	i 番目のシーケンス ($\boldsymbol{x}_i = \{x_i(1), \ldots, x_i(n)\}$)
$P_i(t)$	i 番目のキーワードの時刻 t における 注目度
$C_i(t)$	i 番目のキーワードの時刻 t における 推定量
$\boldsymbol{p}$	初期注目度 ：$\{p_i\}_{i=1}^{d}$
$\boldsymbol{r}$	成長率 ：$\{r_i\}_{i=1}^{d}$
$\boldsymbol{K}$	環境収容力 ：$\{K_i\}_{i=1}^{d}$
$\mathbf{A}$	競合行列 ($d \times d$) ：$\mathbf{A} = \{a_{ij}\}_{i,j=1}^{d,d}$
n_p	周期 (1 年 ＝ 52 週)
k	季節活動パターンの数
$\mathbf{W}$	重み行列 ($d \times k$) ：$\mathbf{W} = \{w_{ij}\}_{i,j=1}^{d,k}$
$\mathbf{B}$	季節行列 ($k \times n_p$) ：$\mathbf{B} = \{b_j(\tau)\}_{j,\tau=1}^{k,n_p}$

Xbox $\boldsymbol{x}_1$ のシーケンスは，PlayStation $\boldsymbol{x}_2$ のシーケンスと逆向きの成長のパターンをもつ．例えば 2007 年から 2010 年にかけて，PlayStation の値が増加している一方で，Xbox は減少傾向にある．つまり，これらの二つのキーワードは潜在的な競争関係にあると言える．

最後の要素は，年単位の周期性 (**G3**) である．例えば，図 1 (a) において，全ての時系列シーケンスが，クリスマスのスパイクを有する．これは，ユーザの行動パターンが季節やイベントに応じて変化しているためである．自然界における野生動物についても，天候の変化や季節に応じて行動を変化させる現象が見られる．

表 2 は，ジャングルと Web における生態系のアナロジーを示している．本研究では，Web 上のユーザの活動とアクティビティの推移を，ジャングルの生態系における種内・種間競争として捉えることで，潜在的なユーザ資源をキーワード間でどのように共有，競合しているかを表現する．

次節では，提案モデルの詳細を示す．

4.1　EcoWeb-基本モデル (G1)

ここでは，最も簡単な場合として，種間競争の生じない状況，つまり，独立した単一のキーワードに対するシーケンスのモデル化について述べる．K を，単一のキーワードに対するユーザ資源の上限（環境収容

力），p を初期状態（時刻 $t=0$）において既に消費されているユーザ資源の量とする．提案モデルでは，次のルールに従いアクティビティが推移すると仮定する．

- 種内・種間競争が生じない場合には，現時刻の注目度（つまり，ユーザの興味の量）が維持される．
- 時刻 t において，新たなユーザ資源を得ることで，成長率 r で注目度が成長する．

$P(t)$ を時刻 t におけるキーワードの注目度とすると，単一のキーワードに対する時系列パターンの推移は次の差分方程式で表現される．

$$P(t+1)=P(t)\left[1+r\left(1-\frac{P(t)}{K}\right)\right], \tag{2}$$

更に，$P(0)=p$ は初期状態を示し，

- $P(t)$: 時刻 t におけるキーワードの注目度（各ユーザにおけるキーワードへの注目量の総和）
- p: 初期状態，時刻 $t=0$ における注目度．
- r: 成長率，キーワードのユーザに対する魅力の強さ．
- K: 環境収容力，つまり，キーワードに対するユーザ資源（注目度）の上限値．

ここで，$\left[1+r\left(1-\frac{P(t)}{K}\right)\right]$ は，現時刻における注目度 $P(t)$ が，次の時刻における成長へ貢献する率を表し，$\left(1-\frac{P(t)}{K}\right)$ は，キーワードの時刻 t におけるユーザ資源の残量率を示す．例えば，もしキーワードに対するユーザ資源が尽きた場合（つまり $P(t)=K$）には，注目度は定数へと向かう．更に，式 (2) は，単一の種の場合 ($d=1$) のロトカ・ボルテラ競争モデル（式 (1)）を離散的に表したものと同等である．

4.2 EcoWeb-競合関係モデル (G2)

次に，複数のキーワード間の競合関係 (**G2**) について述べる．一般に，関連するキーワードは共通のユーザ資源をもつ．例えば，Xbox と PlayStation の間には明らかな競合関係が存在しており，大抵の場合，ユーザは，価格や好みのゲームタイトル等に応じて一つのゲーム機を選択，購入する．

［**モデル 1**］（EcoWeb-競合関係） $P_i(t)$ を，i 番目のキーワードの時刻 t における注目度とする．EcoWeb-競合関係モデルは，次の式で表現される．

$$P_i(t+1)=P_i(t)\left[1+r_i\left(1-\frac{\sum_{j=1}^{d}a_{ij}P_j(t)}{K_i}\right)\right], \quad (i=1,\cdots,d), \tag{3}$$

ここで，$r_i>0, K_i>0, a_{ii}=1, a_{ij}\geq 0$, $P_i(0)=p_i$ とする．

モデル 1 では，複数のキーワードが共通のユーザ資源に対し競合していると仮定する．時刻 t における，i 番目のキーワードの潜在的なユーザ資源の割合は，次式で表現される．

$$\left(1-\frac{\sum_{j=1}^{d}a_{ij}P_j(t)}{K_i}\right), \tag{4}$$

競合関係係数 a_{ij} は，i 番目のキーワードに対し j 番目のキーワードが影響を与える割合を示す．ここで，もし i 番目と j 番目のキーワード間に競合がない場合には（つまり，$a_{ij}=0$ $(i\neq j)$），中立関係が成り立ち，このモデルは式 (2) と同等である．逆に，$a_{ij}=a_{ji}=1$ である場合には，二つのキーワードは完全に同じユーザ資源のグループを共有する強い競合関係をもち，更に，$a_{ij}=1, a_{ji}=0$ だった場合には，片害関係をもつ．この場合には，i 番目のキーワードが j 番目のキーワードに強い影響を与えられる一方で，j 番目のキーワードは i 番目のキーワードからの干渉を受けない．

4.3 EcoWeb-季節活動モデル (G3)

これまでは，d 個のシーケンス集合が与えられた場合の長期的な非線形ダイナミックスを表現するモデルについて議論したが，Web 上の日々の詳細な活動パターンを表現するには，不十分である．Xbox や Amazon 等のキーワードは，常に一定数のユーザが興味をもっているが，実際のユーザの行動は様々なイベントにより変化している．例えば，Black Friday には Amazon.com へのアクセスが集中し，普段より多くのユーザが見込まれる．これは，自然界の生態系においても見られる現象である．例えば，大抵の場合，サルは，暖かで明るい日中に活動し，夜は睡眠をとる．そして，このような季節性を伴う行動は，他の種（キーワード）と関連していることが多い．例えば，Amazon を含む多くの小売店では，Black Friday に最も売上が集中する．そこで本研究では，上記の現象を反映するため，季節活動パターン (**G3**) のための次のモデルを提案する．

［**モデル 2**］（EcoWeb-full） $C_i(t)$ を i 番目のキーワードの時刻 t における推定量とする．提案モデルは次式を用いてユーザの季節性を伴う活動を表現する．

$$C_i(t)=P_i(t)\left[1+e_i(t)\right] \quad (i=1,\cdots,d), \tag{5}$$

ここで $e_i(t)$ は，i 番目のキーワードの季節活動パターンを表現する．

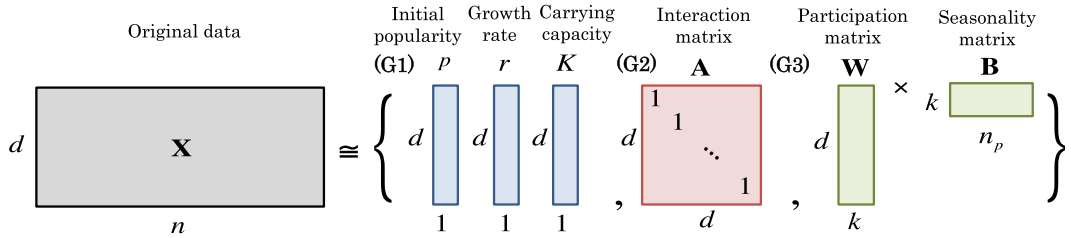


図 3　EcoWeb の概要．提案手法は d 次元のシーケンス X から重要なパターン (**G1**) (**G2**) (**G3**) を自動抽出する．

Fig. 3　Illustration of EcoWeb structure. Given a set of d sequences X of length n, we extract important properties of online activities: (**G1**) (**G2**) (**G3**).

推定量 $C_i(t)$ は，i 番目のキーワードが時刻 t において出現した回数を示し，潜在的な注目度 $P_i(t)$ と季節活動パターン $\mathbf{E} = \{e_i(t)\}_{i,t=1}^{d,n}$ に依存する．$\mathbf{E}$ の各要素は，潜在的な注目度と実際のキーワードの出現回数の相対値を示し，祝日等の季節のイベントに関連する．もし，i 番目のキーワードが時刻 t における季節活動パターンをもたない場合には（つまり $e_i(t) = 0$），キーワードの出現回収は潜在的な注目度と一致し，$C_i(t) = P_i(t)$ となる．

季節活動パターンの表現と圧縮. データ集合 X における全ての季節活動パターン $\mathbf{E}$ を表現するためには，合計で $d \times n$ 個のパラメータを必要であり，このままでは冗長で扱いにくい．更に，ここでの重要な目的として，(a) Black Friday 等の周期的なパターンの発見，(b) 小売店セール等の潜在的な季節活動グループの発見を行いたい．そこで，本研究では，より純度の高いモデル化を行うために，$\mathbf{E}$ を分解する手法を提案する．より具体的には，提案手法は，$\mathbf{E}$ を次の二つの行列に圧縮，及び分解する：季節行列 $\mathbf{B}$ (サイズ $(k \times n_p)$)，重み行列 $\mathbf{W}$ (サイズ $(d \times k)$)．ここで，$\mathbf{B}$ は長さ（周期）n_p の k 個の潜在的な季節行列の成分であり，$\mathbf{W}$ は季節行列の成分が，各シーケンスに及ぼす影響（重み）を表す．まとめると，季節活動 $\mathbf{E} = \{e_i(t)\}_{i,t=1}^{d,n}$ は，次式で表現される．

$$e_i(t) \simeq f(i,t|\mathbf{W},\mathbf{B}) = \sum_{j=1}^{k} w_{ij} b_j(\tau) \qquad (6)$$
$$(\tau = [t \mod n_p])$$

ここで，

- n_p: 周期（1 年 =52 週）．
- k: 季節行列における成分の数．
- $\mathbf{W} = \{w_{ij}\}_{i,j=1}^{d,k}$: 重み行列, i 番目のキーワードの j 番目の成分への影響（重み）．
- $\mathbf{B} = \{b_j(\tau)\}_{j,\tau=1}^{k,n_p}$: 季節行列, 時刻 τ における j 番目の成分の大きさ．

更に，季節行列における成分の数 k は，自動的に推定するべきであり，詳細については次章で述べる．

EcoWeb- パラメータ集合. 図 3 は，提案モデルの概要を示す．d 次元のシーケンス集合 X が与えられたとき，本研究の目的は，次の三つの重要な要素を自動抽出することである．(**G1**) 基本要素：初期注目度: $\boldsymbol{p} = \{p_i\}_{i=1}^{d}$, 成長率: $\boldsymbol{r} = \{r_i\}_{i=1}^{d}$, 環境収容力: $\boldsymbol{K} = \{K_i\}_{i=1}^{d}$; (**G2**) 競合行列: $\mathbf{A} = \{a_{ij}\}_{i,j=1}^{d,d}$; (**G3**) k 個の季節活動パターン (重み行列 $\mathbf{W}$, 季節行列 $\mathbf{B}$).

[**定義 1**]（**EcoWeb のパラメータ集合**） $\mathcal{S}$ を X を表現する全パラメータ集合 $\mathcal{S} = \{\boldsymbol{p}, \boldsymbol{r}, \boldsymbol{K}, \mathbf{A}, \mathbf{W}, \mathbf{B}\}$) とする．

5. 最適化アルゴリズム

本章では，モデルの学習アルゴリズムである EcoWeb-Fit について述べる．より具体的には，次の二つの課題：(1) 最適な季節活動パターン（つまり $\mathbf{W}, \mathbf{B}$）の発見，(2) データ X に対する最適なパラメータ集合 $\mathcal{S}$ の推定に取り組む．

5.1 季節活動パターンの自動分析

まず，一つ目の課題として，最適な季節活動パターン $\mathbf{W}, \mathbf{B}$ の発見について述べる．この課題は，次の二つの部分問題に分解される．

- 季節活動パターンの発見：成分の個数（サイズ）k が固定された上での季節活動行列 $\mathbf{W}$, $\mathbf{B}$ の最適化．
- 自動成分分析：最適な成分の個数（サイズ）k の推定 $(k = 1, 2, \ldots)$．

季節活動パターンの発見. まず，最も単純な部分問題として，シーケンス X と基本パラメータ集合

$\{\boldsymbol{p}, \boldsymbol{r}, \boldsymbol{K}, \mathbf{A}\}$ が与えられている場合を考える．モデル 2 に基づき，季節活動パターン $\mathbf{E}$ は，次式で計算される：

$$e_i(t) = \frac{x_i(t) - P_i(t)}{P_i(t)} \quad (i = 1, \ldots, d; t = 1, \ldots, n). \tag{7}$$

次に提案手法は，計算した $\mathbf{E}$ に対し，要約情報として $\mathbf{W}, \mathbf{B}$ を計算する．ここで，$\mathbf{W}, \mathbf{B}$ の最も簡易的な計算方法として考えられるのは，長さ n の d 個の各シーケンスを，$k = d$ 個の異なる季節活動パターンとして表現することである．しかし，この方法は，全シーケンス X を表現するために，$d \times n$ 個の独立したパラメータ集合が必要となる．この方法は，複数のシーケンス間の共通するパターンを表現することができないため，モデルとして不十分である．

そこで本研究では，与えられた X に対する最適な季節活動パターンを発見するための新たなアルゴリズムを提案する．図 4 は，提案アルゴリズムの概要を示している．季節活動 $\mathbf{E}$ が与えられたとき，提案アルゴリズムは，各シーケンスを長さ n_p の部分シーケンス集合に分割し，行列 $\hat{\mathbf{E}}$ を生成する．ここで，$\hat{\mathbf{E}}$ のサイズは $[d \times \lceil n/n_p \rceil] \times n_p$ である．次に，行列 $\hat{\mathbf{E}}$ から k 個の成分を発見し，季節行列 $\mathbf{B}$ を生成する．重み行列 $\mathbf{W}$ は，式 (6) を用いて，行列 $\mathbf{E}$ の復元エラー（$\mathbf{E} \simeq f(\mathbf{W}, \mathbf{B})$）が最小となるものを選ぶ．ここで重要な課題として，$\hat{\mathbf{E}}$ の中から最適な季節活動パターン $\mathbf{B}$ を抽出する手法が必要である．まず挙げられるのは，主成分分析 (PCA: principal component analysis) を用いた方法である．しかしながら，PCA は時系列データの重要な成分を発見することができる一方で，直行成分しか扱えないという短所がある．そこで本研究では，独立成分分析 (ICA: independent component analysis) を用いた情報抽出手法を行う．ICA は，PCA とは異なり，ガウス性をもたないシーケンスに対し，統計的に独立な成分を k 個発見することができる．

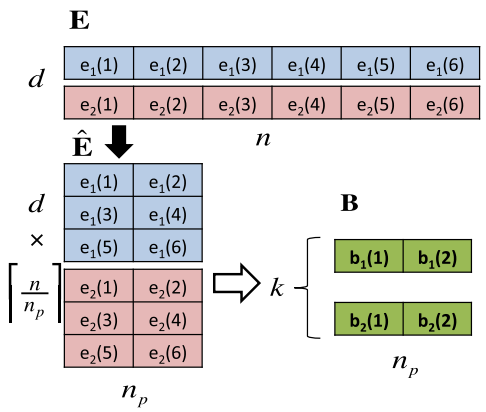


図 4　季節活動パターンの発見の概要（$n_p = 2$ の場合）
Fig. 4　Illustration of seasonal component analysis (for $n_p = 2$).

自動成分分析. 次に，成分の最適な個数 k の推定方法について議論する．本研究では，季節活動パターンを表現する二つの行列 $\mathbf{W}$, $\mathbf{B}$ の最適なサイズを自動的に推定するために，最小記述長 (MDL: minimum description length) に基づく符号化スキームを導入する．直感的には，データがより圧縮できれば，よりよいモデルであるとみなす．

EcoWeb の全パラメータ集合 $\mathcal{S}$ の表現コストは以下の要素から構成される：次元数 d，及び，時系列の長さ n：$\log^*(d) + \log^*(n)$ ビット(注5)．初期注目度, 成長率, 環境収容力 (つまり $\{\boldsymbol{p}, \boldsymbol{r}, \boldsymbol{K}\}$) と競合行列 $\mathbf{A}$ にそれぞれ，$d \times 3$ と $(d \times d - d)$ のパラメータを要する．つまり，$Cost_M(\boldsymbol{p}, \boldsymbol{r}, \boldsymbol{K}) + Cost_M(\mathbf{A}) = c_F \cdot d(3 + d - 1)$ となり，c_F は浮動小数点のコストを示す(注6)．同様に，k 個の季節活動成分は，$Cost_M(k, \mathbf{W}, \mathbf{B}) = \log^*(k) + \log^*(n_p) + c_F(dk + kn_p)$ で表現される．

モデルパラメータ集合 $\mathcal{S}$ が与えられた上での X の符号化コストは，ハフマン符号を用いた情報圧縮により，負の対数ゆう度を用いて次のように表現することができる：

$$Cost_C(X|\mathcal{S}) = \sum_{i,t=1}^{d,n} \log_2 p^{-1}_{Gauss(\mu,\sigma^2)}(x_i(t) - C_i(t)),$$

ここで $x_i(t)$ と $C_i(t)$ はぞれぞれ，i 番目のキーワードの時刻 t におけるオジリナルデータと推定値 (モデル 2) を示す．更に，μ, σ^2 はそれぞれ，オリジナルデータの値と，推定値の間の平均，分散を示す(注7)．まとめると，パラメータ集合 $\mathcal{S}$ に対するデータ X の符号長は次の式で与えられる．

$$Cost_T(X; \mathcal{S}) = \log^*(d) + \log^*(n) + Cost_M(\boldsymbol{p}, \boldsymbol{r}, \boldsymbol{K}) + Cost_M(\mathbf{A})$$

(注5)：ここで，$\log^*$ は整数のユニバーサル符号長を表す．
(注6)：本論文では浮動小数点を $c_F = 8$ ビットにデジタイズした．
(注7)：ここで，μ, σ^2 は $2c_F$ ビットを要するが，これらのコストは定数であるため，モデル推定の際には除外することができる．

$$+Cost_M(k,\mathbf{W},\mathbf{B}) + Cost_C(X|\mathcal{S}) \qquad (8)$$

結論として，提案アルゴリズムは，季節活動成分の最適な個数 k_{opt} を，次式で求めることができる：$k_{opt} = \arg\min_k Cost_T(X;\mathcal{S})$.

5.2 学習アルゴリズム

ここまでは，シーケンス X に対する基本パラメータ集合 $\{\boldsymbol{p},\boldsymbol{r},\boldsymbol{K},\mathbf{A}\}$ が与えられている場合に，どのようにして季節活動行列 $\{\mathbf{W},\mathbf{B}\}$ を発見するかについて述べた．次の課題は，全てのパラメータ集合 $\mathcal{S}$ を高速かつ効率的に推定することである．具体的には，(**G1**) 基本パラメータ集合 $\{\boldsymbol{p},\boldsymbol{r},\boldsymbol{K}\}$，(**G2**) 競合行列 $\mathbf{A}$，(**G3**) 季節活動パターン $\{\mathbf{W},\mathbf{B}\}$ を同時に学習したい．

最もシンプルな方法として挙げられるのは，$\mathcal{S}$ 内の全てのパラメータの最適解を見つける方法である．しかしながら，これは $(3d+(d^2-d)+k(d+n_p))$ 個のパラメータを同時に推定する必要があり，現実的でない．更に，季節活動パラメータを最適化するために，k $(1 \le k \le d)$ 個の異なる解のコストを計算しなくてはいけない．

そこで本研究では，全パラメータ集合 $\mathcal{S}$ を二つの部分集合 $\{\boldsymbol{p},\boldsymbol{r},\boldsymbol{K},\mathbf{A}\}$，と $\{\mathbf{W},\mathbf{B}\}$ に分けて反復的にパラメータを推定する手法として，StepFit を提案する．アルゴリズム 1 は StepFit の処理の流れを示す．提案アルゴリズムはまず，季節活動パターンが存在しない状態（$k=0$）を仮定し，基本パラメータ集合を学習する．次に，**5.1** で示したように，自動成分分析を用いて最適な $\mathbf{B}$ と $\mathbf{W}$ を発見する．コスト関数（式 (8)）の最小化には，非線形性を有する学習に適したレーベンバーグ・マルカート (LM: Levenberg-Marquardt) 法を用いる．

しかしながら，依然として StepFit は，d 個のシーケンス集合のための基本パラメータ $\{\boldsymbol{p},\boldsymbol{r},\boldsymbol{K}\}$ に加え，サイズ $(d\times d)$ の競合行列 $\mathbf{A}$ の更新コストを必要とする．言い換えれば，StepFit は，d 個のキーワード全ての間の関係性 (競合関係) を同時に推定しなくてはならない．一般に，非線形モデルにおける多数のパラメータの同時推定は，局所解に陥りやすく最適な値の学習が非常に難しく，収束に要する計算コストも高い．そこで本研究では，全パラメータ集合 $\mathcal{S}$ を高速かつ効果的に推定するためのアルゴリズムとして EcoWeb-Fit を提案する．

EcoWeb-Fit. アルゴリズム 2 は，EcoWeb-Fit の処理の流れを示す．EcoWeb-Fit は，StepFit を拡張し，大規模なシーケンス集合 X の中から重要なパターンを高速に学習する手法である．具体的には，StepFit が X のパラメータを同時に学習するのに対し，EcoWeb-Fit はまず種間競争が存在しない状態を仮定した上で（つまり，$\mathbf{A}=\mathbf{I}_d$, $a_{ij}=0$ $(i\neq j)$）として），各キーワード x_i $(i=1,\ldots,d)$ を独立シーケンスとして扱い，各モデルパラメータ $\mathcal{S}_i=\{p_i,r_i,K_i,w_{ii},b_i\}$ を個別に学習する．ここで，$\mathcal{S}_i$ の学習には StepFit を用いる．次に，i 番目と j

Algorithm 1 StepFit (X)

Input: Co-evolving sequences X $(d\times n)$
Output: Full parameter set, i.e., $\mathcal{S}=\{\boldsymbol{p},\boldsymbol{r},\boldsymbol{K},\mathbf{A},\mathbf{W},\mathbf{B}\}$
$\mathbf{W}=\mathbf{B}=0$; /* Initialize seasonal activities($k=0$) */
while improving the parameters **do**
  /* (I) Base parameter fitting (**G1**), (**G2**) */
  $\{\boldsymbol{p},\boldsymbol{r},\boldsymbol{K},\mathbf{A}\} = \arg\min_{\boldsymbol{p}',\boldsymbol{r}',\boldsymbol{K}',\mathbf{A}'} Cost_T(X;\boldsymbol{p}',\boldsymbol{r}',\boldsymbol{K}',\mathbf{A}',\mathbf{W},\mathbf{B})$;
  /* (II) Seasonal parameter fitting (**G3**) */
  $\{\mathbf{W},\mathbf{B}\} = \arg\min_{\mathbf{W}',\mathbf{B}'} Cost_T(X;\boldsymbol{p},\boldsymbol{r},\boldsymbol{K},\mathbf{A},\mathbf{W}',\mathbf{B}')$;
end while
return $\mathcal{S}=\{\boldsymbol{p},\boldsymbol{r},\boldsymbol{K},\mathbf{A},\mathbf{W},\mathbf{B}\}$;

Algorithm 2 EcoWeb-Fit (X)

Input: Co-evolving sequences X $(d\times n)$
Output: Full parameter set, i.e., $\mathcal{S}=\{\boldsymbol{p},\boldsymbol{r},\boldsymbol{K},\mathbf{A},\mathbf{W},\mathbf{B}\}$
$\mathbf{A}=\mathbf{I}_d$; /* Initialize $\mathbf{A}$, i.e., identity matrix of size $(d\times d)$ */
/* (I) Single fitting (**G1**), (**G3**) */
for $i=1:d$ **do**
  /* Estimate individual parameters of keyword i */
  $\mathcal{S}_i = \text{StepFit}(\boldsymbol{x}_i)$;
end for
/* (II) Pair fitting (**G1**), (**G2**), (**G3**) */
while improving the parameters **do**
  /* Find the most unfitted sequence $\boldsymbol{x}_i$ */
  $i = \arg\max_{1\le i'\le d} Cost_T(\boldsymbol{x}_{i'};\mathcal{S})$;
  /* Estimate parameters of pair (i,j) */
  for $j=1:d$ **do**
    $\mathcal{S}'_{ij} = \text{StepFit}(\boldsymbol{x}_i,\boldsymbol{x}_j)$;
  end for
  /* Find the most affecting sequence $\boldsymbol{x}_j$ on $\boldsymbol{x}_i$ */
  $j = \arg\min_{j'} Cost_T(\boldsymbol{x}_i,\boldsymbol{x}_{j'};\mathcal{S}'_{ij'})$;
  /* Update best pair parameters */
  Update $\mathcal{S}_{ij}=\mathcal{S}'_{ij}$;
end while
/* (III) Full fitting (**G1**), (**G2**), (**G3**) */
$\mathcal{S} = \arg\min_{\mathcal{S}'} Cost_T(X;\mathcal{S}')$;
return $\mathcal{S}$;

番目の二つの異なるキーワードに対し，競合関係がある場合を仮定し，最適な競合ペア (x_i, x_j) として，コスト関数 $Cost_T(x_i, x_j|\mathcal{S}_{ij})$ を最小化するような全てのシーケンスペアを見つける．最後に，全体のシーケンス集合 X に対し，StepFit を用いて全パラメータ集合 $\mathcal{S}$ を最適化する．

6. 評価実験

本論文では EcoWeb の有効性を検証するため，実データを用いた実験を行った．具体的には，本章では以下の項目について検証する．

Q1　オンライン活動データに関する提案手法の有効性
Q2　提案アルゴリズムの精度の検証
Q3　パターン抽出に対する計算時間の検証

6.1 Q1: 提案手法の有効性

本節では，大規模オンライン活動データに対する EcoWeb の情報抽出の効果を検証する．本研究では，*Google* における次の五つのカテゴリーに関連するキーワード（アクティビティ）集合に対し解析を行った：ビデオゲーム (#1)，プログラミング言語 (#2)，ソーシャルメディア (#3)，アパレル企業 (#4)，リテール企業 (#5)．ここで，各データは 2004 年から 2014 年にかけての週ごとのクエリ検索数から構成される．

以下では，それぞれのカテゴリーに対する解析結果の考察を行う．

#1. ビデオゲーム. 既に図 1 において示したように，提案手法は，Xbox, PlayStation, Wii の三つのゲーム機の長期的な時間発展と，Android の出現を表現し，Black Friday やクリスマス等の季節性を伴うイベントのパターンを発見した．

#2. プログラミング言語. 図 5 (a) は，プログラミング言語に関連する三つのキーワード：“C”, “R”, “MATLAB” に対する解析結果を示している．

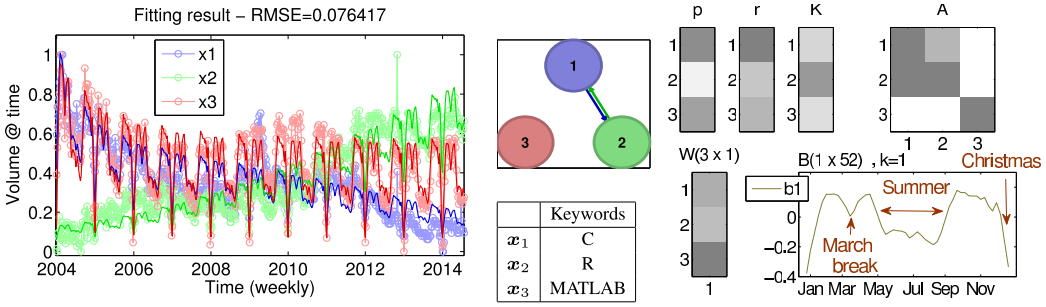
- 長期的な時間発展と競合関係: 図 5 (a-i) は，EcoWeb のモデル学習に基づく推定量（線）とオリジナルシーケンス（丸印）を示す．図に示すとおり，提案モデルはオリジナルの時系列パターンを柔軟に表現している．図 5 (a-ii) は，競合関係ネットワークを表現しており，“C” と “R” の間に潜在的な関係がある一方で，“MATLAB” は独立していることを示している．図 5 (a-i) に示すように，2004 年から現在にかけ，ビッグデータ解析ツールとしての利用等により “R” のアクティビティは増加傾向にある一方，“C” は著しく下降している．
- 季節活動パターン: 図 5 (a-iii) は，EcoWeb の全パラメータ集合を示している．ここで，濃度の高いパラメータは，高い値を示す．興味深いことに，プログラミング言語に関連する季節活動パターン (図 5 (a-iii) 下段：**W**, **B**) は，大学のアカデミックカレンダーと強い相関関係にある．明らかに，これらのクエリの入力ユーザの多くは大学生であり，プロのエンジニア，プログラマの活動パターンとは異なる．例えば，春休み，夏休み，冬休みの間，各キーワード（とりわけ MATLAB）の検索数は，著しく低下している．つまりこの結果から，大学生である彼らの大半は，コーディングの手を止め，休暇を楽しんでいるということがわかる．

#3. ソーシャルメディア. 図 5 (b) は，ソーシャルメディア関連の三つのキーワード：“Tumblr”, “Facebook” and “LinkedIn” に対する解析結果である．

- 長期的な時間発展と競合関係: 多くのソーシャルメディアサイトは，2008 年以降に成長している（$\boldsymbol{p} \approx 0$）．例えば，ブログプラットホームの Tumblr は 2007 年のサービス開始以来，多くのユーザの興味を集め，非常に高い成長率 r となっている．図 5 (b-ii) は，Tumblr と Facebook の間の競合関係を示している．
- 季節活動パターン: 図 (b-iii) 下段は，ソーシャルメディアサイトにおける季節活動パターンを表現している．クリスマスと新年において，Facebook のユーザが多い一方で，LinkedIn は反対に，ユーザの一時的な現象が見られる．これは，Facebook が個人目的でのサイトなのに対し，LinkedIn がビジネス志向の SNS であることが理由である．

#4. アパレル企業. 図 5 (c) は，四つの主要なアパレル関連企業：*Nordstrom*（高級百貨店），*Kohl's*（ディスカウントストア），*JCPenney*（大衆向け百貨店）*Forever21*（レディース専門店）の解析結果である．

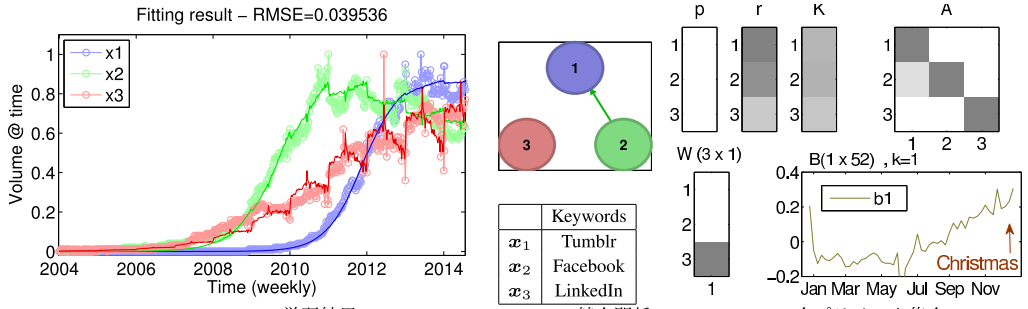
- 長期的な時間発展と競合関係: EcoWeb は，Kohl's と Nordstrom, JCPenney と Forever21 の間の潜在的な競合関係を発見した．2008 年以降の景気後退の影響により，多くの顧客が高級百貨店 Nordstrom から離れているが，一方で同時期において，ディスカウントストア Kohl's の人気が上昇した．Kohl's は，シニア向けの割引サービスや，自社クレジットカード導入等の様々な企



	Keywords
x_1	C
x_2	R
x_3	MATLAB

(a-i) EcoWeb の学習結果　(a-ii) 競合関係　(a-iii) 全パラメータ集合

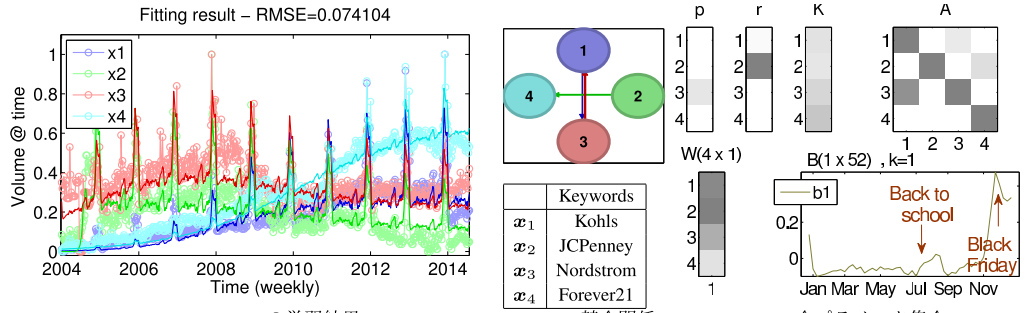
(a) プログラミング言語に関するパターン発見



	Keywords
x_1	Tumblr
x_2	Facebook
x_3	LinkedIn

(b-i) EcoWeb の学習結果　(b-ii) 競合関係　(b-iii) 全パラメータ集合

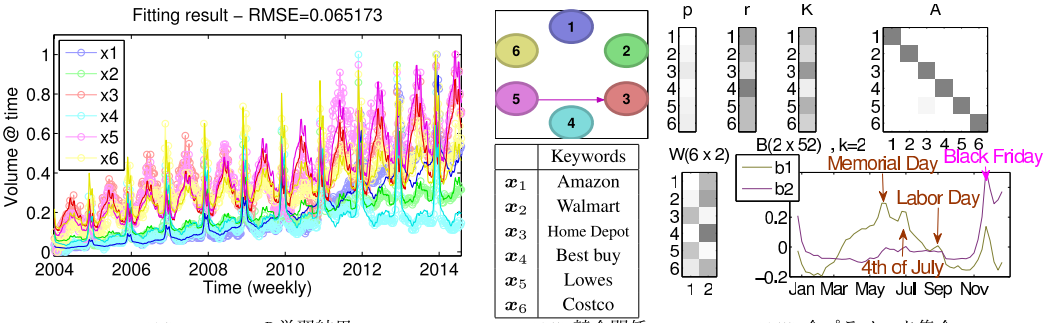
(b) ソーシャルメディアに関するパターン発見



	Keywords
x_1	Kohls
x_2	JCPenney
x_3	Nordstrom
x_4	Forever21

(c-i) EcoWeb の学習結果　(c-ii) 競合関係　(c-iii) 全パラメータ集合

(c) アパレル企業に関するパターン発見



	Keywords
x_1	Amazon
x_2	Walmart
x_3	Home Depot
x_4	Best buy
x_5	Lowes
x_6	Costco

(d-i) EcoWeb の学習結果　(d-ii) 競合関係　(d-iii) 全パラメータ集合

(d) リテール企業に関するパターン発見

図 5　四つのカテゴリーに関連するキーワードに対する EcoWeb の学習結果

Fig. 5　Fitting results of EcoWeb for four areas, i.e., (a) Programming languages, (b) social media, (c) apparel and (d) retail companies.

業戦略により，安定した成長を続けている．同様に，レディース専門店 Forever21 は，ラージサイズの拡張等を始めとする積極的な戦略により，JCPenney の顧客を引き込み，規模拡大が進んでいる．一方，JCPenney は，当時の CEO である Ron Johnson による不適切な経営判断により，著しい後退をみせた．

- 季節活動パターン: アパレル企業関連のキーワードは，共通の周期性をもつ．年間で最も大規模なスパイクが Black Friday セールであり，更に，中規模のスパイクが 8 月（“back to school” セール）にある．

#5. リテール企業. 図 5 (d) 六つの主要なリテール企業に関連するキーワードに対する解析結果である．

- 長期的な時間発展と競合関係: Amazon を始めとするオンラインサービスの成功により，主なリテール企業は，著しく成長している一方で，家電量販店 Best Buy は横ばいとなっている．日曜大工 (DIY: do it yourself) 専門店である Home Depot と Lowes の関係を除き，各キーワード間には，明らかな競合関係は存在しない．
- 季節活動パターン: 図 (d-iii) は，EcoWeb が発見した $k = 2$ 個の季節活動パターンを示している．第一成分 (b_1，黄色) は，Home Depot と Lowes のパターン，第二成分 (b_2，紫色) は，Amazon, Walmart, Best Buy, Costco のパターンを表現している．両成分における，Black Friday のスパイクに加え，Home Depot と Lowes (b_1) には，アメリカ合衆国内の祝日（Memorial Day，5 月の最終月曜日；独立記念日，7 月 4 日；労働者の日，9 月の第一月曜日）に相当するスパイクが確認できる．

6.2 Q2: 提案手法の精度

次に，提案モデルの精度を検証するため，既存手法であるロトカ・ボルテラ (*LV*) モデルとの比較を行った．本研究では更に，提案アルゴリズムの効果を検証するために，モデル学習に StepFit のみを用いた手法として EcoWeb-Plain との比較も行った．図 6 は，五つのオリジナルデータセット (#1-#5) と推定値との 2 乗平均誤差 (RMSE: root mean square error) を示している．*LV* モデルは季節性を伴う活動パターンを表現できないため，周期的に発生するスパイクに影響され，長期的な時間発展を正しく学習することができない．EcoWeb-Plain は，周期性を表現する能力

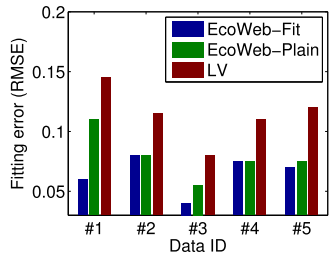


図 6 EcoWeb-Fit と既存手法の精度比較
Fig. 6 Accuracy of EcoWeb-Fit.

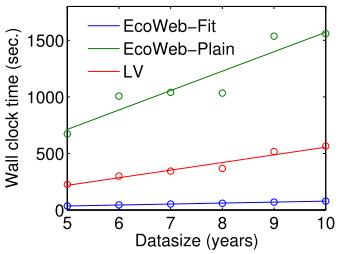


図 7 EcoWeb-Fit の計算コスト
Fig. 7 EcoWeb-Fit scales linearly.

を有するが，複雑な競合関係パラメータの学習ができない．図に示すとおり，*LV* モデルと EcoWeb-Plain と比較し，提案手法は高い精度でのデータの学習に成功した．

6.3 Q3: 提案手法の学習時間

次に，EcoWeb-Fit の計算時間を検証する．図 7 はデータのサイズ n を変化 (5 年間から 10 年間) させた上での提案手法の計算時間を示している．図に示すとおり，EcoWeb-Fit は高速かつ高精度に重要なパターンを発見することができる．ここで，EcoWeb-Fit はデータの入力サイズ n に対し線形であり，EcoWeb-Plain と比較し最大で 20 倍，*LV* と比較し 7 倍の性能向上を達成している．

7. アプリケーション

本章では，EcoWeb の最も重要なアプリケーションであるオンライン活動データの将来予測について述べる．図 8 はビデオゲーム (#1) に対する予測結果を示している．ここでは，シーケンス全体の 2/3（黒線）を学習データとして用い，その後の 2012 年以降を予測した．提案モデルの精度を検証するため，自己回帰モデル (AR: autoregressive model) との比較を行った．公平な比較をするために，AR の回帰係数は EcoWeb のパラメータ数と同様になるように設定し

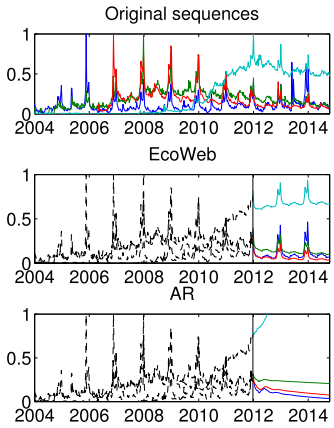


図 8　EcoWeb による将来予測 (ビデオゲーム (#1))
Fig. 8　Forecasting future evolutions.

た．図 8 上段はオリジナルデータ，中下段はそれぞれ，EcoWeb と AR の予測結果を示している．図に示したとおり，提案手法は非線形性を有する長期的なアクティビティの推移を予測するとともに，季節性を伴うスパイクも柔軟に表現することができる．

8. むすび

本論文では，大規模オンライン活動データのための非線形解析手法として EcoWeb について述べた．EcoWeb は，Web 上のオンラインユーザ活動の推移パターンを，自然界の生態系における種内・種間競争として捉えることで，潜在的なユーザ資源 (ユーザの興味) を共有，または競合することで時間発展していくような非線形動的システムとして表現する．実データを用いた実験では，EcoWeb が Web 上の様々な種類のオンライン活動において，長期的な成長や競合パターンや季節性等の重要なトレンドを発見し，更に，長期的な将来予測を高精度に行うことを確認した．今後の課題として，多種多様なキーワード間におけるより複雑な推移パターンや競合関係を抽出するための手法として，食物連鎖や進化，突然変異等をはじめとする高度な現象のモデル化について検討していく予定である．

謝辞　本研究の一部は JSPS 科研費 JP15H02705, JP16K12430, JP26280112, JP26730060, JST さきがけ及び総務省 SCOPE (受付番号 162110003) の助成を受けたものです．The authors would like to thank Christina Cowan for her help with interpreting the patterns of apparel companies. This material is based upon work supported by the National Science Foundation under Grants No. CNS-1314632 and IIS-1408924; and by the Army Research Laboratory (ARL) under Cooperative Agreement Number W911NF-09-2-0053; and by a Google Focused Research Award. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, ARL, or other funding parties. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

文　献

[1] J. Ginsberg, M. Mohebbi, R. Patel, L. Brammer, M. Smolinski, and L. Brilliant, “Detecting influenza epidemics using search engine query data,” Nature, vol.457, pp.1012–1014, 2009.
[2] J. Leskovec, L. Backstrom, and J.M. Kleinberg, “Meme-tracking and the dynamics of the news cycle,” KDD, pp.497–506, 2009.
[3] Y. Sakurai, Y. Matsubara, and C. Faloutsos, “Mining and forecasting of big time-series data,” SIGMOD, Tutorial, pp.919–922, 2015.
[4] Y. Zhao, N. Sundaresan, Z. Shen, and P.S. Yu, “Anatomy of a web-scale resale market: a data mining approach,” WWW, pp.1533–1544, 2013.
[5] Y. Matsubara, Y. Sakurai, and C. Faloutsos, “The web as a jungle: Non-linear dynamical systems for co-evolving online activities,” WWW, pp.721–731, 2015.
[6] I.N. Davidson, S. Gilpin, O.T. Carmichael, and P.B. Walker, “Network discovery via constrained tensor analysis of fmri data,” KDD, pp.194–202, 2013.
[7] Y. Sakurai, C. Faloutsos, and M. Yamamuro, “Stream monitoring under the time warping distance,” ICDE, pp.1046–1055, 2007.
[8] Y. Sakurai, S. Papadimitriou, and C. Faloutsos, “Braid: Stream mining through group lag correlations,” SIGMOD, pp.599–610, 2005.
[9] Y. Sakurai, M. Yoshikawa, and C. Faloutsos, “Ftw: Fast similarity search under the time warping distance,” PODS, pp.326–337, June 2005.
[10] A. Jain, E.Y. Chang, and Y.-F. Wang, “Adaptive stream resource management using kalman filters,” SIGMOD, pp.11–22, 2004.
[11] L. Li, B.A. Prakash, and C. Faloutsos, “Parsimonious linear fingerprinting for time series,” PVLDB, vol.3, no.1, pp.385–396, 2010.
[12] Y. Tao, C. Faloutsos, D. Papadias, and B. Liu, “Prediction and indexing of moving objects with unknown motion patterns,” SIGMOD, pp.611–622, 2004.

[13] Y. Matsubara, Y. Sakurai, C. Faloutsos, T. Iwata, and M. Yoshikawa, "Fast mining and forecasting of complex time-stamped events," KDD, pp.271–279, 2012.
[14] Y. Matsubara, Y. Sakurai, W.G. van Panhuis, and C. Faloutsos, "FUNNEL: Automatic mining of spatially coevolving epidemics," KDD, pp.105–114, 2014.
[15] Y. Matsubara and Y. Sakurai, "Regime shifts in streams: Real-time forecasting of co-evolving time sequences," KDD, pp.1045–1054, 2016.
[16] Y. Matsubara, Y. Sakurai, and C. Faloutsos, "Autoplait: Automatic mining of co-evolving time sequences," SIGMOD, pp.193–204, 2014.
[17] T. Rakthanmanon, B.J.L. Campana, A. Mueen, G.E.A.P.A. Batista, M.B. Westover, Q. Zhu, J. Zakaria, and E.J. Keogh, "Searching and mining trillions of time series subsequences under dynamic time warping," KDD, pp.262–270, 2012.
[18] J. Yang, J.J. McAuley, J. Leskovec, P. LePendu, and N. Shah, "Finding progression stages in time-evolving event sequences," WWW, pp.783–794, 2014.
[19] A. Beutel, B.A. Prakash, R. Rosenfeld, and C. Faloutsos, "Interacting viruses in networks: can both survive?," KDD, pp.426–434, 2012.
[20] M. Eirinaki and M. Vazirgiannis, "Web mining for web personalization," ACM Trans. Internet Techn., vol.3, no.1, pp.1–27, 2003.
[21] Y. Koren, "Factorization meets the neighborhood: A multifaceted collaborative filtering model," KDD, pp.426–434, 2008.
[22] R. Kumar, M. Mahdian, and M. McGlohon, Dynamics of conversations. KDD, pp.553–562, 2010.
[23] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins, "Microscopic evolution of social networks," KDD, pp.462–470, 2008.
[24] Y. Lu, P. Tsaparas, A. Ntoulas, and L. Polanyi, "Exploiting social context for review quality prediction," WWW, pp.691–700, 2010.
[25] B.A. Prakash, D. Chakrabarti, M. Faloutsos, N. Valler, and C. Faloutsos, "Threshold conditions for arbitrary cascade models on arbitrary networks," ICDM, pp.537–546, 2011.
[26] E. Shmueli, A. Kagian, Y. Koren, and R. Lempel, "Care to comment?: recommendations for commenting on news stories," WWW, pp.429–438, 2012.
[27] D. Gruhl, R. Guha, R. Kumar, J. Novak, and A. Tomkins, "The predictive power of online chatter," KDD, pp.78–87, 2005.
[28] H. Choi and H.R. Varian, "Predicting the present with google trends," The Economic Record, vol.88, no.s1, pp.2–9, 2012.
[29] S. Goel, J. Hofman, S. Lahaie, D. Pennock, and D. Watts, "Predicting consumer behavior with web search," PNAS, 2010.
[30] T. Preis, H.S. Moat, and H.E. Stanley, "Quantifying trading behavior in financial markets using google trends," Sci. Rep., 3, 04 2013.
[31] Y. Matsubara, Y. Sakurai, B.A. Prakash, L. Li, and C. Faloutsos, "Rise and fall patterns of information diffusion: model and implications," KDD, pp.6–14, 2012.
[32] F. Figueiredo, J.M. Almeida, Y. Matsubara, B. Ribeiro, and C. Faloutsos, "Revisit behavior in social media: The phoenix-r model and discoveries," PKDD, pp.386–401, 2014.
[33] B. Ribeiro, "Modeling and predicting the growth and death of membership-based websites," WWW, pp.653–664, 2014.
[34] B.A. Prakash, A. Beutel, R. Rosenfeld, and C. Faloutsos, "Winner takes all: competing viruses or ideas on fair-play networks," WWW, pp.1037–1046, 2012.
[35] R.M. May, "Qualitative stability in model ecosystems," Ecology, vol.54, no.3, pp.638–641, 1973.
[36] F. Brauer and C. Castillo-Chavez, Mathematical models in population biology and epidemiology, vol.40, Springer Verlag, New York, 2001.
[37] R.M. Anderson and R.M. May, Infectious Diseases of Humans Dynamics and Control, Oxford University Press, 1992.
[38] E. Jackson, Perspectives of Nonlinear Dynamics, Cambridge University Press, 1992.
[39] M. Nowak, Evolutionary Dynamics, Harvard University Press, 2006.
[40] A.M.D. Livera, R.J. Hyndman, and R.D. Snyder, "Forecasting time series with complex seasonal patterns using exponential smoothing," Journal of the American Statistical Association, vol.106, no.496, pp.1513–1527, 2011.
[41] S. Papadimitriou, A. Brockwell, and C. Faloutsos, "Adaptive, hands-off stream mining," VLDB, pp.560–571, 2003.
[42] E.J. Keogh, S. Chu, D. Hart, and M.J. Pazzani, "An online algorithm for segmenting time series," ICDM, pp.289–296, 2001.
[43] P. Wang, H. Wang, and W. Wang, "Finding semantics in time series," SIGMOD Conference, pp.385–396, 2011.
[44] E. Odum and G. Barrett, Fundamentals of Ecology, Thomson Brooks/Cole, 2005.
[45] J. Murray, Mathematical Biology II: Spatial Models and Biomedical Applications, Intercisciplinary Applied Mathematics: Mathematical Biology, Springer, 2003.

（平成 28 年 7 月 1 日受付，10 月 28 日再受付，
29 年 1 月 6 日早期公開）

松原　靖子　（正員）

2006 年お茶の水女子大学理学部情報科学科卒業．2009 年同大学院博士前期課程修了．2012 年京都大学大学院情報学研究科社会情報学専攻博士後期課程修了．博士（情報学）．2012 年 NTT コミュニケーション科学基礎研究所 RA．2013 年熊本大学大学院自然科学研究科日本学術振興会特別研究員（PD）．2014 年より同大学院助教．この間，カーネギーメロン大学客員研究員．2016 年 12 月より国立研究開発法人科学技術振興機構さきがけ研究員．2016 年度日本データベース学会上林奨励賞，山下記念研究賞受賞．大規模時系列データマイニングに関する研究に従事．ACM，日本データベース学会各会員．

櫻井　保志　（正員）

1991 年同志社大学工学部電気工学科卒業．1991 年日本電信電話（株）入社．1999 年奈良先端科学技術大学院大学情報科学研究科博士後期課程修了．博士（工学）．2004～2005 年カーネギーメロン大学客員研究員．2013 年熊本大学大学院自然科学研究科教授．本会平成 19 年度論文賞，情報処理学会平成 18 年度長尾真記念特別賞，平成 16 年度及び平成 19 年度論文賞，日本データベース学会上林奨励賞，ACM KDD best paper awards（2008，2010）等受賞．データマイニング，データストリーム処理，センサーデータ処理，Web 情報解析技術の研究に従事．ACM，情報処理学会，日本データベース学会各会員．

Christos　Faloutsos

カーネギーメロン大学教授．1989 年アメリカ国立科学財団 Presidential Young Investigator Award 受賞．2006 年 IEEE ICDM Research Contributions Award 受賞．2010 年 ACM SIGKDD Innovations Award 受賞．22 件の論文賞，及び 4 件の test of time award を受賞．KDD / SCS dissertation award 6 件受賞．学術論文 350 件，著書 17 件，特許 7 件，チュートリアル講演 40 件．大規模データマイニング，グラフ，時系列，テンソルデータとフラクタルデータ解析技術の研究に従事．ACM フェロー，SIGKDD executive committee.