情報処理学会論文誌　データベース　Vol.10　No.3　1–15 (Oct. 2017)

大規模オンライン活動データの特徴自動抽出

松原 靖子[1,2,a)]　櫻井 保志[1,b)]　Christos Faloutsos[3,c)]

受付日 2017年3月7日, 採録日 2017年7月3日

概要：本論文では，大規模オンライン活動データのための特徴自動抽出手法である CompCube について述べる．CompCube は，(*activity, location, time*) の三つ組で構成される様々なオンライン活動データに対し，重要な時系列パターンや外れ値を統合的に解析，要約し，将来の長期的なイベント予測を実現する．たとえば，“Nokia/Nexus/Kindle” あるいは “CNN/BBC” 等のオンライン検索キーワードの各地域（国）における 2004 年から 2015 年にかけての出現件数に関する時系列データが与えられたとき，提案手法は，(a) 基本的な非線形動的パターン，(b) 各アクティビティ間の潜在的な関連性や競合性（Nokia vs. Nexus 等），(c) クリスマスや旧正月等の各地域における季節性，(d) 単発的なイベントや外れ値等の重要なパターンを自動的に抽出する．本論文ではさらに，重要な特徴を自動的かつ高速に抽出するためのアルゴリズムとして CompCube-Fit を提案する．実データを用いた実験では，CompCube が様々なオンライン活動データの中から有用なパターンを正確に発見することを確認し，さらに，最新の既存手法と比較し提案手法が大幅な精度，性能向上を達成していることを明らかにした．

キーワード：時系列データ，非線形動的システム，特徴自動抽出，将来予測

Automatic Mining of Competing Local Activities

Yasuko Matsubara[1,2,a)]　Yasushi Sakurai[1,b)]　Christos Faloutsos[3,c)]

Received: March 7, 2017, Accepted: July 3, 2017

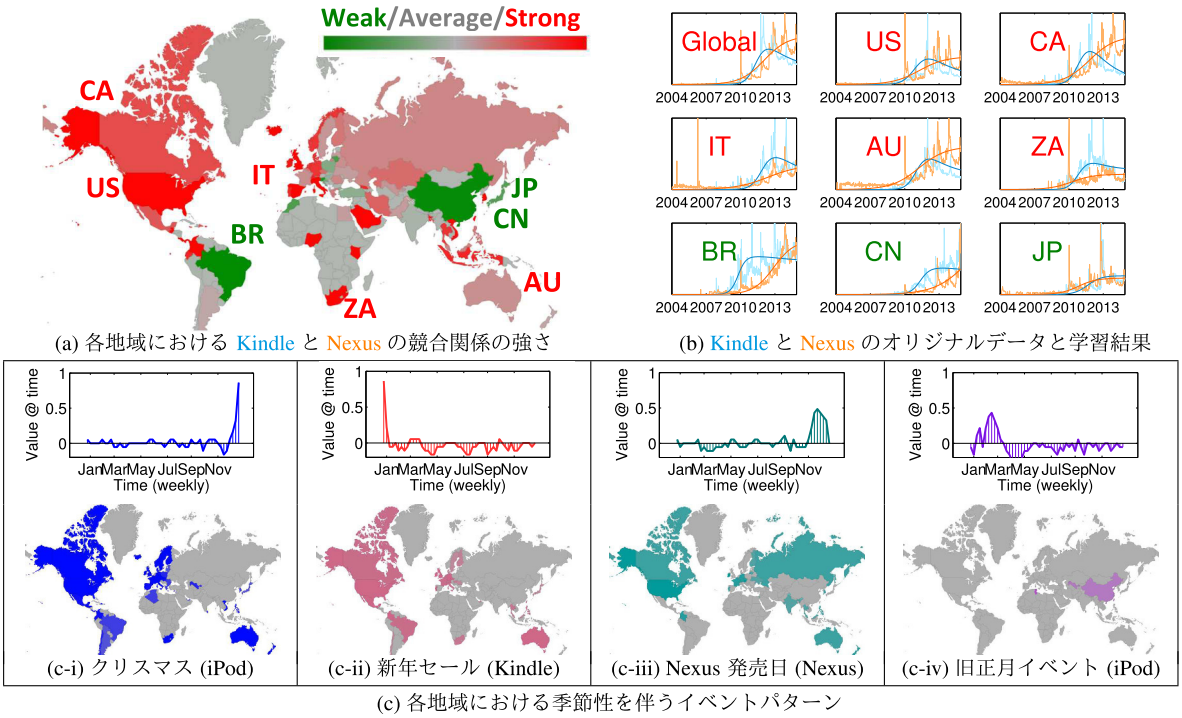
Abstract: Given a large collection of time-evolving activities, such as Google search queries, which consist of d keywords/activities for m locations of duration n, how can we analyze temporal patterns and relationships among all these activities and find location-specific trends? How do we go about capturing non-linear evolutions of local activities and forecasting future patterns? For example, assume that we have the online search volume for multiple keywords, e.g., “Nokia/Nexus/Kindle” or “CNN/BBC” for 236 countries/territories, from 2004 to 2015. We present CompCube, a unifying non-linear model, which provides a compact and powerful representation of co-evolving activities; and also a novel fitting algorithm, CompCube-Fit, which is *parameter-free* and *scalable*. Our method captures the following important patterns: (**B**)asics, i.e., non-linear dynamics of co-evolving activities, signs of (**C**)ompetition and latent interaction, e.g., Nokia vs. Nexus, (**S**)easonality, e.g., a Christmas spike for iPod in the U.S. and Europe, and (**D**)eltas, e.g., unrepeated local events such as the U.S. election in 2008. Thanks to its concise but effective summarization, CompCube can also forecast long-range future activities. Extensive experiments on real datasets demonstrate that CompCube consistently outperforms the best state-of-the-art methods in terms of both accuracy and execution speed.

Keywords: time series, non-linear, parameter free, forecasting

1. まえがき

オンラインニュース，ブログ，SNS をはじめとする様々な Web サービスの普及にともない，オンライン活動に基づく市場調査や社会行動分析に注目が集まっている．たとえば，“Kindle”，“Nexus” 等の製品にあげられるようなエレクトロニクス産業や，“CNN”，“BBC”，“Yahoo! News” 等のニュースメディア，その他の様々な企業において，グロー

[1] 熊本大学大学院先端科学研究部
Faculty of Advanced Science and Technology, Kumamoto University, Chuo, Kumamoto 860–8555, Japan
[2] 国立研究開発法人科学技術振興機構さきがけ
[3] Department of Computer Science, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213-3891, America
[a)] yasuko@cs.kumamoto-u.ac.jp
[b)] yasushi@cs.kumamoto-u.ac.jp
[c)] christos@cs.cmu.edu



(a) 各地域における Kindle と Nexus の競合関係の強さ

(b) Kindle と Nexus のオリジナルデータと学習結果

(c-i) クリスマス (iPod)　(c-ii) 新年セール (Kindle)　(c-iii) Nexus 発売日 (Nexus)　(c-iv) 旧正月イベント (iPod)

(c) 各地域における季節性を伴うイベントパターン

図 1　エレクトロニクス産業関連キーワード集合における CompCube の特徴自動抽出と出力例

Fig. 1　Modeling power of CompCube for the consumer electronics market (i.e., iPhone, Samsung Galaxy, Nexus, HTC, iPad, BlackBerry, Nokia, iMac, iPod, Kindle).

バル化にともなうシェア獲得競争が激化しており，Web 上における各国の顧客の動向分析と将来予測は非常に重要な課題である．

本研究の目的は，(*activity, location, time*) の三つ組で構成される様々なオンライン活動データに対し，重要な時系列パターンや外れ値を統合的に解析，要約し，将来のイベント予測を行うことである．たとえば，236 の地域（国）における “Nokia/Nexus/Kindle” あるいは “CNN/BBC” 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，これらの中から 3 つの要素：*activity*, *location*, *time* に関する重要なトレンドを自動的に発見したい．本論文では，大規模オンライン活動データのための特徴自動抽出手法である CompCube について述べる [22] [*1]．より具体的には以下の問題を扱う．

問題 1　d 個のアクティビティ（キーワード），m の地域（国），長さ n の期間から生成される大規模オンライン活動データ集合 $\mathcal{X} \in \mathbb{N}^{d \times m \times n}$ が与えられたとき，(a) 以下の重要なパターンを自動抽出する．

- アクティビティ間の競合関係
- 地域別の季節性パターン
- 外れ値や局所的なイベント

さらに，(b) 上記の抽出パターンに基づき高速かつ自動的に将来のイベント予測を行う．

具体例．図 1 は，エレクトロニクス産業関連キーワードに対する CompCube の出力例を示している．具体的には，*Google Search* [*2] における $d = 10$ 種のキーワード（iPhone, Kindle, Nexus 等）の $m = 236$ の地域（国）における 2004 年 1 月 1 日から現在にかけてのオンライン検索数を用いている．使用したキーワード集合の詳細については，図 3 (*#1*) *Products* において後述する．

アクティビティ間の競合関係： CompCube は与えられた d 個のキーワードの中から関連性の強い競合相手を自動的に発見することができる．たとえば，提案手法は Kindle と Nexus の間に潜在的な関連性があることを発見している．図 1 (a) は，各地域（国）における Kindle と Nexus の間の競合関係の強さを示している．赤色の地域（United States (US)，Canada (CA) 等）は強い競合関係を示し，緑色の地域（Brazil (BR)，China (CN)，Japan (JP) 等）は弱い競合関係を示す．図 1 (b) は，各国における Kindle（水色）と Nexus（橙色）の 2 つのキーワードの検索数（週ごと）を示しており，オリジナルデータは薄色，提案モデルの学習結果を濃色で示している．図のように，提案手法は各地域における成長，減衰パターンを柔軟に表現することができる．たとえば，United States (US)，Canada (CA)，Italy (IT)，Australia (AU)，South Africa (ZA) では強い競合関係があり，Nexus の人気上昇と同時期に Kindle の減衰パターンが見られる．一方で，Brazil (BR)，China (CN)，Japan (JP) のような地域では傾向が異なり，Kindle が依

[*1] http://www.cs.kumamoto-u.ac.jp/~yasuko/software.html

[*2] http://www.google.com/trends/

然として成長を続けている．これらの情報は，マーケティングにおける意思決定支援において非常に重要となる．

季節パターン：図 1 (c) は，エレクトロニクス市場における主要な周期的パターンを示している．たとえば，図 1 (c-i) はクリスマスに関連するパターンを示している．地図は各国における iPod のクリスマススパイクの強さを示しており，濃色であれば強い周期性を持ち，灰色であればその地域にはクリスマススパイクがないことを示す．地図中に見られるように，米国，欧州，豪州等のキリスト圏において強いクリスマススパイクが見られるが，一方で，中国やインド等の地域では存在しない．同様にして，図 1 (c-ii) は Kindle の新年セールを示す．図 1 (c-iii) は，Nexus の典型的な季節パターンを示しており，米国，欧州，ロシア等において，多くのユーザが冬に Nexus に興味を持つことが分かった．これはおそらく，Nexus の新作モデルが毎年 11 月ごろ発売されることが要因であると思われる．さらに，CompCube は地域特有の季節パターンを抽出することができる．たとえば，図 1 (c-iv) は，中国における iPod の旧正月イベントを示している．

本論文の貢献．本研究では，大規模オンライン活動データの特徴自動抽出手法である CompCube を提案する．CompCube は次の特長を持つ．

(1) 大規模オンライン活動データから 3 方向（activity, location, time）の重要なパターンを抽出する．
(2) 様々な種類の実データの中から，競合関係，季節性，外れ値等の様々な特徴を柔軟に表現する．さらに，長期的なイベント予測を実現する．
(3) 提案手法はパラメータ設定を必要としない．ユーザの介入を必要とせず，重要なパターンを自動的に抽出することができる．
(4) 計算コストは入力データサイズに対し線形である．

2. 関連研究

近年，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が活発化している [2], [7], [14], [15], [18], [35], [40]．文献 [24] 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 [6], [34] 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 [31] は，ネットワーク上において，2 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．FUNNEL [25] 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り，EcoWeb [21] は，Web 上のユーザ活動を生態系を用いて解析した．Gruhl ら [9] はブログ等のオンライン活動と Amazon.com における売上げの関係性に着目し，Ginsberg ら [7] 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 [5], [8], [32] 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．

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

	基礎/ツール			線形		非線形モデル				
	DWT/DFT	PARAFAC	AutoPlait	ARIMA/++	PLiF	SpikeM	LV/WTA	EcoWeb	FUNNEL	CompCube
競合関係	-	-	-	-	-	-	√	√	-	√
周期性	√	-	√	√	√	√	-	√	√	√
局所性	-	-	-	-	-	-	-	-	√	√
自動化	-	-	√	-	-	√	√	√	√	√
将来予測	-	-	-	√	√	√	-	√	√	√
外れ値	√	√	√	-	√	√	-	-	√	√

大規模時系列データの解析も非常に有用な技術として注目されている [19], [30], [36], [37], [38]．AutoPlait [20] は多次元時系列シーケンスのための特徴自動抽出手法であり，文献 [23] 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を提案した．Rakthanmanon らは文献 [33] 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いる．

関連研究と本研究の位置づけ．表 1 は，既存手法と CompCube の能力の比較である．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持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．本研究で扱うオンライン活動データはテンソルとして表現することができる．PARAFAC，Tucker をはじめとするテンソル解析手法 [13] は，与えられたテンソルデータの圧縮と 3 方向解析の能力を有するが，一方で，周期性やドメイン知識を表現せず，非線形的な動的パターンの予測能力を有していない．AutoPlait [20]，pHMM [39] は，時系列シーケンスのダイナミクスを表現し，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クスを表現することができない．

AR, LDS, SARIMA, TBATS [17], あるいはその他の関連する予測手法である AWSOM [29], PLiF [16], TriMine [23] は，すべ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 [37]．さらに，これらの手法はパラメータの設定を要する．

ロトカ・ボルテラ（LV: Lotka-Volterra）モデル [26], ロジスティック方程式（LF: logistic function）[4], SI（susceptible-infected）モデル [1], SpikeM [24], WTA [31], EcoWeb [21], FUNNEL [25] や他の非線形方程式 [11], [28] は，ドメイン

知識に基づくが，ユーザの局所的な活動パターンや周期的なパターンを表現できない．

3. 提案モデル

本章では提案モデルである CompCube について述べる．本研究で扱うデータは (*activity, location, time*) の三つ組で表現され，それぞれ，d 個のアクティビティ（キーワード），m の地域（国），長さ n の期間（1 週間単位）から構成される．このオンライン活動データは，3 階のテンソル $\mathcal{X} \in \mathbb{N}^{d \times m \times n}$ として表現することができ，$\mathcal{X}$ の要素 $x_{il}(t)$ は時刻 t において i 番目のアクティビティ/キーワード（activity）が l 番目の地域/国（location）に出現した頻度を示す．たとえば，(CNN, US, 01-01-2015; 100) の場合，アメリカ合衆国内の “CNN” というキーワードの検索/クリック回数が 2015 年 1 月 1 日に 100 件報告されたことを表す．

本論文の目的は，与えられたオンライン活動データ $\mathcal{X}$ に対し，重要なパターンを要約，抽出することである．具体的には，以下の 4 つの特徴を発見したい．

- (**B**)asics：個々のキーワードの非線形パターン（潜在的な人気度や成長率等）．
- (**C**)ompetition：異なるキーワード間の潜在的な関連性（Nokia vs. Nexus 等）．
- (**S**)easonality：ユーザ活動の周期性や季節性イベント（クリスマスや夏休み等）．
- (**D**)eltas：突発的なイベントや周期性のない外れ値等（アメリカ大統領選挙等）．

ここで，本研究において最も重要な点として，上述の特徴は，以下の 2 方向から抽出する必要がある．

- **(Global)**：世界規模でのトレンド，共通パターン
- **(Local)**：地域（国）レベル，局所的なトレンド

次節において，提案モデルの詳細について順に説明する．議論を単純化するために，ここではまず，(1) 1 つの地域（つまり $m = 1$）におけるオンライン活動データのモデル化を行い，提案手法がどのようにして基本的な成長パターンやキーワード間の競合関係を表現するかについて述べる．次に，(2) 各地域における局所的なパターンに着目し，与えられたテンソル $\mathcal{X}$ の中から上述の 4 つの特徴を表現する方法について述べ，最後に，(3) テンソル $\mathcal{X}$ をグローバル，ローカルの両方向から解析し，重要なパターンの抽出を行う方法について述べる．

3.1 単一の地域における動的パターンと競合関係

最も単純な場合として，単一の地域におけるオンライン活動のモデル化について述べる．具体的には，d 個のキーワード，長さ n，単一の地域（$m = 1$）で構成されるシーケンス集合を考える．

背景：生態系モデルにおける競合関係． d 個のシーケンス集合が与えられたとき，(a) 個々のシーケンスの非線形的な時間発展，そして (b) 異なるシーケンス間の潜在的な相互作用や競合関係を表現したい．たとえば，図 1 (b) において，Nexus の成長パターンと Kindle の減衰パターンが同時期に起きていることから，これらのキーワードがユーザの興味を取り合い競合しているようにみえる[*3]．

それでは，d 個のキーワード間の隠れた競合関係を表現するには，どうすればよいだろうか．競合関係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・ボルテラの競争モデル（LVC: Lotka-Volterra population model of competition）があげられる [27]．LVC モデルは，生態系における d 種の競争関係を表現し，i 番目の種の個体数 P_i が時間発展していく様子を，次の非線形微分方程式を用いて表現する．$\frac{dP_i}{dt} = r_i P_i \left(1 - \frac{\sum_{j=1}^{d} c_{ij} P_j}{K_i}\right)$，$(i = 1, 2, \ldots, d)$，ここで，$r_i$ は i 番目の種の成長率（$r_i \geq 0$），K_i は，i 番目の種の環境収容力（$K_i \geq 0$），c_{ij} は，競合係数，つまり，異種間の相互作用の強さ（$c_{ij} \geq 0$）を示す[*4]．

上記のモデルは時系列シーケンス集合に対し重要な非線形パターンを表現することができるが，本論文で扱うデータを表現するには不十分である．次の課題は，どのようにして地域別の（局所的な）時系列パターンを表現するか，そして，季節性をともなう周期的パターンおよび外れ値を発見するかである．次節において詳細を示す．

3.2 CompCube-dense

d のアクティビティ（キーワード），m の地域（国），長さ n のタイムスタンプで構成されたテンソル $\mathcal{X}$ が与えられたとき，次の目標は，各地域における各アクティビティに対する重要なパターンを抽出することである．具体的には，4 つの重要な特徴：(**B**)asics，(**C**)ompetition，(**S**)easonality，(**D**)eltas を同時に発見したい．以下では各項目の詳細について順に述べる．

(B)asics, (C)ompetition. テンソル $\mathcal{X}$ が与えられたとき，最初のステップは，潜在的な人気度 $P_{il}(t)$ を i 番目のアクティビティ（キーワード），l 番目の地域，時刻 t において，それぞれ推定することである．ここで人気度は，各アクティビティ，各地域における個々のユーザの興味や注目の強さを示し，時間発展していくものとする．たとえば，発売されたばかりの商品（たとえば Nexus）が魅力的だった場合，多くのユーザがこの商品に対し，より多くの時間をかけたり，友人に紹介することによって新たなユーザを

[*3] もちろん，実際の時系列シーケンスを見る限りでは，これらの 2 つの製品が実際に競合しているかどうかを判断することはできないが，このような時系列パターン上の現象を競合としてとらえ，モデル化することができる．本論文では，これらのシーケンス間の振舞いを競合関係と呼ぶ．

[*4] 本論文では，種内競争に対し一定の強さ $c_{ii} = 1$ を仮定し，また，種間競争については，競合，中立，片害作用が存在する場合（$c_{ij} \geq 0$）について議論する．

獲得していき，結果として，人気度の指数関数的な成長につながる．同様にして，本研究では，2 つの異なるアクティビティ間に潜在的な競合関係を仮定する．たとえば，たいていの場合，ユーザは個人の嗜好や価格に応じて，Nexus, iPhone，Kindle，iPad 等の様々な類似製品の中から，1 つを選んで使用する．ここでさらに強調すべき点として，これらの傾向やパターンは，地域によって差が見られる場合がある．図 1 (b) で見たように，各国は独自の傾向や時系列パターンを有し，それらはその国の習慣，教育，経済をはじめとする様々な要因によって変化する．

モデル 1　$P_{il}(t)$ をアクティビティ i の l の地域における時刻 t の潜在的な人気度とする．提案する基本モデルは次の式で表現される．

$$P_{il}(t) = P_{il}(t-1)\left[1 + r_{il}\left(1 - \frac{\sum_{j=1}^{d} c_{ijl} \cdot P_{jl}(t-1)}{K_{il}}\right)\right], \quad (i = 1, \cdots, d;\ l = 1, \cdots, m;\ t = 1, \cdots, n) \qquad (1)$$

ここで，$r_{il} > 0$，$K_{il} > 0$，$c_{iil} = 1$，$c_{ijl} \geq 0$，$P_{il}(0) = p_{il}$.

モデル 1 は，次のパラメータ集合で構成される．

- p_{il}：初期状態，つまり，アクティビティ i の l 番目の地域における時刻 $t = 0$ の人気度（$P_{il}(0) = p_{il}$）．
- r_{il}：成長率，つまり，アクティビティ i の l 番目の地域における魅力の強さ．
- K_{il}：環境収容力，つまり，アクティビティ i の l 番目の地域におけるユーザ資源の量．
- c_{ijl}：競合関係の係数，つまり，l 番目の地域において，アクティビティ j がアクティビティ i に与える影響力の強さ．

ここでは，競合するアクティビティが共通のユーザ資源を取り合う関係を仮定している．生態系における食料資源と同様に，Web 上でのユーザの総数，およびユーザ資源は有限である．具体的に，ユーザ資源とは，ユーザの興味や注目，あるいは消費した時間や金額を指す．ユーザは，同時に複数の目的に対し時間やお金を使用することができない．アクティビティ i の l 番目の地域における時刻 t の潜在的なユーザ資源の割合は $\left(1 - \frac{\sum_{j=1}^{d} c_{ijl} P_{jl}(t)}{K_{il}}\right)$ として表され，c_{ijl} は競合関係の係数，つまり，アクティビティ j がアクティビティ i に対し l 番目の地域で与える影響の強さを示す．もし $c_{ijl} = 0$（$i \neq j$）の場合，l 番目の地域において，アクティビティ i と j の相互作用は存在せず，中立関係となる．対照的に，もし $c_{ijl} = c_{jil} = 1$ の場合，l 番目の地域における 2 つのアクティビティは，まったく同じユーザ資源のグループを共有していることを意味する．もし $c_{ijl} = 1$，$c_{jil} = 0$ の場合には，一方的な競合関係，つまり片害作用を表現する．この場合，アクティビティ i がアクティビティ j に強く影響を受ける一方，アクティビティ j はほとんどアクティビティ i から干渉を受けない．

(S)easonality, (D)eltas.　続いて，季節性をともなう周期パターンと外れ値の表現について議論する．Kindle，CNN 等の各アクティビティは，つねに一定の人気（注目）を有しているが，その一方で，Web 上のユーザの振舞いはクリスマス，夏休み等，季節性や周期的なイベント，習慣によって動的に変化していく．さらに，2008 年のオバマ大統領選挙等，外部要因によって発生する単発型のイベントや外れ値も発生する．そこで本研究では，これらの現象を表現するために，2 つの新たなパラメータとして，$s_{il}(t)$：季節性，$\delta_{il}(t)$：外れ値を導入する．

モデル 2　$V_{il}(t)$ を l 番目の地域におけるアクティビティ i の時刻 t における推定値とする．提案モデルは次式で表現される．

$$V_{il}(t) = P_{il}(t)\left[1 + s_{il}(t \bmod n_p)\right] + \delta_{il}(t) \quad (i = 1, \cdots, d;\ l = 1, \cdots, m;\ t = 1, \cdots, n) \qquad (2)$$

ここで n_p は，周期の長さ（ここでは $n_p = 52$ 週）を示す．

推定値 $V_{il}(t)$ はアクティビティ i が l 番目の地域において時刻 t に何回出現したかの回数を示す．これは，潜在的な人気度 $P_{il}(t)$ と，次の 2 種の新たなパラメータから計算される．

- $s_{il}(t \bmod n_p)$：季節/周期的なトレンド，つまり，人気度 $P_{il}(t)$ と実際の推定値 $V_{il}(t)$ に対する相対値．
- $\delta_{il}(t)$：外れ値，つまり，周期性をともなわない，外部要因に基づく単発的なイベント．

もし，l 番目の地域において，時刻 t のアクティビティ i が季節性および外れ値を持たない場合（つまり，$s_{il}(t \bmod n_p) = \delta_{il}(t) = 0$）には，推定値は人気度と一致する（$V_{il}(t) = P_{il}(t)$）．

パラメータ集合 - CompCube-dense.　図 2 は，提案モデルの概要を示している．テンソル $\mathcal{X}$（図 2 (a)）が与えられたとき，提案手法ははじめに 4 つの密なテンソル（図 2 (b)）を抽出する．これを CompCube-dense と呼ぶ．

定義 1（CompCube-dense）　$\mathcal{M} = \{\mathcal{B}, \mathcal{C}, \mathcal{S}, \mathcal{D}\}$ を CompCube-dense のパラメータ集合とする．ここで，

- $\mathcal{B}$ $(d \times 3 \times m)$：個々のアクティビティの基本的な動的パターン．初期値，成長率，環境収容力から構成される（$\mathcal{B} = \{p_{il}, r_{il}, K_{il}\}_{i,l=1}^{d,m}$）．
- $\mathcal{C}$ $(d \times d \times m)$：アクティビティ i と j に対する l 番目の地域における競合関係係数（$\mathcal{C} = \{c_{ijl}\}_{i,j,l=1}^{d,d,m}$）．
- $\mathcal{S}$ $(d \times n_p \times m)$：アクティビティ i，l 番目の地域，時刻 t における季節性パターン（$\mathcal{S} = \{s_{il}(t)\}_{i,t,l=1}^{d,n_p,m}$）．
- $\mathcal{D}$ $(d \times n \times m)$：外れ値（$\mathcal{D} = \{\delta_{il}(t)\}_{i,t,l}^{d,n,m}$）．

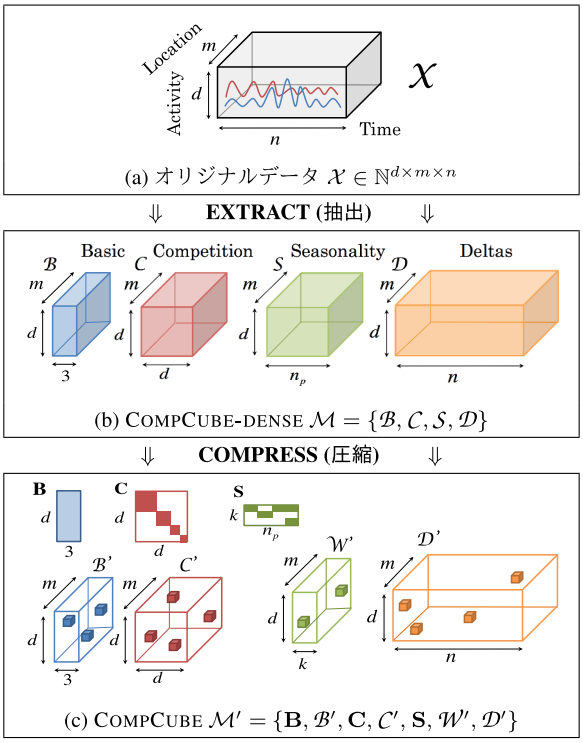


図 2　CompCube の概要
Fig. 2　Illustration of CompCube.

3.3　CompCube

ここまでの議論では，各地域における個別の時系列パターンの表現として，4 つの要素 $\mathcal{B}$，$\mathcal{C}$，$\mathcal{S}$，$\mathcal{D}$ について述べた．本論文の最終目的は，これらの 4 つの要素に対し，**(Global)**，**(Local)** の両方向に対しパターンを発見することである．つまり，与えられたデータに対し，世界規模の共通の振舞いと，特定の地域における局所的な傾向を同時に表現したい．さらに，図 2 (b) に示したとおり，CompCube-dense はすべての時系列シーケンス集合 $\mathcal{X}$ を表現するために，膨大な数の（非ゼロの）パラメータが必要となり，冗長なモデルとなってしまう．そこで本研究では，与えられたデータをシンプルかつコンパクトに表現するためのモデルとして CompCube を提案する．

情報圧縮と特徴抽出． 図 2 (c) は，提案モデルの様子を示す．与えられた CompCube-dense の密なパラメータ集合（つまり，テンソル集合：$\mathcal{B}$，$\mathcal{C}$，$\mathcal{S}$，$\mathcal{D}$）を次のようなスパースな成分に分解する．

$$\mathcal{B} \simeq \mathbf{B} \cdot 2^{\mathcal{B}'},\ \mathcal{C} \simeq \mathbf{C} \cdot 2^{\mathcal{C}'},\ \mathcal{S} \simeq \mathbf{S} \cdot \mathcal{W}',\ \mathcal{D} \simeq \mathcal{D}'. \quad (3)$$

具体的には，与えられた密なテンソル $\mathcal{B}$，$\mathcal{C}$ に対し，行列 $\mathbf{B}$，$\mathbf{C}$ とスパーステンソル $\mathcal{B}'$，$\mathcal{C}'$ に分解，圧縮する．ここで，$\mathbf{B}$，$\mathbf{C}$ は，d 種のアクティビティに対する統合的（global）なトレンドを表現し，$\mathcal{B}'$，$\mathcal{C}'$ は，局所的（local）なトレンドを示す．より具体的には，次式のように，グローバルとローカルトレンドの相対値の対数を表す：$\mathcal{B}' = \log(\mathcal{B}/\mathbf{B})$，$\mathcal{C}' = \log(\mathcal{C}/\mathbf{C})$．つまり，テンソル $\mathcal{B}'$，$\mathcal{C}'$ の各要素は，各地域（local）のトレンドが全体（global）とどのくらい異なるかを表現する．たとえば，もしすべての d 種のアクティビティ，m カ所すべての地域（国）において，$\mathcal{B}' = \mathbf{0}$ だった場合には，テンソルはグローバルのパターンと一致する（$\mathcal{B} = \mathbf{B} \cdot 2^{\mathbf{0}} = \mathbf{B}$）．同様にして，季節性パターン（$\mathcal{S}$）に対して，提案手法は季節行列 $\mathbf{S}$ と 季節テンソル $\mathcal{W}'$ に分解，圧縮する．ここで，$\mathbf{S}$ は長さ n_p の k 個の成分から構成され，各成分はクリスマスや夏休み等の個々の季節パターンを表現する．$\mathcal{W}'$ は各アクティビティ，各地域における各季節パターンの重みの強さを示す．さらに，テンソル $\mathcal{D}'$ についてもスパースにし，重要なイベントや外れ値を自動的に発見したい．これらの詳細については次章において述べる．

パラメータ集合 - CompCube. 図 2 (c) は，提案モデルの様子を示し，次の要素で構成される．

定義 2（CompCube） $\mathcal{M}'$ を CompCube のパラメータ集合とする：$\mathcal{M}' = \{\mathbf{B}, \mathcal{B}', \mathbf{C}, \mathcal{C}', \mathbf{S}, \mathcal{W}', \mathcal{D}'\}$．ここで，

- $\mathbf{B}$ $(d \times 3)$：各アクティビティのグローバルな基本トレンド（初期値，成長率，環境収容力）．
- $\mathcal{B}'$ $(d \times 3 \times m)$：各地域におけるローカルな基本トレンド．
- $\mathbf{C}$ $(d \times d)$：d 種のすべてのアクティビティ間におけるグローバルな競合関係．
- $\mathcal{C}'$ $(d \times d \times m)$：各地域におけるローカルな競合関係．
- $\mathbf{S}$ $(k \times n_p)$：長さ n_p の k 個の季節パターン．
- $\mathcal{W}'$ $(d \times k \times m)$：各地域における季節パターンの重み．
- $\mathcal{D}'$ $(d \times n \times m)$：外れ値，突発的なイベントを表現するスパーステンソル．

4.　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 CompCube-Fit について述べる．提案アルゴリズムの目的は，与えられた大規模オンライン活動データに対し，重要なパターンを自動抽出することである．

問題 2　d 個のアクティビティ，m の地域，長さ n の期間から生成されるテンソル $\mathcal{X} \in \mathbb{N}^{d \times m \times n}$ が与えられたとき，$\mathcal{X}$ を表現する最適なモデルパラメータ集合 $\mathcal{M}' = \{\mathbf{B}, \mathcal{B}', \mathbf{C}, \mathcal{C}', \mathbf{S}, \mathcal{W}', \mathcal{D}'\}$ を発見する．

4.1　モデル学習とデータ圧縮

本章の目的は，問題 2 において述べたパラメータ集合 $\mathcal{M}'$ の自動推定である．具体的には，与えられた $\mathcal{X}$ に対し，適切なモデルパラメータを学習すると同時に，季節性パターンの数 k を自動推定し，外れ値 $\mathcal{D}'$ も取り除きたい．さらに，前章で述べたように，得られたパラメータ集合 $\mathcal{M}'$ はできるだけ圧縮しシンプルかつコンパクトに表現したい．本研究では，大規模テンソル $\mathcal{X}$ を適切に表現・モデル化するために，最小記述長（MDL: minimum description

length）に基づく新たな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，より良いモデルであると見なす．

モデル表現コスト：CompCube のモデルパラメータ表現コスト $Cost_M(\mathcal{M}')$ は以下の要素から構成される．

- アクティビティの総数 d，地域の総数 m，時系列の長さ n に $\log^*(d)+\log^*(m)+\log^*(n)$ ビット要する[*5]．
- ($\underline{\mathbf{B}}$)asics：$Cost_M(\mathbf{B}) = d\cdot 3\cdot c_F$，$Cost_M(\mathcal{B}') = |\mathcal{B}'|\cdot(\log(d)+\log(3)+\log(m)+c_F)+\log^*(|\mathcal{B}'|)$
- ($\underline{\mathbf{C}}$)ompetition：$Cost_M(\mathbf{C}) = |\mathbf{C}|\cdot(\log(d)+\log(d)+c_F)+\log^*(|\mathbf{C}|)$，$Cost_M(\mathcal{C}') = |\mathcal{C}'|\cdot(\log(d)+\log(d)+\log(m)+c_F)+\log^*(|\mathcal{C}'|)$
- ($\underline{\mathbf{S}}$)easonality：$Cost_M(\mathbf{S}) = |\mathbf{S}|\cdot(\log(k)+\log(n_p)+c_F)+\log^*(|\mathbf{S}|)+\log^*(k)$，$Cost_M(\mathcal{W}') = |\mathcal{W}'|\cdot(\log(d)+\log(k)+\log(m)+c_F)+\log^*(|\mathcal{W}'|)$
- ($\underline{\mathbf{D}}$)eltas：$Cost_M(\mathcal{D}') = |\mathcal{D}'|\cdot(\log(d)+\log(n)+\log(m)+c_F)+\log^*(|\mathcal{D}'|)$

ここで，$|\cdot|$ は非ゼロ要素の総数，c_F は浮動小数点のコストを示す[*6]．

データの符号化コスト．与えられたモデルパラメータ集合 $\mathcal{M}'$ に対する $\mathcal{X}$ の符号化コストをハフマン符号を用いて次のように表現することができる [3]：$Cost_C(\mathcal{X}|\mathcal{M}') = \sum_{i,l,t=1}^{d,m,n}\log_2 p^{-1}_{Gauss(\mu,\sigma^2)}(x_{il}(t)-V_{il}(t))$，ここで，$x_{il}(t)$，$V_{il}(t)$ はそれぞれ，アクティビティ i の l 番目の地域，時刻 t におけるオリジナルデータと推定値を示す（モデル 2）[*7]．

符号化コスト関数．モデルパラメータ集合 $\mathcal{M}'$ が与えられたときの $\mathcal{X}$ の符号長は次のように表現される．

$$Cost_T(\mathcal{X};\mathcal{M}') = Cost_M(\mathcal{M}') + Cost_C(\mathcal{X}|\mathcal{M}') \quad (4)$$

したがって，本論文の次の目標は，上記のコスト関数を最小化するようなパラメータ集合 $\mathcal{M}'$ を推定することである．

4.2 提案アルゴリズム

本節では，最適なパラメータ集合 $\mathcal{M}'$ を効果的かつ効率的に推定するための手法として，CompCube-Fit を提案する．提案手法は，次のアルゴリズムから構成される．

(1) **GlobalFit**：グローバルな成長パターンと競合関係 $(\mathbf{B},\mathbf{C})$，季節パターンと外れ値 $(\mathbf{S},\mathbf{D})$ を抽出する．
(2) **LocalFit**：ローカルなトレンドおよび，季節性，局所的な外れ値や外部イベント $(\mathcal{B},\mathcal{C},\mathcal{S},\mathcal{D})$ を発見する．
(3) **AutoCompress**：$\mathcal{X}$ の特徴を圧縮，要約し，最適解 $\mathcal{M}'$ を求める．

Algorithm 1 CompCube-Fit($\mathcal{X}$)

1: **Input:** Tensor $\mathcal{X}$ $(d\times m\times n)$
2: **Output:** Full parameter set $\mathcal{M}' = \{\mathbf{B},\mathcal{B}',\mathbf{C},\mathcal{C}',\mathbf{S},\mathcal{W}',\mathcal{D}'\}$.
3: /* Parameter fitting for global-level activities */
4: $\{\mathbf{B},\mathbf{C},\mathbf{S},\mathbf{D}\}$ =GlobalFit($\mathcal{X}$);
5: /* Parameter fitting for local-level activities */
6: $\{\mathcal{B},\mathcal{C},\mathcal{S},\mathcal{D}\}$ =LocalFit($\mathcal{X},\mathbf{B},\mathbf{C}$);
7: /* Automatic model compression */
8: $\{\mathcal{B}',\mathcal{C}',\mathbf{S},\mathcal{W}',\mathcal{D}'\}$ = AutoCompress($\mathcal{X},\mathbf{B},\mathbf{C},\mathcal{B},\mathcal{C},\mathcal{S},\mathcal{D}$);
9: **return** $\mathcal{M}' = \{\mathbf{B},\mathcal{B}',\mathbf{C},\mathcal{C}',\mathbf{S},\mathcal{W}',\mathcal{D}'\}$;

Algorithm 1 に CompCube-Fit の概要を示す．与えられたテンソル $\mathcal{X}$ に対し，提案手法はまず，グローバルなパラメータ集合 $\{\mathbf{B},\mathbf{C},\mathbf{S},\mathbf{D}\}$ を推定する．次に，テンソル $\mathcal{X}$ と推定したグローバルパラメータ $\{\mathbf{B},\mathbf{C}\}$ を用いて，4 つの密なテンソル（$\{\mathcal{B},\mathcal{C},\mathcal{S},\mathcal{D}\}$）で構成されるローカルなトレンドを抽出する．最後に，コスト関数（式 (4)）に基づきパラメータ圧縮を行い，$\mathcal{M}'$ を出力する．

以下では，各アルゴリズムについて詳細を述べる．

4.2.1 GlobalFit

与えられたテンソル $\mathcal{X}$ に対し，GlobalFit はグローバルなパラメータ集合を推定する．Algorithm 2 は GlobalFit の処理の様子を示す．$\boldsymbol{X}$ を d 種のアクティビティ，m の地域，長さ n の活動値の全体（global）の平均とし，$\boldsymbol{X} = \{\bar{\boldsymbol{x}}_i\}_{i=1}^{d}$，$\bar{\boldsymbol{x}}_i = \{\frac{1}{m}\sum_{l=1}^{m}x_{il}(t)\}_{t=1}^{n}$ とする．与えられたグローバルな活動値 $\boldsymbol{X}$ に対し，GlobalFit は，各アクティビティ i に対するパラメータを反復法を用いて推定する．

ここで $\mathbf{M}_i$ をアクティビティ i に対するグローバルなパラメータ集合とする（$\mathbf{M}_i = \{\mathbf{B}_i,\mathbf{C}_i,\mathbf{S}_i,\mathbf{D}_i\}$）[*8]．GlobalFit は，まず，(I) d 種のアクティビティ間に競合関係が存在しないと仮定し（つまり，$c_{ij}=0$ $(i\neq j)$），各シーケンス $\bar{\boldsymbol{x}}_i$ $(i=1,\ldots,d)$ に対し独立かつ別々にパラメータ $\mathbf{M}_i$ を推定する．次に，(II) 2 つのアクティビティ $\bar{\boldsymbol{x}}_i$，$\bar{\boldsymbol{x}}_j$ の間に競合関係が存在すると仮定し，係数を推定する．具体的には，各アクティビティ $\bar{\boldsymbol{x}}_i$ に対し，コスト関数（式 (4)）を最小化するような競合相手 $\bar{\boldsymbol{x}}_j$ を発見する．

GlobalFit は 2 つの基本的なアイディア：(1) TetraFit，(2) SubsetCollection から構成される．

(1) TetraFit. 与えられたグローバルなシーケンス $\boldsymbol{X}$ に対し，基本的なパラメータ $\mathbf{B}$，$\mathbf{C}$ を最適化すると同時に，季節パターン $\mathbf{S}$，外れ値 $\mathbf{D}$ を取り除く過程を考える．一般に，これらのモデル成分は，非常に多くのパラメータか

[*5] ここで，$\log^*$ は整数のユニバーサル符号長を表す．
[*6] 本論文では，符号長が最適となるよう浮動小数点をデジタル化（離散化）した．ここで，$c_F = \log(b_F)$ であり，b_F はバケットの数を示し，$b_F = argmin_{b'_F} Cost_T(\mathcal{X};\mathcal{M}')$ となり，c_F の上限は 4×8 とする．
[*7] ここで，μ，σ^2 はオリジナルデータと推定値の間の距離の平均と分散を示す．
[*8] $\mathbf{B}_i = \{p_i, r_i, K_i\}$，$\mathbf{C}_i = \{c_{ij}\}_{j=1}^{d}$，$\mathbf{S}_i = \{s_i(t)\}_{t=1}^{n_p}$，$\mathbf{D}_i = \{\delta_i(t)\}_{t=1}^{n}$

Algorithm 2 GlobalFit($\mathcal{X}$)

Input: Tensor $\mathcal{X}$ $(d \times m \times n)$
Output: Global-level parameters $\mathbf{M} = \{\mathbf{B}, \mathbf{C}, \mathbf{S}, \mathbf{D}\}$;
Compute average volumes $\boldsymbol{X}$ $(d \times n)$, i.e., $\boldsymbol{X} = \{\bar{\boldsymbol{x}}_i\}_{i=1}^{d}$
Initialize parameter set $\mathbf{M}$
/* (I) Estimate individual parameters */
for $i = 1 : d$ **do**
  $\mathbf{M}_i = \text{TetraFit}(i, \bar{\boldsymbol{x}}_i, \mathbf{M})$; // Fitting for i using $\bar{\boldsymbol{x}}_i$;
end for
/* (II) Estimate competition among all d activities */
while improving the parameters **do**
  /* Select the most unfitted sequence $\bar{\boldsymbol{x}}_i$ */
  $i = \arg\max_{1 \le i' \le d} Cost_T(\bar{\boldsymbol{x}}_{i'}; \mathbf{M})$;
  /* Estimate parameter set $\mathbf{M}'_{ij}$ for each activity $\boldsymbol{x}_j$ */
  for $j = 1 : d$ **do**
    /* Find subset of sequences that have competition with i, j */
    $\boldsymbol{X}_{[i,j]} = \text{SubsetCollection}(\{\bar{\boldsymbol{x}}_i, \bar{\boldsymbol{x}}_j\}, \mathbf{C})$;
    $\mathbf{M}'_{ij} = \text{TetraFit}(i, \boldsymbol{X}_{[i,j]}, \mathbf{M})$; // Fitting with $\boldsymbol{X}_{[i,j]}$;
  end for
  /* Find the best competitor $\boldsymbol{x}_j$ of $\boldsymbol{x}_i$, and update $\mathbf{M}_i$ */
  $j = \arg\min_{1 \le j' \le d} Cost_T(\boldsymbol{X}; \mathbf{M}'_{ij'})$; $\mathbf{M}_i = \mathbf{M}'_{ij}$;
end while
return $\mathbf{M} = \{\mathbf{B}, \mathbf{C}, \mathbf{S}, \mathbf{D}\}$;

ら構成されているため，1 度にすべてのモデルパラメータの推定を行うことは困難となる．そこで本研究では，効果的かつ効率的なアルゴリズムとして，TetraFit を提案する．TetraFit は，各アクティビティ i に対するパラメータ集合 $(\mathbf{B}_i, \mathbf{C}_i, \mathbf{S}_i, \mathbf{D}_i)$ を独立かつ交互に繰り返し推定することで，高速かつ高精度にパラメータ推定するための手法である．Algorithm 3 は，TetraFit の詳細を示す．シーケンス集合 $\boldsymbol{X}$，現在のパラメータ集合 $\mathbf{M}$，そして，インデックス i が与えられたとき，提案手法はアクティビティ i に関するパラメータのみを最適化する．ここで，レーベンバーグ・マルカート（LM: Levenberg-Marquardt）法を用いてコスト関数（式 (4)）の最適化を行った．

しかしながら，上記の手法はナイーブな方法に比べ効率的である一方，依然として，コスト関数の計算に $O(d^2 n)$ の時間を要してしまう．式 (1) で示したとおり，提案手法はデータ内のすべての競合関係の組合せ（$\{c_{ij}\}_{i,j=1}^{d,d}$）の計算が必要である．ここで重要な点として，競合関係行列 $\mathbf{C}$ は通常スパースであり，たいていの要素はゼロ（$c_{ij} = 0$）となる．このため，関連性の低いアクティビティのペア (i, j) の要素は無視することができる．このアイディアに基づき，新たに以下を導入する．

Algorithm 3 TetraFit(i, $\boldsymbol{X}$, $\mathbf{M}$)

Input: (a) Index i, (b) Sequences $\boldsymbol{X}$, (c) Current parameters $\mathbf{M}$
Output: Optimal parameters for i, $\mathbf{M}_i = \{\mathbf{B}_i, \mathbf{C}_i, \mathbf{S}_i, \mathbf{D}_i\}$
while improving the parameters **do**
  /* (I) Base and competition parameter fitting, i.e., $\mathbf{B}_i, \mathbf{C}_i$ */
  $\{\mathbf{B}_i, \mathbf{C}_i\} = \arg\min_{\mathbf{B}'_i, \mathbf{C}'_i} Cost_T(\boldsymbol{X}; \mathbf{B}, \mathbf{C}, \mathbf{S}, \mathbf{D})$;
  /* (II) Seasonal parameter fitting, i.e., $\mathbf{S}_i$ */
  $\{\mathbf{S}_i\} = \arg\min_{\mathbf{S}'_i} Cost_T(\boldsymbol{X}; \mathbf{B}, \mathbf{C}, \mathbf{S}, \mathbf{D})$;
  /* (III) Find deltas, i.e., $\mathbf{D}_i$ */
  $\{\mathbf{D}_i\} = \arg\min_{\mathbf{D}'_i} Cost_T(\boldsymbol{X}; \mathbf{B}, \mathbf{C}, \mathbf{S}, \mathbf{D})$;
end while
return $\mathbf{M}_i = \{\mathbf{B}_i, \mathbf{C}_i, \mathbf{S}_i, \mathbf{D}_i\}$;

(2) SubsetCollection. より効率的にモデル学習をするための手法として，SubsetCollection を提案する．SubsetCollection は，あるアクティビティ i に対し，相互作用の存在する部分集合のみを抽出するための手法である．具体的には，TetraFit の各ステップにおいて，アクティビティ i に関するモデルパラメータを推定する際，$\boldsymbol{X}$ を用いて d 個のすべてのペア (i, j)（$j = 1, \cdots, d$）を学習する代わりに，アクティビティ i に関係する（競合している）部分集合 $\boldsymbol{X}_{[i]} \subset \boldsymbol{X}$ のみを用いて高速にモデル推定を行う．ここで，部分集合 $\boldsymbol{X}_{[i]}$ は，直接的に（あるいは間接的に）i 番目のシーケンス $\bar{\boldsymbol{x}}_i$ と関連している（つまり，$c_{ij} > 0$ となるような）j 番目のシーケンス $\bar{\boldsymbol{x}}_j$ をまとめた集合とする．ここで，$\boldsymbol{X}_{[i]}$ は再帰関数 $f(\cdot)$ を用いて求め，$\boldsymbol{X}_{[i]} = f(\bar{\boldsymbol{x}}_i)$，$f(\bar{\boldsymbol{x}}_i) = \{\bar{\boldsymbol{x}}_i \cup \bar{\boldsymbol{x}}_j \cup f(\bar{\boldsymbol{x}}_j) | {}^{\forall j} c_{ij} > 0\}$ とする．これにより，$\bar{\boldsymbol{x}}_i$ を始点として，$c_{ij} > 0$ となるようなすべての競合関係シーケンス $\bar{\boldsymbol{x}}_j$ の集合を発見する．なお，もしアクティビティ i に競合相手が存在しない場合には（つまり ${}^{\forall j} c_{ij} = 0$ $(i \neq j)$），$\boldsymbol{X}_{[i]} = \bar{\boldsymbol{x}}_i$ となる．部分集合 $\boldsymbol{X}_{[i]}$ は，オリジナル全体の集合に比べ，非常に少ない数のシーケンスから構成されるため，この手法は TetraFit におけるモデル推定の処理を飛躍的に高速化することができる．

4.2.2 LocalFit

続いて，各地域における活動パターンを抽出する手法である LocalFit について述べる．ここでは，与えられたテンソル $\mathcal{X}$ に対し，ローカルなパラメータ $\mathcal{M} = \{\mathcal{B}, \mathcal{C}, \mathcal{S}, \mathcal{D}\}$ を推定する．ここで，最適な $\mathcal{M}$ を推定するための方法について考える．最も簡単なものとしては，すべての m の地域に対し独立して，GlobalFit を単純適応する手法があげられる．この場合，各地域 l に対し個別にパラメータ推定を行う：$\mathcal{M} = \{\mathbf{M}_l\}_{l=1}^{m}$（ここで，$\mathbf{M}_l = \{\mathbf{B}_l, \mathbf{C}_l, \mathbf{S}_l, \mathbf{D}_l\}$）．

しかしながら，この方法は，すべての地域において別々

Algorithm 4 LocalFit($\mathcal{X}, \mathbf{B}, \mathbf{C}$)

Input: Tensor $\mathcal{X}$, Global parameters $\mathbf{B}, \mathbf{C}$
Output: Local-level parameters i.e., $\mathcal{M} = \{\mathcal{B}, \mathcal{C}, \mathcal{S}, \mathcal{D}\}$
/* For each i-th activity in l-th location, $\boldsymbol{x}_{il}$ */
for $l = 1 : m$ **do**
  for $i = 1 : d$ **do**
    // (I) Find subset of sequences that have competition with i
    $\boldsymbol{X}_{[il]} =$ SubsetCollection($\boldsymbol{x}_{il}, \mathbf{C}$);
    /* (II) Estimate parameters for $\boldsymbol{x}_{il}$ */
    $\mathcal{M}_{il} =$ TetraFit($i, \boldsymbol{X}_{[il]}, \{\mathbf{B}, \mathbf{C}\}$);
  end for
end for
return $\mathcal{M} = \{\mathcal{B}, \mathcal{C}, \mathcal{S}, \mathcal{D}\}$;

にすべての競合関係のペア（$d \times d$）についてモデル推定を行う必要がある．さらに，実データにおいては，いくつかの地域において，データがスパースである場合が多いため，個別のシーケンスに対する学習では適切なパラメータ推定ができない可能性が高い．同時に，本研究において重要な点として，グローバルなパターンに加え，各地域における局所的な傾向や外れ値等，相対的な差異や特徴についても自動抽出したい．たとえば，Nokia vs. Nexus のような競合関係において，世界的なトレンドと比較し，各地域がどのくらい異なるかを確認したい．このとき，各国のパターンを個別に独立に学習してしまうのではなく，全体と比較したうえでの相対的なトレンドを自動的に発見したい．

そこで本研究では，より重要なパターン発見を行うため，m 個の地域におけるグローバルな競合関係を共有化することによって，モデル学習を行う．

具体的な方法として，もし m カ所の全地域において，2 つのアクティビティ i と j の間にローカルな競合関係がない場合には，グローバルな競合関係もないものと見なす．つまり，グローバルな競合関係行列 $\mathbf{C}$ が与えられたとき，競合関係の存在しないペア i, j（$\forall i, j, c_{ij} = 0$）に対しては，ローカルな競合関係の係数を計算する必要がなく，$c_{ij} > 0$ の場合にのみ係数を計算すればよい．Algorithm 4 は LocalFit の処理の流れを示す．各アクティビティ i の l 番目の地域において，LocalFit は最適なモデルパラメータ $\mathcal{M}_{il}$ を推定する．ここで効率的なモデル学習のために，SubsetCollection，TetraFit を用いる．

4.2.3 AutoCompress

テンソル $\mathcal{X}$ と密なパラメータ集合 $\{\mathbf{B}, \mathbf{C}, \mathcal{B}, \mathcal{C}, \mathcal{S}, \mathcal{D}\}$ が与えられたとき，本研究の最終目標は，$\mathcal{X}$ を要約，表現するモデルパラメータ $\mathcal{M}'$ を自動抽出することである．大規模なテンソル $\mathcal{X}$ の中から，冗長なパターンを除去しながらも，重要なトレンドをすべて発見したい．ここでの課題は，(a) 最適な季節性パターン（$\mathbf{S}$）をどのように発見するか，そして (b) 季節性パターンの最適な個数 k，および，それぞれのテンソルの中の非ゼロ要素の最適な個数（$|\mathcal{B}'|$，$|\mathcal{C}'|$，$|\mathcal{W}'|$，$|\mathcal{D}'|$）をどのようにして推定するか，の 2 点である．

第 1 の課題については，図 2 (b) において示したように，季節テンソル $\mathcal{S}$ は長さ n_p のシーケンスが $d \times m$ 個の集合で構成され，各シーケンスが，i 番目のアクティビティ，l 番目の地域における周期パターンを（Kindle のアメリカ合衆国における 12 月のクリスマスセール等）を表現している．しかしながら，この季節テンソル $\mathcal{S}$ はこのままでは冗長であり，複数のアクティビティ，複数の地域における共通のパターンを表現することができない．そこで本研究では，与えられた季節テンソル $\mathcal{S}$ の中から最適な季節パターン $\mathbf{S}$ を発見する手法を考える．具体的には，式 (3) において示したように，元の季節テンソル $\mathcal{S}$ に対し，k 個の独立な成分から構成される長さ n_p のシーケンス集合 $\mathbf{S}$（$k \times n_p$ のサイズ）を抽出する．ここでは，復元エラー（$\mathcal{S} \simeq \mathbf{S} \cdot \mathcal{W}'$）が最小となるような k 個の独立な成分を抽出する．本研究では，独立成分分析（ICA: independent component analysis）[10] を用いた情報抽出手法を行う．ICA は，主成分分析（PCA: principal component analysis）[12] とは異なり，ガウス性を持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．

続いて第 2 の課題については，適切な個数 k の季節性パターンを自動的に推定する手法が必要となる．さらに，各テンソルについても，非ゼロ要素の最適な個数も推定しなくてはならない．そこで本研究では，式 (4) に示したコスト関数を用いて，適切な数 k を推定すると同時に，各テンソルをできる限り圧縮しスパースにする手法を提案する．たとえば，もし $\mathcal{C}'$ に含まれる要素 c_{ijl} が非常に小さく無視できる値だった場合には，コスト関数に基づき，値をゼロ（$c_{ijl} = 0$）にすることができる．

Algorithm 5 は AutoCompress の処理の流れを示す．アルゴリズムは，まず式 (3) に基づき $\mathcal{B}'$，$\mathcal{C}'$ を計算する．次に，ICA を用いて $k = 1, 2, \cdots$ 個の季節性パターン $\mathbf{S}$ と季節テンソル $\mathcal{W}'$ に分解（Decompose）し，コスト関数（式 (4)）に基づきパラメータ集合を圧縮（Sparse）する．その後，最適な季節性パターンの個数 k をコスト関数に基づき決定する．同様に，その他の各テンソル $\mathcal{B}'$，$\mathcal{C}'$，$\mathcal{D}'$ についても，コスト関数に基づき圧縮（Sparse）される．

5. 評価実験

本論文では CompCube 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　実データの特徴抽出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

Algorithm 5 AUTOCOMPRESS($\mathcal{X}, \mathbf{B}, \mathbf{C}, \mathcal{B}, \mathcal{C}, \mathcal{S}, \mathcal{D}$)

Input: Tensor $\mathcal{X}$, Dense parameters $\mathbf{B}, \mathbf{C}, \mathcal{B}, \mathcal{C}, \mathcal{S}, \mathcal{D}$
Output: Compressed parameters i.e., $\{\mathcal{B}', \mathcal{C}', \mathbf{S}, \mathcal{W}', \mathcal{D}'\}$
$\mathcal{B}' = \log(\mathcal{B}/\mathbf{B})$;　$\mathcal{C}' = \log(\mathcal{C}/\mathbf{C})$;　// Compute $\mathcal{B}', \mathcal{C}'$;
$k = 1$; // Find k seasonal components ($k = 1, 2, \cdots$);
while improving the cost **do**
　$\{\mathbf{S}, \mathcal{W}'\} = $ DECOMPOSE($\mathcal{S}, k$);
　$\{\mathbf{S}, \mathcal{W}'\} = $ SPARSE($\mathbf{S}, \mathcal{W}'$);
　$Cost_k = Cost_T(\mathcal{X}; \mathbf{S}, \mathcal{W}', k)$;
　if $Cost_k < Cost_{best}$ **then**
　　/* Update best candidate set */
　　$Cost_{best} = Cost_k$, $k_{best} = k$;
　　$\{\mathbf{S}_{best}, \mathcal{W}'_{best}\} = \{\mathbf{S}, \mathcal{W}'\}$
　end if
　$k++$;
end while
$\{\mathbf{S}, \mathcal{W}'\} = \{\mathbf{S}_{best}, \mathcal{W}'_{best}\}$
/* Compress parameters */
$\{\mathcal{B}', \mathcal{C}', \mathcal{D}'\} = $ SPARSE($\mathcal{B}', \mathcal{C}', \mathcal{D}'$);
return $\mathcal{M}' = \{\mathcal{B}', \mathcal{C}', \mathbf{S}, \mathcal{W}', \mathcal{D}'\}$;

5.1　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 COMPCUBE の情報抽出の効果を検証する．本論文では，*GoogleTrends* における次の 8 つのドメイン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：(*#1*) *Products*，(*#2*) *News sources*，(*#3*) *Beer*，(*#4*) *Cocktails*，(*#5*) *Car companies*，(*#6*) *Social media sites*，(*#7*) *Financial companies*，(*#8*) *Software*．各ドメインに対し，主要なキーワード上位 $d = 10$ 件を選び，$m = 236$ カ所の地域（国）に対する 2004 年から 2015 年にかけての週ごとのクエリ検索数を用いた．ここで，各シーケンスは，上限が 1.0 になるよう正規化処理を行っている．

図 3 は，COMPCUBE を用いた特徴自動抽出の様子を示している．与えられた 8 種類のドメインにおける長期的なオンライン活動データに対し，(**B**)asics：成長パターン，(**C**)ompetition：異なるアクティビティ間の競合による減衰パターン，(**S**)easonality：周期的な活動パターン，(**D**)eltas：突発的なスパイク（図中赤丸のイベント等）の 4 つの重要な特徴を適切に表現している．以下では，COMPCUBE によるパターン抽出結果例を 4 つの特徴に分類して順に紹介する．

(B)asics. 図 3 (a) は，8 種類のドメインから生成された各データセットに対する提案手法の学習結果を示している．オリジナルデータは薄色の線，提案モデルによる推定値は濃色の線でそれぞれ示している．COMPCUBE は，各データセットに対し，指数的な成長パターンや長期的な時間発展を正確に表現している．たとえば，図 3 (a) に示すとおり，(*#1*) *Products* は，iPhone や Nexus をはじめとする多くのキーワードが，2004 年から 2012 年にかけて飛躍的な成長をとげているが，一方で Nokia，iPod は異なる傾向を示している．これは，おそらく Android 関連の製品が出現したことが要因であると考えられる．

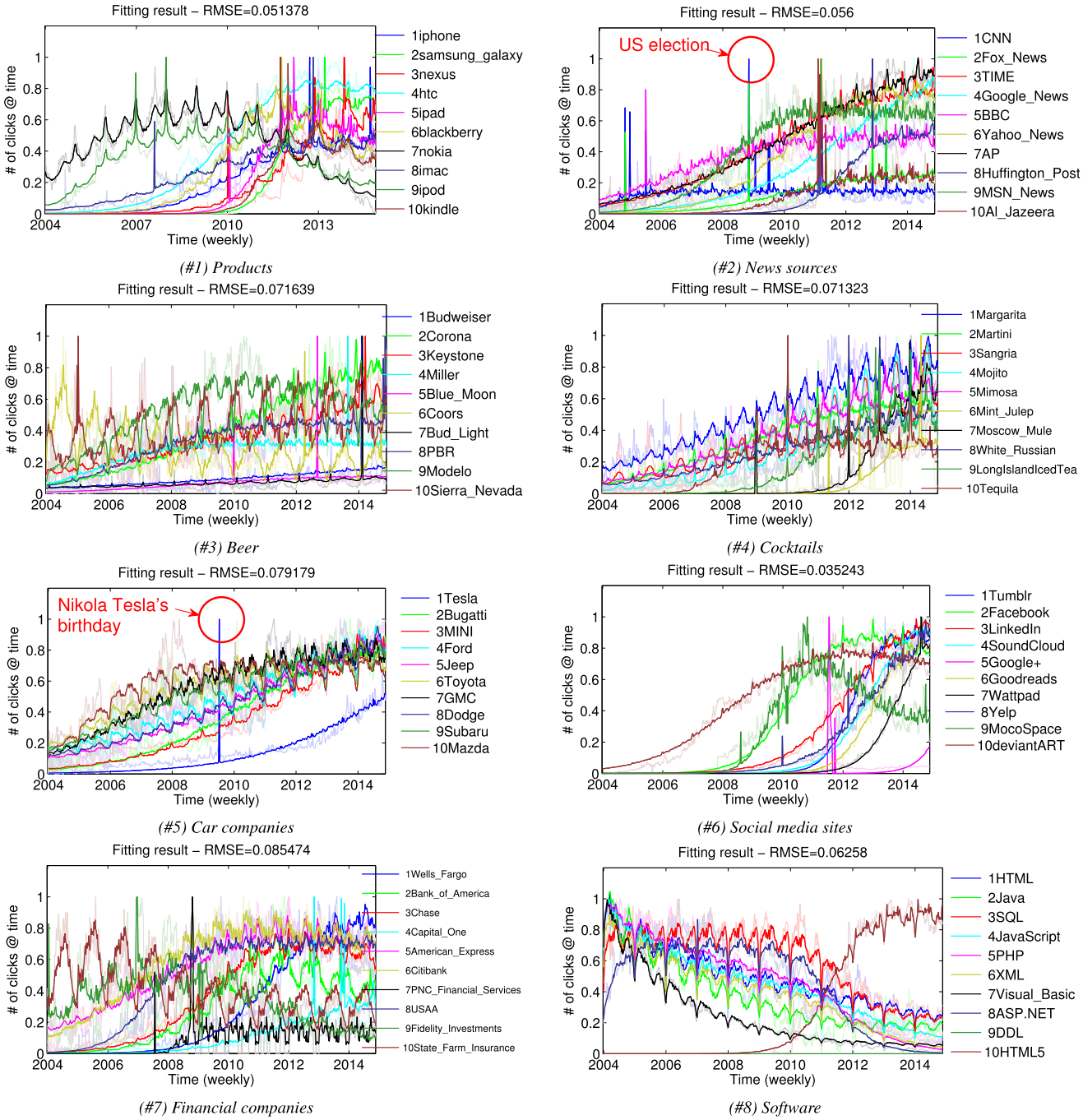
(C)ompetition. COMPCUBE は複数のキーワード間の潜在的な競合関係を自動的に発見することができる．図 3 (b) は，各データセットにおける潜在的な競合関係の例を示している．より具体的には，提案手法によって自動抽出された $\mathbf{C} = \{c_{ij}\}_{i,j}^{d,d}$ のうち，高い値を持つペアを示している．たとえば，(*#1*) *Products* において，Kindle，Nokia，Nexus の間に隠れた相互作用が含まれていることを検出した．さらに，1 章においてすでに述べたとおり，提案手法はその地域（国）特有の局所的な競合関係の強さも表現することができる．図 1 (a) は，提案手法で抽出した各地域における Kindle と Nexus の間の競合関係の強さ $\mathcal{C}'$ を示しており，赤色の地域（United States（US），Canada（CA）等）は強い競合関係，緑色の地域（Brazil（BR），China（CN），Japan（JP）等）は弱い競合関係を示している．

同様に，**図 4** (a-i), (a-ii) は，(*#3*) *Beer* における Modelo と Corona の間の地域別競合関係を示している．Modelo と Corona は *Grupo Modelo* 社が製造する著名なメキシコビールである．図を見ると，Corona が長期的に成長している一方，Modelo はユーザの興味が減少しており，特に，Mexico（MX）や Brazil（BR）においてこの傾向が顕著である．しかしながら，Guatemala（GT），Chile（CL）等の地域においては異なる傾向が見られた．

図 4 (b-i)，(b-ii) は (*#8*) *Software* における HTML と HTML5 の間の地域別の競合関係を示している．他のドメインとは異なり，これらのキーワード間には，局所的なトレンドやパターンは見られず，各地域が同様の振舞いをしている．たとえば，どの地域においても，HTML5 のアクティビティ上昇と同時に HTML の検索数が減少傾向にあることが分かる．

(S)easonality. 図 3 (a) に示すように，COMPCUBE は，すべてのデータセットにおいて，年周期のパターンを柔軟にとらえている．たとえば，(*#1*) *Products*，(*#3*) *Beer*，(*#8*) *Software* においては，クリスマスや夏休み等，様々な季節性イベントのパターンが発見されている一方で，(*#2*) *News sources* には明確な季節性が存在しないことが分かった．地域別の季節性イベントに関しては，図 1 (c) において，4 つの代表的なキーワードに対する各地域の季節性を示している．具体的には，上段が抽出された季節性パターン $\mathbf{S}$，下段が各地域における季節性の強さ $\mathcal{W}'$ を示している．図に示すように，クリスマスや新年等の広域で見られるパターンから，旧正月のような局所イベントも自動的に発見することができた．

図 4 (a-iii) は，Coors における各地域の季節性パターン



(a) 8 種のデータセットに対する CompCube の学習結果

(#1) Products	
Keyword	Competitor
iPod	Samsung Galaxy
Kindle	Nexus
iPad	Nexus
Nokia	Nexus

(#2) News sources	
MSN News	Yahoo News
(#3) Beer	
Modero	Corona
(#4) Cocktails	
Mimosa	Mojito

(#7) Financial companies	
Bank of America	Wells Fargo
Chase	Wells Fargo
(#8) Software	
HTML	HTML5
JavaScript	HTML5
PHP	HTML5

(b) 各データセットにおける主要な競合関係の例

図 3　8 種類のオンライン活動データ（$d = 10$）に対する CompCube の学習結果

Fig. 3　Fitting results of CompCube for eight datasets (here, $d = 10$ for each dataset).

を示している．Coors はアメリカ合衆国コロラド州に拠点を置くビールの銘柄である．毎年，US，Canada（CA）において，ビールの季節である夏を中心に Coors の人気が上昇していることが分かる．

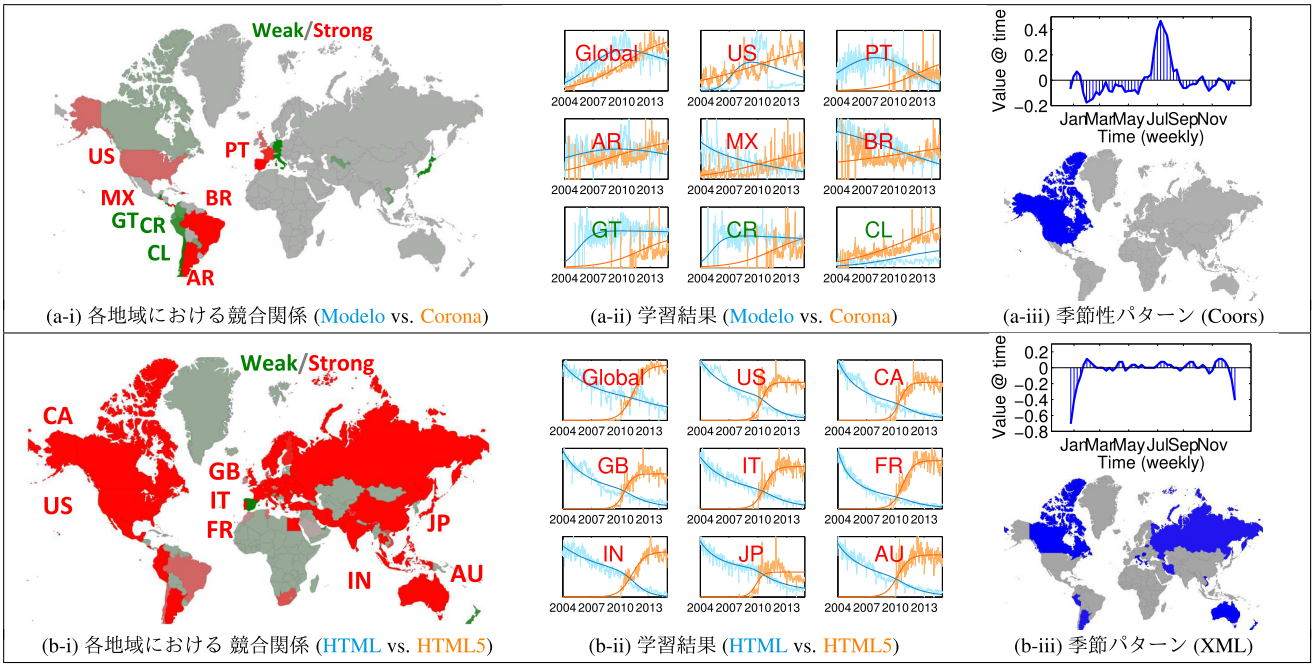


図 4　(#3) *Beer*, (#8) *Software* に対する CompCube のローカルパターンの出力例
Fig. 4　Discovery of local patterns for (#3) *Beer* and (#8) *Software*.

図 4 (b-iii) は (#8) *Software* における主要な季節パターンとして，新年休みを示している．下図は，XML における新年イベントの各地域の強さを示す．図に示すとおり，プログラマやエンジニアは，休暇中にこれらのキーワード検索をしていないことが分かる．この傾向は，Java や SQL においても見られた．

(D)eltas. 図 5 は，世界的なイベントの地域別傾向を示している．具体的には，(a) CNN における 2008 年のアメリカ大統領選挙のニュース，(b) 2009 年 7 月 10 日の Nikola Tesla の誕生日に関する Google Doodle を示す（図 3 (#2) *News sources*, (#5) *Car companies* において赤丸で表示）．図 5 では，各イベントに対し，地域ごとの影響力の強さ $\mathcal{D}'$ を示している．濃色であれば強いスパイクを持ち，灰色であればその地域にはイベントの影響がないことを意味する．図 5 (a) は，各国においてアメリカ合衆国関連の政治ニュースがどのくらい注目されているかが分かる．ここでは，US，Canada，Europe，South Africa，Japan，Korea 等，英語圏の地域や政治的，経済的に強い関わりのある国が強い関心を示していることが分かる．同様にして，図 5 (b) は，歴史上最も重要な発明家の 1 人である Nikola Tesla に対し，技術立国（US，Canada，India，Brazil 等）において強い関心が集まっている．

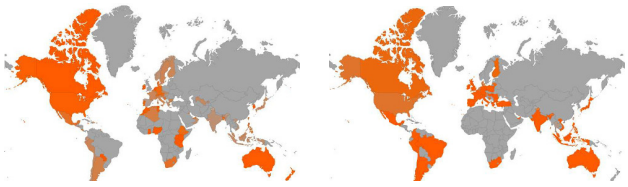

(a) CNN: Nov 2-8, 2008 (US election)　(b) Tesla: Jul 5-11, 2009 (Google Doodle)

図 5　CompCube における外部イベント（(D)eltas）抽出例
Fig. 5　CompCube automatically detects world-wide events.

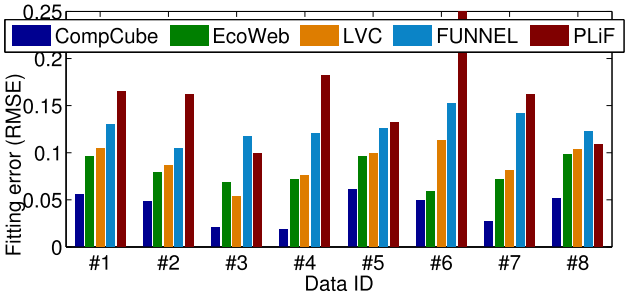


図 6　CompCube と既存手法の精度比較（RMSE）
Fig. 6　Average fitting error (RMSE) between original and fitted volumes.

5.2　Q2：提案手法の精度

次に，CompCube の学習精度を検証するため，以下の技術との比較を行った．(a) PLiF [16]：時系列シーケンスのための線形動的システム，(b) FUNNEL [25]：疫病データのための非線形モデリング，(c) LVC [26]：ロトカ・ボルテラの競争モデル，(d) EcoWeb [21]：単一地域におけるオンライン活動の競争モデル．**図 6** は 8 つのデータセット (#1–#8) におけるオリジナルデータと推定値との二乗平均誤差（RMSE: root mean square error）を示している．より低い値はより良い学習精度を示す．図に示すとおり，提案手法は高い精度を実現している一方で，(a) PLiF は線形モデルであり，非線形的な時間発展を表現できない．

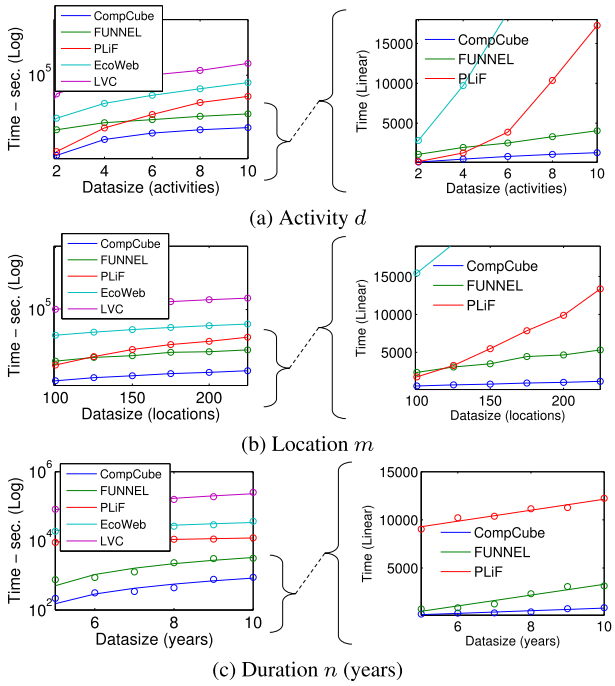


図 7　CompCube の計算コスト

Fig. 7　Wall clock time vs. dataset size, i.e., (a) activity d, (b) location m, (c) duration n.

(b) FUNNEL は非線形的な成長パターンや外部ショックによるスパイクを表現するが，異なるシーケンス間の競合関係をとらえることができない．(c) LVC は，異なるキーワード間の競合関係を表現できるが，季節性パターンを抽出できず，(d) EcoWeb は，競合関係と季節性パターンは表現できるが，外れ値や各地域における局所的な傾向を発見できない．図に示すとおり，既存手法と比較し，提案手法は高い精度でのデータの学習に成功した．

5.3　Q3：提案手法の学習時間

最後に，CompCube の計算時間を検証する．**図 7** はデータのサイズを変化させたうえでの提案手法の計算時間を示している．ここでは，それぞれ，(a) アクティビティの総数 d，(b) 地域の総数 m，(c) シーケンスの長さ n（年単位）を変化させている．左図は linear-log，右図は linear-linear のスケールで示している．PLiF については，$k = 5$ 個の隠れ値を設定し，$iter = 20$ とし，入力テンソルを $d \times m$ 個のシーケンス集合として扱った．図に示すとおり，CompCube はデータの入力に対し線形の計算時間で重要なパターンを発見することができる．結論として，提案手法は計算コストと学習精度の両面において，既存手法と比較し大幅な能力の向上を達成した．

6.　ディスカッション

前章で述べたように，CompCube は様々なドメインにおけるオンライン活動データを多方面から柔軟に表現，解析することができる．そこで本章では，本研究の最も重要な

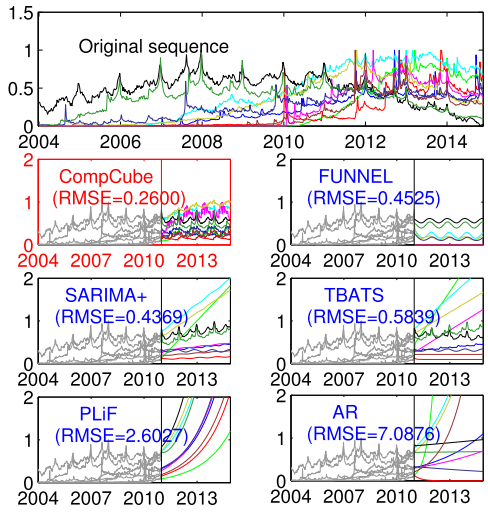


図 8　(#1) *Products* における CompCube の将来予測の例

Fig. 8　Forecasting power of CompCube for (#1) *Products*.

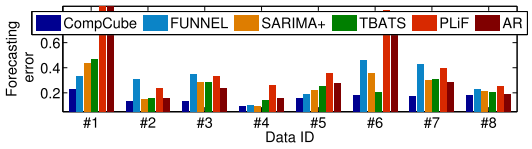


図 9　CompCube による将来予測の精度

Fig. 9　Forecasting error of CompCube.

アプリケーションとして，オンライン活動分析に基づく将来イベント予測について紹介する．ここでは，CompCube の予測能力について，以下の既存の予測技術と比較を行う：(a) FUNNEL [25], (b) SARIMA+, (c) TBATS [17], (d) PLiF [16] ($k = 5$), (e) AR．ここで，SARIMA+について周期性を CompCube と同様の $n_p = 52$ とした．パラメータの数は AIC を用いて決定した．TBATS についても同様に $n_p = 52$ とし，AR の係数を $p = 5$ とした．**図 8** は，(#1) *Products* における CompCube の将来予測の例を示している．本実験では，データセット全体の 2/3 の長さ（2004 年～2010 年，図中灰色線）を用いてモデルの学習を行い，その後の 1/3 の期間（2011 年～，図中色線）について予測を行った．2 章で議論したように，SARIMA+，TBATS，PLiF，AR は線形モデルに基づく予測技術であるため，非線形性を有する時系列パターンの表現には不十分であり，予測結果が発散している様子が見られる．FUNNEL は，非線形モデルに基づく予測手法であり，単純な周期性なら表現できるが，その一方で，複数のシーケンス間の相互関係が表現できないため，長期的な成長，減衰パターンが予測できない．

図 9 は，主要な地域（国）における予測精度の比較としてオリジナルデータと予測値の間の誤差（RMSE）を示している．低い値はより良い予測精度を意味する．図のとおり，提案手法による予測は，すべてのデータセットにおいて高い予測精度を有していることが分かる．

7. 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特徴自動抽出手法としてCompCubeについて述べた．CompCubeは，大規模なオンライン活動データの中から，競合関係，地域性，季節性や外れ値等の重要なパターンを自動的に抽出し，長期的な将来予測の能力を有する．様々なドメインのオンライン活動データを用いて実験を行い，CompCubeが最新の解析手法と比べてより高い精度と性能を持つことを示した．今後の課題として，多種多様なオンラインアクティビティの間のより複雑な推移パターンの学習と予測を行うための手法として，捕食，共生関係や食物連鎖をはじめとする高度な相互作用を持つ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, JP17H04681, JP16K12430, JSTさきがけ, 総務省SCOPE（受付番号162110003）及び国立研究開発法人日本医療研究開発機構（AMED）の臨床研究等ICT基盤構築研究事業の助成を受けたものです．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.

参考文献

[1] Anderson, R.M. and May, R.M.: *Infectious Diseases of Humans Dynamics and Control*, Oxford University Press (1992).
[2] Armenatzoglou, N., Pham, H., Ntranos, V., Papadias, D. and Shahabi, C.: Real-time multi-criteria social graph partitioning: A game theoretic approach, *SIGMOD*, pp.1617–1628 (2015).
[3] Böhm, C., Faloutsos, C., Pan, J.-Y. and Plant, C.: RIC: Parameter-free noise-robust clustering, *TKDD*, Vol.1, No.3 (2007).
[4] Brauer, F. and Castillo-Chavez, C.: *Mathematical models in population biology and epidemiology*, Vol.40, Springer Verlag, New York (2001).
[5] Choi, H. and Varian, H.R.: Predicting the present with google trends, *The Economic Record*, Vol.88, No.s1, pp.2–9 (2012).
[6] Figueiredo, F., Almeida, J.M., Matsubara, Y., Ribeiro, B. and Faloutsos, C.: Revisit behavior in social media: The phoenix-r model and discoveries, *PKDD*, pp.386–401 (2014).
[7] Ginsberg, J., Mohebbi, M., Patel, R., Brammer, L., Smolinski, M. and Brilliant, L.: Detecting influenza epidemics using search engine query data, *Nature*, Vol.457, pp.1012–1014 (2009).
[8] Goel, S., Hofman, J., Lahaie, S., Pennock, D. and Watts, D.: Predicting consumer behavior with web search, *PNAS* (2010).
[9] Gruhl, D., Guha, R., Kumar, R., Novak, J. and Tomkins, A.: The predictive power of online chatter, *KDD*, pp.78–87 (2005).
[10] Hyvärinen, A. and Oja, E.: Independent component analysis: Algorithms and applications, *Neural Netw.*, Vol.13, No.4-5, pp.411–430 (2000).
[11] Jackson, E.: *Perspectives of Nonlinear Dynamics*, Cambridge University Press (1992).
[12] Jolliffe, I.: *Principal Component Analysis*, Springer Verlag (1986).
[13] Kolda, T.G. and Bader, B.W.: Tensor decompositions and applications, *SIAM Review*, Vol.51, No.3, pp.455–500 (2009).
[14] Kumar, R., Mahdian, M. and McGlohon, M.: Dynamics of conversations, *KDD*, pp.553–562 (2010).
[15] Leskovec, J., Backstrom, L., Kumar, R. and Tomkins, A.: Microscopic evolution of social networks, *KDD*, pp.462–470 (2008).
[16] Li, L., Prakash, B.A. and Faloutsos, C.: Parsimonious linear fingerprinting for time series, *PVLDB*, Vol.3, No.1, pp.385–396 (2010).
[17] Livera, A.M.D., Hyndman, R.J. and Snyder, R.D.: Forecasting time series with complex seasonal patterns using exponential smoothing, *Journal of the American Statistical Association*, Vol.106, No.496, pp.1513–1527 (2011).
[18] Mathioudakis, M., Koudas, N. and Marbach, P.: Early online identification of attention gathering items in social media, *WSDM*, pp.301–310 (2010).
[19] Matsubara, Y. and Sakurai, Y.: Regime shifts in streams: Real-time forecasting of co-evolving time sequences, *KDD*, pp.1045–1054 (2016).
[20] Matsubara, Y., Sakurai, Y. and Faloutsos, C.: Autoplait: Automatic mining of co-evolving time sequences, *SIGMOD* (2014).
[21] Matsubara, Y., Sakurai, Y. and Faloutsos, C.: The web as a jungle: Non-linear dynamical systems for co-evolving online activities, *WWW* (2015).
[22] Matsubara, Y., Sakurai, Y. and Faloutsos, C.: Non-linear mining of competing local activities, *WWW*, pp.737–747 (2016).
[23] Matsubara, Y., Sakurai, Y., Faloutsos, C., Iwata, T. and Yoshikawa, M.: Fast mining and forecasting of complex time-stamped events, *KDD*, pp.271–279 (2012).
[24] Matsubara, Y., Sakurai, Y., Prakash, B.A., Li, L. and Faloutsos, C.: Rise and fall patterns of information diffusion: Model and implications, *KDD*, pp.6–14 (2012).
[25] Matsubara, Y., Sakurai, Y., van Panhuis, W.G. and Faloutsos, C.: FUNNEL: Automatic mining of spatially coevolving epidemics, *KDD*, pp.105–114 (2014).
[26] May, R.M.: Qualitative stability in model ecosystems, *Ecology*, Vol.54, No.3, pp.638–641 (1973).
[27] Murray, J.: *Mathematical Biology II: Spatial Models and Biomedical Applications*, Interdisciplinary Applied Mathematics: Mathematical Biology, Springer (2003).
[28] Nowak, M.: *Evolutionary Dynamics*, Harvard University Press (2006).
[29] Papadimitriou, S., Brockwell, A. and Faloutsos, C.: Adaptive, hands-off stream mining, *VLDB*, pp.560–571 (2003).

[30] Papadimitriou, S. and Yu, P.S.: Optimal multi-scale patterns in time series streams, *SIGMOD*, pp.647–658 (2006).
[31] Prakash, B.A., Beutel, A., Rosenfeld, R. and Faloutsos, C.: Winner takes all: Competing viruses or ideas on fair-play networks, *WWW*, pp.1037–1046 (2012).
[32] Preis, T., Moat, H.S. and Stanley, H.E.: Quantifying trading behavior in financial markets using google trends, *Sci. Rep.*, Vol.3, No.4 (2013).
[33] Rakthanmanon, T., Campana, B.J.L., Mueen, A., Batista, G.E.A.P.A., Westover, M.B., Zhu, Q., Zakaria, J. and Keogh, E.J.: Searching and mining trillions of time series subsequences under dynamic time warping, *KDD*, pp.262–270 (2012).
[34] Ribeiro, B.: Modeling and predicting the growth and death of membership-based websites, *WWW*, pp.653–664 (2014).
[35] Sakaki, T., Okazaki, M. and Matsuo, Y.: Earthquake shakes twitter users: Real-time event detection by social sensors, *Proc. 19th International Conference on World Wide Web*, *WWW '10*, pp.851–860 (2010).
[36] Sakurai, Y., Faloutsos, C. and Yamamuro, M.: Stream monitoring under the time warping distance, *ICDE*, pp.1046–1055 (Apr. 2007).
[37] Sakurai, Y., Matsubara, Y., and Faloutsos, C.: Mining and forecasting of big time-series data, *SIGMOD, Tutorial*, pp.919–922 (2015).
[38] Sakurai, Y., Yoshikawa, M. and Faloutsos, C.: FTW: Fast similarity search under the time warping distance, *PODS*, Baltimore, Maryland, pp.326–337 (June 2005).
[39] Wang, P., Wang, H. and Wang, W.: Finding semantics in time series, *SIGMOD Conference*, pp.385–396 (2011).
[40] Zhu, L., Galstyan, A., Cheng, J. and Lerman, K.: Tripartite graph clustering for dynamic sentiment analysis on social media, *SIGMOD*, pp.1531–1542 (2014).

松原 靖子　（正会員）

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，電子情報通信学会，日本データベース学会各会員．

櫻井 保志　（正会員）

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 18 年度長尾真記念特別賞，平成 16 年度および平成 19 年度論文賞，電子情報通信学会平成 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 年）等受賞．データマイニング，データストリーム処理，センサデータ処理，Web 情報解析技術の研究に従事．ACM，電子情報通信学会，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．24 件の論文賞，および 5 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.

（担当編集委員　北本 朝展）