

# 多次元複合データ分析から時空間多次元集合データ解析技術へ

本稿ではWeb上のデータやアプリログ、IoT（Internet of Things）センサデータなど、複数の属性を含む多種多様なデータから、データ間に現れる有意義な横断的特徴を効率的に抽出することのできる多次元複合データ分析技術について紹介します。基本的なアルゴリズムの仕組みと口コミサイトデータへの適用事例について紹介するとともに、時空間の関係性をモデル化し、事象の発生個所と時期を予測する時空間多次元集合データ解析技術への展開について紹介します。

な や ふとし†1 さ わ だ ひろし†2  
納谷 太 / 澤田 宏

NTTコミュニケーション科学基礎研究所<sup>†1</sup>  
NTTサービスエボリューション研究所<sup>†2</sup>

## 多種多様なデータを組み合わせた分析

ECサイト、スマートフォンアプリ、IoT（Internet of Things）デバイスなどの普及に伴い、ヒトの行動やモノの動きなどに付随する多種多様なデータ

が蓄積されています。例えば、ECサイトでの購買ログには、いつ、どこで、どのような年代、性別の顧客が、どのような商品を購入したのかといったデータが含まれており、膨大に蓄積されたこれらのデータから特徴的な顧客層や商品群を抽出する分析が行われて

います（図1（a））。また、最近のスマートフォンアプリの中には、ユーザの許諾を得たうえで、アプリを利用している際の位置情報をGPSやWi-Fi、ビーコンなどで取得するものもあり、購買ログと組み合わせることにより、外出先なのか自宅なのかなど、どのような

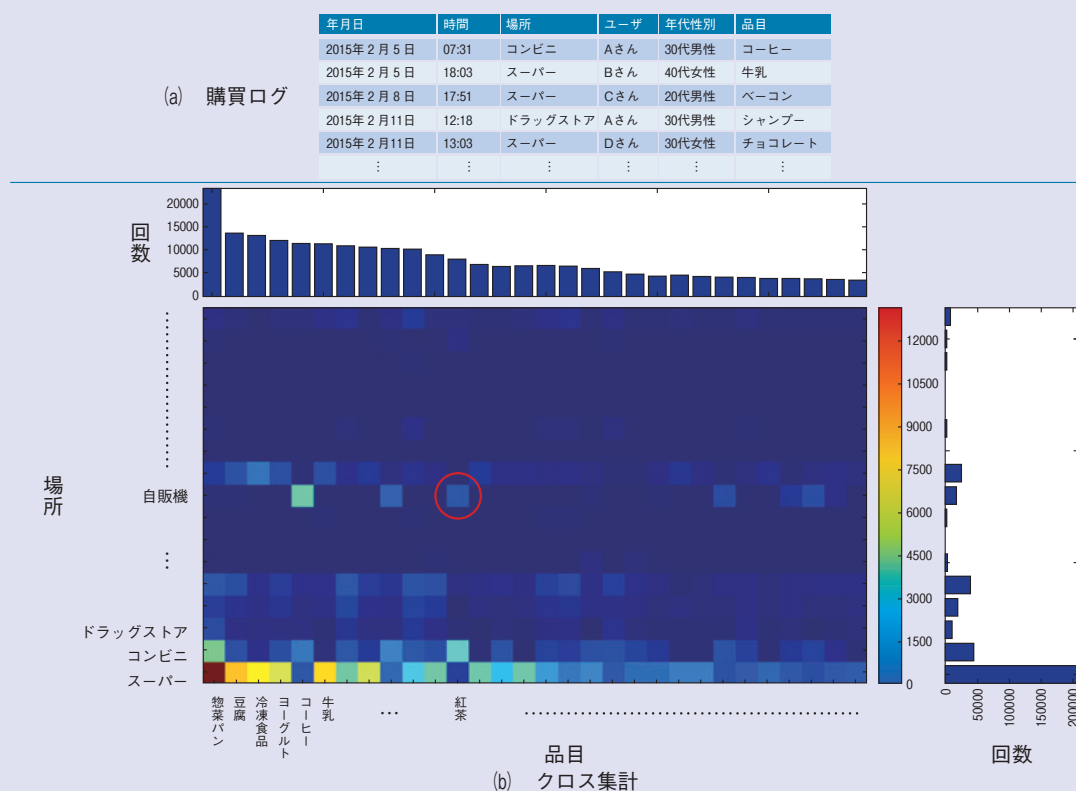
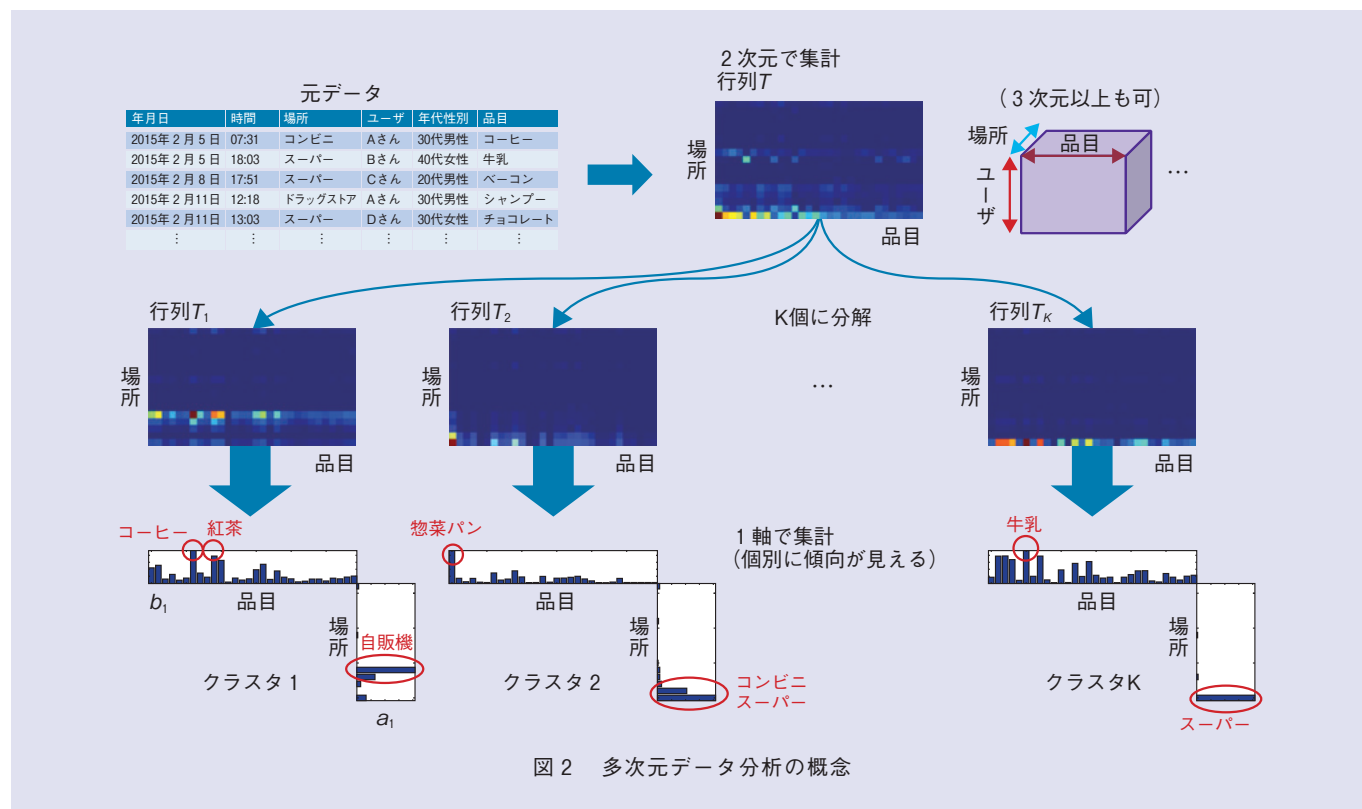


図1 購買ログデータのクロス集計例



状況で購買がなされるのかといった詳細な分析も可能になりつつあります。

従来、このような顧客層や商品群の特徴を抽出する手法として、クロス集計が活用されています。クロス集計では、例えば、場所と品目の各組合せで売上個数が集計され、その結果は表形式、もしくは行列で表現できます(図1(b))。店舗の種類や、商品の品目などを「属性」または「軸」と呼び、図1の場合は、場所と品目の2種類の属性、すなわち2次元の軸を持つデータといえます。また、各属性はさまざまな値(場所ならばスーパーやコンビニ、品目ならコーヒーなど)を持ちます。集計結果からは、ある特定の場所に注目してどのような品目の売上個数が多いのか、または、ある特定の品目がどのような場所で売れているのかなど、単一の属性値での傾向を把握することはできます。しかし、属性の取得値が増えるほど結果は複雑にな

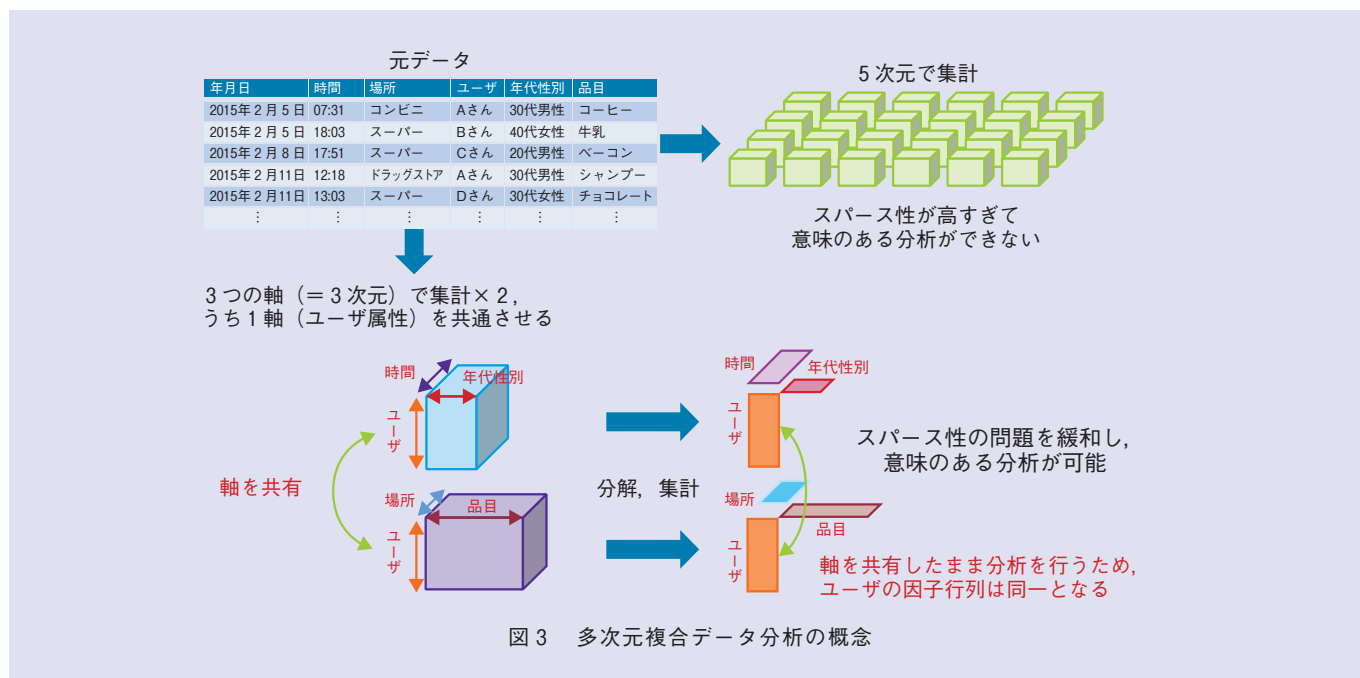
り、集計数の多い組合せのみにデータが集中するため、例えば図1(b)の赤丸で示すような、「紅茶は自販機でも比較的売上が多い」といった特徴が抽出しにくくなってしまいます。

### 多次元データ分析技術

前述の問題を解決する手法の1つが多次元データ分析技術です。多次元データ分析技術では、2次元の集計データである行列をK個の行列の和となるように分解します(図2)。K個に分解された個々の行列 $T_1 \sim T_K$ のデータをそれぞれクラスタと呼びます。多次元データ分析技術によって分解された個々のクラスタは、特定の場所や特定の品名に集計値が偏るデータになるため、例えばクラスタ1では自販機で良く売れている品目が抽出され、クラスタ2では惣菜パンの売上はスーパーが大半でコンビニがその半数程度、クラスタKではスーパーの売上

では牛乳がトップなど、各クラスタでどのような傾向があるのかがより把握しやすくなります。

ここで行列の分解の仕方がポイントになります。元の集計データを $I$ 行 $J$ 列の行列 $T$ とし、これを $T_1 \sim T_K$ のK個の行列の和で近似できるように分解とします( $T \cong T_1 + T_2 + \dots + T_K$ )。上記例では、 $I$ は場所の数、 $J$ は品目の数です。このとき、分解した個々の $T_i$  ( $i = 1 \dots K$ )も $I$ 行 $J$ 列の行列です。このような分解の仕方は任意性がありますが、多次元データ分析では、分解後の各行列 $T_i$  ( $i = 1 \dots K$ )を、 $I$ 行1列の列ベクトル $a_i$ と1行 $J$ 列の行ベクトル $b_i$ の積であるように分解します。すなわち、 $T_i = a_i \times b_i$ 。ただし、各ベクトルの要素は非負(0以上の値)を取るような制約を課します。すると、 $T_1 + T_2 + \dots + T_K = a_1 \times b_1 + a_2 \times b_2 + \dots + a_K \times b_K = (a_1 \ a_2 \ \dots \ a_K) \times (b_1 \ b_2 \ \dots \ b_K) = A \times B$ のように表現できます。これは、元の



$I$ 行 $J$ 列の行列 $T$ を,  $I$ 行 $K$ 列の行列 $A$ および $K$ 行 $J$ 列の行列 $B$ の積で近似することにほかなりません ( $T \approx A \times B$ )。ここで,  $K$ は $I$ および $J$ よりも非常に小さく取ります ( $K \ll I, J$ )。ここで $A, B$ を因子行列と呼びます。このような因子行列を求めるには,  $A$ と $B$ の因子行列に最初ランダムな値を設定し,  $A \times B = T'$ として再構成した行列 $T'$ と, 元の行列 $T$ との誤差が少なくなるような基準に基づき, 因子行列 $A$ と $B$ を更新します。上記の作業を誤差が収束するまで繰り返すことにより最終的な因子行列 $A, B$ を得ます。上述の分解例では $A$ が「場所に相当する因子」,  $B$ が「品目に相当する因子」として抽出されます。上記手法は, 機械学習技術の1つで非負値行列因子分解法 (NMF: Non-negative Matrix Factorization) と呼ばれ, データマイニングの分野で広く活用されています<sup>(1)</sup>。

また, 上記の考え方は2次元の行列データだけでなく, 属性数を3つ以上で集計した高次元のデータ (テンソル) にも適用可能であり, その手法は非負

値テンソル因子分解法 (NTF: Non-negative Tensor Factorization) と呼ばれます。

### 多次元「複合」データ分析技術への拡張

NTFは, さまざまな属性の組合せで集計した行列 (テンソル) データを, 因子行列の積として分解することにより, データに潜む特徴的なクラスタを抽出することができます。しかし, 各属性の「値」のバリエーションが多いほど, 集計データには偏りが生じやすくなります。また, 場所と品目に加えて, ユーザの年代という新たな属性を加えて集計したデータを分解することにより, どのようなユーザ層がどのような場所でどのような品目を購入する傾向があるのかといった顧客クラスタを抽出することもできますが, このような属性の数を増やせば増やすほど, その組合せ数が爆発し, ほとんどの組合せで集計値が0となる疎な (スパースな) データとなります。図1に示す購買ログを, 「時間 (1時間ごとの24

通り)」×「場所 (100通り)」×「ユーザ (1000人)」×「品名 (100種類)」の4つの属性 (軸) と値の組合せで集計する場合, 組合せ数は  $24 \times 100 \times 1000 \times 100 = 2$  億4000万通りとなり, データの種類にもよりますが, 4つ以上の属性の組合せでは, 実に99.9%以上の要素が0となります。つまり, 集計データの中で0でない要素は全体の0.1%に満たず, このようなスパースな集計データを因子分解しても満足のいく分析結果を得ることが難しくなります。

NTT研究所では, このスパース性の問題を解決する手法として, 複合非負値テンソル因子分解法 (NMTF: Non-negative Multiple Tensor Factorization) を開発しました。この手法は, 元データを高次元 (ここでは5次元) のテンソルとして集計する代わりに, 属性 (軸) 数を3つ以下に絞ったスパースではない (0でない要素の多い) 集計データ (テンソルまたは行列) を複数作成し, 集計データ間で共通する属性 (図ではユーザの軸) に相当する因子行列を仮定して分解を行います (図3)。

複数の集計データを組み合わせることで多次元のデータを分析するため、多次元複合データ分析技術と呼んでいます。

## Yelpデータセットを用いた解析例

多次元複合データ分析技術NMTFの有効性を検証するため、米国のレビューサイトYelp（図4）のオープンデータセットに適用した事例を紹介します。Yelpデータセットは学術目的で一般公開<sup>(2)</sup>されており、それぞれの商業施設ごとの口コミ情報として、店舗名、カテゴリ、店舗位置、ユーザ、レーティング、レビューに含まれる単語、チェックインの曜日、時間など、多様な属性データを含んでいます。このデータから、例えば、次の3種類のテンソルデータを集計することができます。

- ① レビューテンソル：ユーザ×店舗×曜日…値はレビュー数
- ② チェックインテンソル：店舗×

時間×緯度経度…値はチェックイン回数

③ 単語頻度テンソル：ユーザ×カテゴリ×単語…値は単語の出現回数  
テンソル間には、ユーザ、店舗といった共通する軸を含んでいます。詳しい分解アルゴリズムや結果は参考文献

(3)に譲りますが、上記3つのテンソルに対してNMTFを適用することにより、図5に示すような「週末午前にレジャー活動」や「日本食レストランでの食事」がさかんな商業施設、利用されることの多い曜日・時間帯、地理的分布、口コミでよく使われる単語な



図4 Yelpデータセット

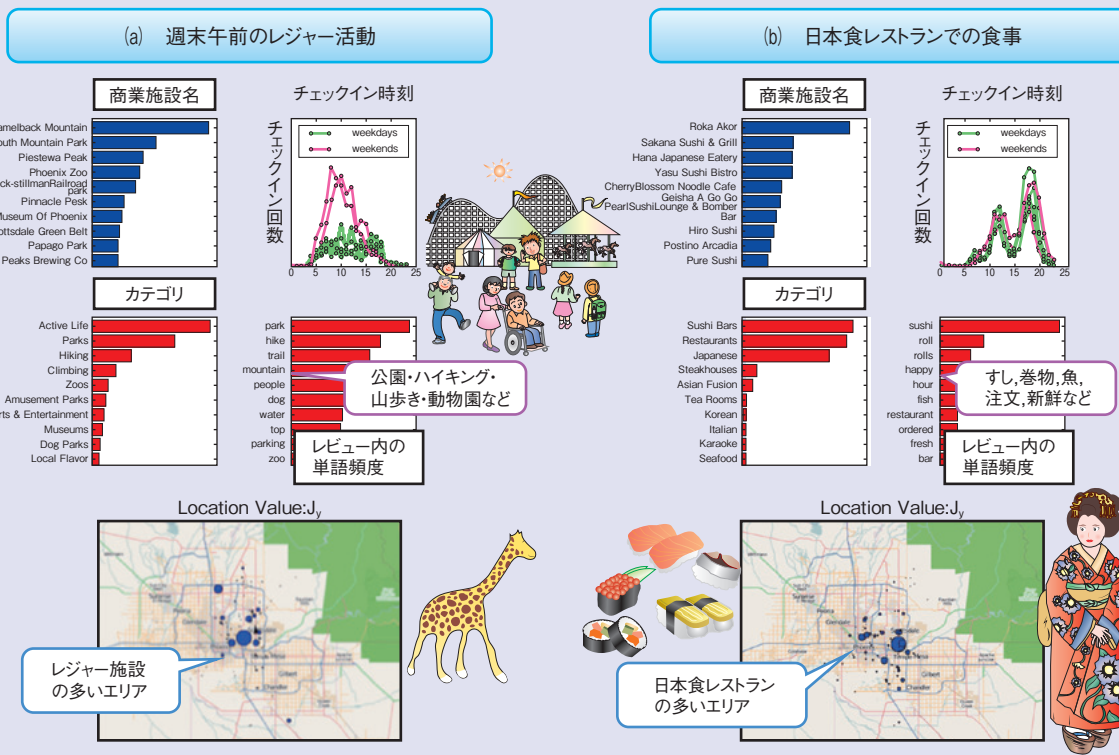
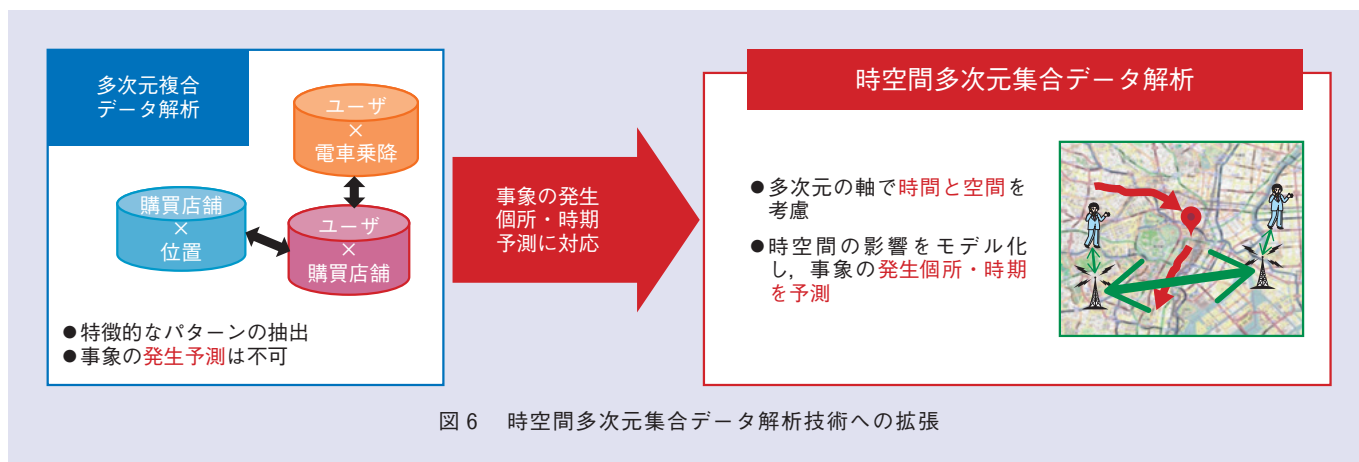


図5 多次元複合データ分析によって得られるクラスタ例





ど、多くの属性の組合せで特徴をとらえたクラスタが抽出できていることが分かります。このように、NMTFはデータの持つ多次元の属性を活用しながらも、スパース性の問題を緩和し、分析結果として意味のあるクラスタを効率的に抽出することができます。

## 今後の展開

NMTFによる多次元複合データ解析は、データに潜む特徴的なクラスタを効率的に抽出できますが、集計したデータ要素間の時間的な因果関係や空間的な隣接性などの関係性を考慮していません。特に、ユーザや車などの位置データの解析では、空間メッシュで区切ったエリアおよび時間帯ごとに観測された人口や車両数をベースに分析が行われることがあります。このようなデータにNMTFを適用しても、例えば混雑などの事象が発生する時期や場所を予測するといった用途に見合う分析結果は得られません。そこで、多次元データの時空間的な関係性をモデル化し、将来の予測を可能にする革新的解析技術として、時空間多次元集合データ解析技術の研究開発を進めています。時空間多次元集合データ解析技術はデータの「時間」「空間」「多次元」「集合」の4要素を考慮し、近未

来の事象を予見・洞察することを目指しています(図6)。「集合」とあるのは、メッシュ当りの人口や車両数などのように、個々を識別できない、集合としての集計データのみからでも、時空間的な人流や交通流の流れを推定することを意図したものです。

2020年に向けて、大規模イベント会場などでの混雑緩和や通信インフラの安定化のために、リアルタイムに観測されたデータを活用し、近未来に起こり得る混雑などの事象を時空間多次元集合データ解析技術によって検知し、先行的に集団を誘導する技術<sup>(4)</sup>の実現に向けて研究開発を進めていきます。

## 参考文献

- (1) 澤田: “非負値行列因子分解NMFの基礎とデータ/信号解析への応用,” 信学誌, Vol.95, No.9, pp.829-833, 2012.
- (2) [https://www.yelp.com/academic\\_dataset](https://www.yelp.com/academic_dataset)
- (3) K. Takeuchi, R. Tomioka, K. Ishiguro, A. Kimura, and H. Sawada: “Non-negative Multiple Tensor Factorization,” Proc. of ICDM 2013, pp.1199-1204, Dallas, U.S.A., Dec. 2013.
- (4) N. Ueda, F. Naya, H. Shimizu, T. Iwata, M. Okawa, and H. Sawada: “Real-time and Proactive Navigation via Spatio-temporal Prediction,” Proc. of the First International Workshop on Smart Cities: People, Technology and Data, in conjunction with Ubicomp2015, pp.1559-1566, Osaka, Japan, Sept. 2015.



(左から) 納谷 太/ 澤田 宏

多種多様なリアルタイムセンサデータから、潜在的な時空間特徴を即時にとらえ、将来を予測し、先行的に人々の誘導や機器を制御し、安心・安全な社会インフラ構築に貢献する技術の確立を目指します。ご意見、ご要望をお待ちしています。

## ◆問い合わせ先

NTTコミュニケーション科学基礎研究所  
協創情報研究部/  
機械学習・データ科学センタ  
TEL 0774-93-5272  
FAX 0774-93-5155  
E-mail [naya.futoshi@lab.ntt.co.jp](mailto:naya.futoshi@lab.ntt.co.jp)