インテリジェントソサエティを支える分析技術

Advanced Analytics for Intelligent Society

● 湯上伸弘　● 井形伸之　● 穴井宏和　● 稲越宏弥

あらまし

富士通研究所では，人や社会の知恵や行動，環境変化などに関する様々な情報を分析・活用することで，より豊かで安心できる社会を実現するためのインテリジェントソサエティの研究開発を進めている。人や社会の動きなどを分析し活用するためには，ブログやTwitterなどのマイクロブログ，SNSのような内容や書き方に関する多様性が非常に大きいテキストデータや，センサデータのように個別の情報量は少ないけれどもリアルタイムで大量に集まるデータといった，これまでビジネスインテリジェンスなどの従来の分析技術が扱ってきた企業内のビジネス情報とは全く異なる性質を持つデータを扱うための新しい分析技術が必要となる。

本稿では，それらの分析技術の中から，ソーシャルメディアを対象とした自然言語処理技術，人の行動や社会の動きを考慮した最適化技術，時刻や位置情報を大量に処理するための時空間データ処理技術の3種類の技術について，簡単な応用例を交えて紹介する。

Abstract

Fujitsu Laboratories is analyzing and utilizing various types of data on the behavior and actions of people and society, as well as environmental change. In this way, it is proceeding with R&D on Intelligent Society to achieve a more prosperous and secure society. This paper focuses on two new types of data. The first one is social media including blogs, Twitter, and social networking services (SNS). The second is data obtained from various types of sensors such as mobile phones, automobiles, and environmental sensors. These data are very different from business data that traditional analytic technologies deal with in business intelligence applications. To realize Intelligent Society, we are researching new and advanced technologies to analyze such data. This paper introduces three of the technologies: social media analysis, optimization and spatiotemporal data processing.

ま　え　が　き

本稿では，インテリジェントソサエティを実現するために必要となる新しい分析技術について，簡単な適用例を交えて紹介する。富士通研究所では，企業内のビジネスデータだけではなく，ソーシャルメディアやセンサデータなど，人や社会の動きに関する大量のデータを分析することで顧客や社会の動きを予測し，その結果に基づいてビジネス計画や社会問題の解決策を立案するための技術の研究開発を行っている。インテリジェントソサエティを実現するためには，従来の代表的な分析技術である統計やデータマイニングに加えて，情報抽出や情報統合，時系列分析と予測，最適化といった様々な技術が必要となる。

本稿ではその中から以下の3種類の技術について解説する。最初に，Twitterなどのソーシャルメディアから様々な人や社会の動きを見つけ出すために必要となる技術について，自然言語処理技術を中心に紹介する。つぎに，予測した結果に基づいて最適なビジネス計画や社会問題の解決策を立案するための最適化技術を紹介する。最後は，時間や空間に関するデータを扱うための時空間データ処理技術である。時刻や位置に関する集計や検索は，人の行動や社会の動きに関する分析を行うために必ず必要となる基盤的な技術であり，今後重要性が高まると考えている。

ソーシャルメディア分析技術

インテリジェントソサエティの大きな目標に，複雑化する社会問題の解決に向けたICTの貢献がある。そこで，著者らは，社会問題の解決に向けた第一歩として，ブログやマイクロブログ（Twitterなど），Mixi，Facebookなどのソーシャルメディアの分析を通して，「人や社会の動き」をとらえる技術の研究開発を行っている。以下では，ソーシャルメディア分析の例として著者らがこれまで研究してきた評判分析と社会俯瞰（ふかん）マップを紹介し，そこで必要となる技術について解説する。

著者らがニフティと共同で開発した評判分析技術は，ブログなどのソーシャルメディアから特定の商品やサービスに関する意見を収集し，その商品やサービスの認知度や強み弱みを分析するための技術である。**図-1**は，評判分析をある飲料水に対して行った結果である。図-1の例では，この飲料水が味に関しては否定的な評価が多いが，「やせる」という機能面からは高く評価されていることが分かる。

図-2は著者らが現在研究を行っている社会俯瞰マップの例である。社会俯瞰マップとは，ソーシャルメディア上の書込みの中から特定の分野（図-2の例では犯罪）に関係する書込みだけを選別し，同時に書込み内容から，「いつ」「どこで」起きた「どんな」出来事に関する書込みかを解析して，地図

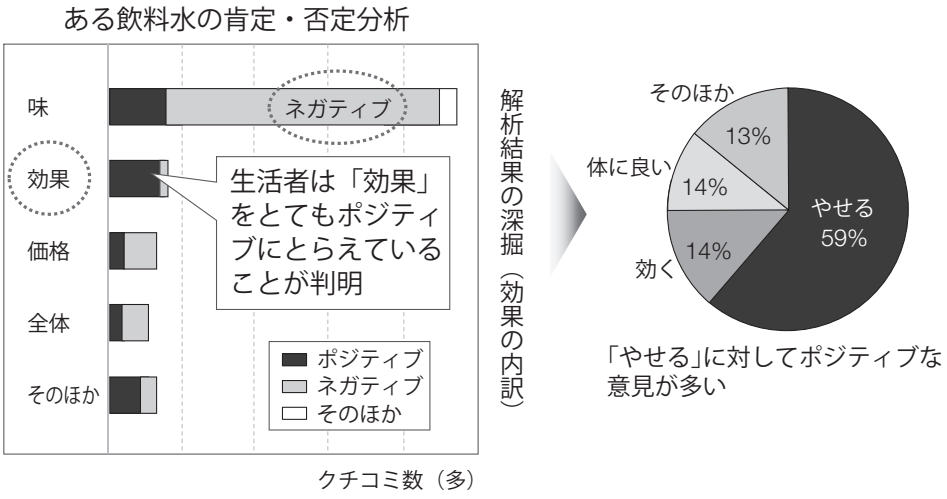


図-1　評判分析の分析例
Fig.1-Example of sentiment analysis.

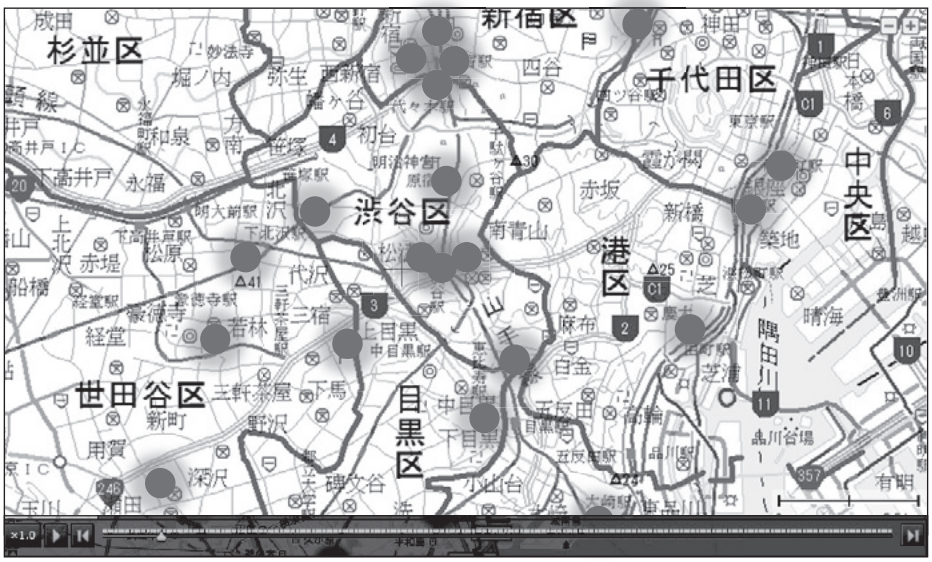


図-2　社会俯瞰マップ(犯罪マップ)
Fig.2-Social overview map (crime map).

上に表示するものである。この技術を様々な分野へ適用することで，住民の意見や行動，社会全体としての動きを，地域依存性や時間変化も含めて知ることができる。

上記のような評判分析や社会俯瞰マップを実現するためには，個別の書込みから必要な情報を抽出するための自然言語処理技術が必要になる。ただし，ソーシャルメディア上の書込みは，新聞記事のように5W1Hが明確に書かれているわけではなく，主に話し言葉により記述され，省略や略称が用いられることが多い。そのため，ある特定の単語をベースにした処理ではなく，よりロバストな自然言語処理が必要となる。

このような問題に対し，著者らは高速で高精度な機械学習を用いた自然言語処理技術の研究開発を行ってきた。[1] 機械学習とは，分類済みの事例の集合から，未分類の事例を正しく分類するための規則を自動的に生成する手法である。これにより，例えば，ある書込みが「不審者情報」かどうかを判定する場合に，単純に「不審者」という単語が含まれているかで判定するのではなく，「不審者」の関連語や文脈を考慮した判定が可能となる。例えば「不審者と間違われたかも…」という書込みは，「不審者」という単語を含むが「不審者情報」ではないと判定される。

また，書込みの内容から「どこで」を抽出する際にも，場所を示す表現として，地名のほかに，店舗名やランドマーク，それらの略称など，様々な表現が用いられることが問題となる。すべての場所表現に対して教師データを準備することは現実的ではないため，著者らは大量のテキストデータに対して統計処理を行い，判定ルールのパラメータとして補強することで，様々な表現に対応できる高精度の抽出技術を開発した。[2]

以上のように，ロバストな自然言語処理を用いた「社会俯瞰マップ」により，まだ断片的ではあるものの，「社会で起きた出来事」の検知・把握が可能となる。

つぎにソーシャルメディアを対象としたネットワーク分析技術について述べる。ソーシャルメディアは，人や書込み，話題などを頂点とする巨大なネットワークと考えることができる。そのため，大規模ネットワークの分析技術，とくにネットワークの構造やその時間変化を高速に分析する技術が必要になる。著者らはカーネギーメロン大学と共同で，大規模ネットワークから特徴的な部分ネットワークを高速に発見する手法の研究を行ってきた。[3] この手法では，ノードの次数やハブスコアと

いった特徴量の分布に着目することにより，ネットワークの大きさ（辺の数）に比例する時間で分析を行うことができる。これはソーシャルメディアのような非常に大規模なネットワークを分析する際の必須の条件である。この技術を使ってソーシャルメディア内の特徴的なコミュニティや話題を抽出し，それを使って社会の中の意識の変化やそれに基づく行動の変化を検出することが期待できる。

今後は，解析可能な分野を拡大するとともに，後述する時空間データ処理技術と組み合わせることで，より豊富で正確性の高い知識の抽出を目指している。

ダイナミック最適化技術

本章では，インテリジェントソサエティ実現のために求められる最適化技術の方向性と著者らの取組みについて紹介する。

最適化技術は，計算機の計算性能の飛躍的な向上と効率的アルゴリズムの進展と相まって，ロジスティクスやものづくりなど広範な分野で活用されている。近年では，インターネットの発達によるWeb情報，ソーシャルメディア，センサデータなどの膨大なデータから，価値ある知見を見出し最適なアクションプランを立案・実行するための基盤技術として，改めて最適化技術への関心が高まっている。

これまで著者らはロジスティックやものづくりなど様々な分野での最適化技術の開発・実用化研究を行ってきた。[4] その一つに，より複雑な制約条件のもとでの設計性能の向上，設計方針決定の効率化を目指した，パラメータ空間アプローチによる最適化技術がある。これは，設計仕様を満たす可能な設計パラメータの領域をすべてとらえて可視化することで，最適性やロバスト性といった解の性質を容易に把握できる最適化技術であり，富士通研究所で独自開発した数式処理技術を活用することで実現している（**図-3**）。この技術をハードディスクや半導体メモリなどの設計プロセスに適用し，ある設計工程では14日間かかっていたものを1日に短縮することができた。

インテリジェントソサエティが対象とする地域エネルギーマネジメントや市場品質マネジメントなどの問題における最適化は，上で述べた最適化技術に加えて新しい課題を解決する必要がある。例えば地域エネルギーマネジメントでは，電力の需給バランスを取るために家庭やオフィスでの電力需要と太陽光発電などによる発電量を事前に予測して最適な需給制御計画を立案する。しかし人の行動を完全に観測したり予測したりすることは不可能なので，予測自体にかなりの不確実性が含まれる。需給計画の最適化はこの不確実性を考慮した上で行う必要がある。また，家庭やオフィスの実際の消費電力や太陽電池の発電量は常に新しいデータが観測されるので，それに合わせて需給計画も変化させていく必要がある。

著者らは，これらの二つの課題，すなわち不確実性への対応と状況の変化への動的な対応を実現するためのダイナミック最適化技術の研究を進めている。予測の不確実性については，不確実性の範囲を事前に設定した上で最悪の事態を想定して最適化を行うロバスト最適化と，確率的な要素を

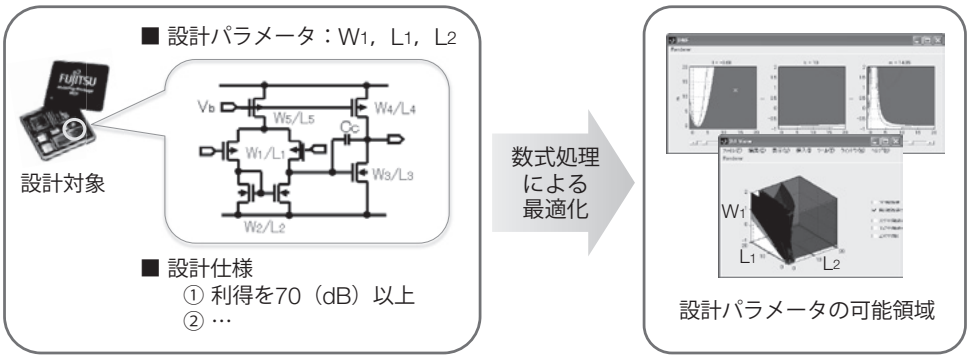


図-3　数式処理による最適化
Fig.3-Optimization by symbolic computation.

含んだ予測モデルに基づく確率的最適化の二つのアプローチに着目し研究を進めている。エネルギーマネジメントの例では，前者は，需要や発電量の変動幅を最適化問題に組み込むことで不確実性に対応する。後者は，変動幅が非常に大きくなり得る場合に対応するため，単一の需要予測や発電量予測を採用するのではなく，複数の予測を確率付きで求め，それを最適化問題に組み込んで需給計画を立案する。状況変化への対応では，状況変化を動的に問題設定に反映し，その都度最適な計画決定を効率的に行うことができる循環型の最適化の枠組みの確立を目指している。

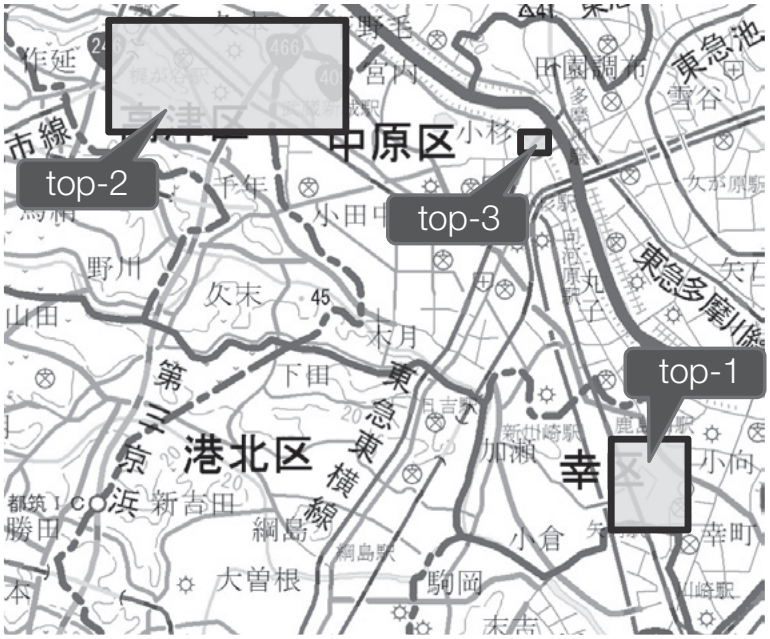
時空間データ処理技術

近年，携帯電話やカメラ，カーナビなどの様々な機器から大量のセンサデータがリアルタイムに収集できるようになってきた。これらのセンサデータには，観測値のほかに観測した時刻と位置に関する情報が含まれており，これらの情報を処理することで，人や社会の動きをこれまでにない精度で，かつ即時に知ることができる。人や社会の動きを知ることはインテリジェントソサエティがターゲットとする様々なソリューションに必須の要素である。そのため，センサデータなどの時間，空間に関するデータを大量に蓄積し，分析の際に必要なデータを検索・集計するための時空間データ処理はインテリジェントソサエティを支える基盤技術と言える。以下では時空間データ処理を支える技術として，スコア最適領域発見技術と非解凍型圧縮データ検索技術を紹介する。

空間情報を扱う代表的な方法として，空間をメッシュ状に分割し，集計や分析を行う方法が知られている。例えば，行政区メッシュを用いた人口統計分析や，固定長メッシュを用いた雨量集計など，簡便で使いやすく，結果が分かりやすいという良さがある。しかし，同じデータに同じ集計をしても，メッシュの切り方が異なるだけで結果が大きく異なる可能性がある。スコア最適領域発見は，集計値（スコア）から領域を決定するという，メッシュ分析とは逆のアプローチをとる。そのため，メッシュを指定する必要がなく，同じデータに対して常に同じ結果を得ることができる。スコア最適領域発見については，これまでは，密度や確率などのスコアを最大とする領域を一つ発見するためのアルゴリズムの研究が行われてきた。しかし現実のアプリケーションでは，着目すべき領域が複数あれば複数の領域を求めることが要求される場合が多い。そのため著者らは，スコアの順に，相互に交差しない複数の領域を高速に発見するSplitRegionSearchアルゴリズムを開発した。[5] この手法は，従来の手法を使って複数の領域を求める場合と比較して数十倍～数百倍高速であり，その効果は対象となるデータが多いほど大きい。例えば川崎市の人口統計データから複数の町をまたがるような人口増加地域を求める問題では，従来手法に比べ500倍以上高速に上位3個の領域を抽出することに成功した（**図-4**）。現在は，凹凸を含む複雑な形状の領域や，様々なスコアに対応したアルゴリズムの研究開発を行っている。

非解凍型圧縮データ検索は，文字どおり，圧縮された大量のデータを解凍せずに圧縮したまま検索を行う技術である。センサデータは連続的に増大し続けるので，それらを蓄積や転送する際には，圧縮してサイズを小さくすることが望ましい。しかし，圧縮したデータは処理する前に再び解凍する必要があり，これに要する計算量やメモリのコストが問題になる。非解凍型圧縮データ検索は，圧縮によるデータサイズ削減と，高速な検索を両立させ，この問題を解決できる可能性がある。そのためには，より小さく圧縮するための辞書の構築と，圧縮されたデータ上での検索の高速化とが必要である。ここでは前者の辞書の構築について述べる。

著者らが北海道大学と共同で開発した非解凍型圧縮データ検索技術[6]ではVF符号化と呼ばれる圧縮方式を採用している。VF符号化では，あらかじめ生成された辞書を用いて可変長の文字列を固定長の符号に置き換えることで圧縮を行う（**図-5**）。そのため限られた数の符号にどの文字列を割り当てるか，すなわち辞書の作り方によって圧縮率が異なる。高い圧縮率を実現するためには圧縮前のデータ中の文字列の出現頻度を考慮して辞書を構築する必要があるが，従来の技術では，辞書を最適化するために圧縮前のデータをすべてメモリ上に格納する必要があり，大規模なファイルに対して適用することができなかった。それに対して著



この背景地図などのデータは，国土地理院の電子国土Webシステムから配信されたものである

図-4　スコア最適領域発見の例(川崎市の人口増加地域)
Fig.4-Discovery of optimal areas with respect to score function (e.g. regions with highest population growth in Kawasaki).

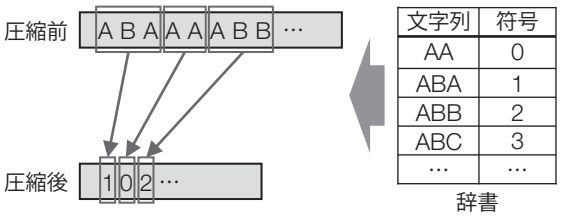


図-5　VF符号による圧縮の仕組み
Fig.5-Text compression by VF coding.

者らの技術では，数回ファイルをスキャンしてデータの特徴を学習することにより，すべてのデータをメモリ上に格納することなしに，すなわちファイルサイズに対する制限なしに，代表的な圧縮ツールであるgzipと同等の圧縮率を実現できる。

むすび

本稿では，インテリジェントソサエティを実現するための分析技術として，ソーシャルメディア分析技術，ダイナミック最適化技術，時空間データ処理技術の3種類の技術を紹介した。今後は，エネルギーマネジメントや交通，ライフログ，リスクマネジメントといった様々な分野への適用を通して，これらの技術をインテリジェントソサエティにおける基盤技術として確立することを目指して研究開発を進める。

参考文献

(1) 岩倉友哉ほか：大規模自然言語処理学習データのための複数弱仮説を生成する弱学習器を用いるAdaBoost手法．電気学会論文誌．C，Vol.130．No.1，p.83-91（2010）．
(2) 岩倉友哉ほか：ラベルなしデータを用いた素性増強による日本語固有表現抽出方法．情報処理学会論文誌，Vol.49，No.10，p.3657-3669（2008）．
(3) K. Maruhashi et al.：Spotting Connection Patterns and Outliers in Large Graphs. *ICDM Workshops 2010*, p.1328-1337（2010）.
(4) 穴井宏和ほか：数式処理を用いた設計技術．*FUJITSU*, Vol.60, No.5, p.514-521（2009）.
(5) 森川裕章ほか：大規模空間データからの最適領域集合の効率的な発見方法．第73回情報処理学会全国大会講演論文集，Vol.1，p.561-562（2011）．
(6) T. Uemura et al.：Training Parse Trees for Efficient VF Coding. *17th edition of the*

Symposium on String Processing and Information Retrieval (SPIRE 2010), LNCS 6393, p.179-184 (2010).

著者紹介

湯上伸弘（ゆがみ　のぶひろ）
ソフトウェアシステム研究所インテリジェントテクノロジ研究部 所属
現在，データマイニング技術および最適化技術などの研究に従事。

穴井宏和（あない　ひろかず）
ITシステム研究所デザインイノベーション研究部 所属
現在，数値・数式ハイブリッド計算技術および最適化と制御の研究に従事。

井形伸之（いがた　のぶゆき）
ソフトウェアシステム研究所インテリジェントテクノロジ研究部 所属
現在，自然言語処理によるソーシャルメディア分析の研究に従事。

稲越宏弥（いなこし　ひろや）
ソフトウェアシステム研究所インテリジェントテクノロジ研究部 所属
現在，文字列照合，時空間データ処理などの大量データ処理技術の研究に従事。