社会の動きを深く横断的に分析するためのビッグデータ処理技術

木俣 豊 Yutaka Kidawara (独) 情報通信研究機構ユニバーサルコミュニケーション研究所
 鳥澤健太郎 Kentaro Torisawa (独) 情報通信研究機構ユニバーサルコミュニケーション研究所
 是津 耕司 Koji Zettsu (独) 情報通信研究機構ユニバーサルコミュニケーション研究所

1 はじめに

インターネットはネットワークのブロードバンド化から、モバイル化、ユビキタス化などの技術革新を経て、いつでも、どこでも、誰もが使うインフラとなった。今や、人々はスマートフォンを介してネットワークを意識することなく、必要な情報にアクセスすることができる。当初は情報を引き出すプル型のICTサービスが主であったが、その場の興味のあるものを写真撮影した上で時空間情報を付与し、Facebook等のSNSやTwitter等のマイクロブログにアップロードすることやLINE等のメッセージ交換ツールで情報交換する双方向型のICTサービスが日常的なものとなっている。その結果、現在では、人々は特定の人だけでなく不特定の人も対象として、自分自身が持つ情報を交換する多様なコミュニケーション手段を手に入れており、そのICTサービス上で流通・蓄積されるデータは膨大なものとなっている。その一方で、人が介在しない機械がネットワークに接続され大量の情報がネットワークを介して機械に集積されるM2M (Machine to Machine) システムが急速に拡大している。従来は環境データをセンシングしたセンサ情報の収集などが主であったが、交通ICカードの利用履歴等の人々の活動を集積するシステム等も拡大しており、高度なICT社会が進めば進むほど、社会の動きがデータとして生成されている。これらのデータが集積され、その量が膨大になるにつれて、そのデータを解析することで社会の動きや環境の変化、人々の行動などの傾向や因果関係、相関関係が分かる可能性が増しており、大規模な情報から経済活動や社会活動、環境変動などを分析・推測するための技術に注目が集まっている。

2 ビッグデータとコミュニケーション

2.1 ビッグデータの種類と利活用領域

ビッグデータは、(a) 人々が能動的に生成・発信するデータ、(b) 人々の社会行動によってICTシステムが獲得するデータ、(c) 環境を計測したデータなどに大別できる。また、それぞれのデータはビジネスや社会活動、自然科学などの分野での経済活動や研究開発等に用いることが想定されている。これらは、ビジネスビッグデータ、ソーシャルビッグデータ、サイエンスビッグデータ等とも呼ばれ、ビジネスや自治体活動、科学の世界で新たな価値創造をもたらすものとして期待が大きくなっている。Cyber Physical System (CPS) は、大規模なデータから実世界の動きを分析することを目的としているが、データも目的も多様であり、目的に合わせてデータを収集した上で、適切な分析手法を利用することが必要となっている。そのため、多様な情報収集手段を管理することや多様な分析手法を柔軟に組み合わせることができ、情報利活用基盤が費用不可欠となっている。

2.2 社会の動きを深く分析するためのビッグデータ

ビッグデータの研究においては、前述のCPSの構築などが注目されがちであるが、センサデータや人流、物流データをそれぞれ分析したとしても、現在の動向や過去のデータと比較した変化等、ローレベルな現象しか検出できず、そうした現象をきちんと解釈し、社会実装可能な一種の施策に結び付けるには非常な困難が予想される。更には、漫然と収集可能なデータを蓄積し、そのデータに出来合いのマイニングアルゴリズムを適用しただけで価値ある情報が得られるとは考えづらい。そもそも、ビッグデータの収集、蓄積、分析は高価な処理であって、どのような目的で分析を行い、更には分析の結果どのような知見が得られるのか、一定の幅を持って想定してお

かない限り、高いコストパフォーマンスで有用な知見を得ることは困難であろう。

筆者らはこうしたセンサデータ等からのマイニングの問題点を補完するため、やはりビッグデータの一部であるネット上のテキスト情報の深い意味的な分析の研究開発を行っており、実際に億単位の Web テキスト、SNS 上のテキストなどを対象に深い意味的分析を行う情報分析システムを構築している。先ほど言及したビッグデータの分析の目的、想定される知見、更には分析結果の解釈なども、ネット上のテキストに表現されている人々の関心や社会の情勢、事件や事故の発生状況や原因などから直接的、間接的に求めることができるのではないかと考えており、筆者らのテキスト分析技術はこうした目的での利用も可能であると考えている。現在、開発している情報分析システムは、自然言語で書かれた複雑な要求を満たす情報を網羅的に取得したり、複数の文書に記載されている断片的な情報を意味的なつながりによって接続して、より意味の明確な情報へと半自動的に編集し、一種の未来予測を行うことなどが可能になっている。このようなシステムが利用可能となることによって、将来的には、データサイエンティストと呼ばれる人々は、まず、テキスト情報から社会の動きに関するハイレベルな仮説を入手し、次いでその内容を検証するために、環境センシングデータや、人流・物流データのマイニングを行うことも可能となる。筆者らの最終的な目標はこうした一連の流れを一環してサポートできるシステムを構築することにある。

筆者らが所属する NICT ユニバーサルコミュニケーション研究所では、このようなビッグデータ時代に対応する高度な情報分析技術や情報利活用技術を研究開発を推進している。第3章では、テキストビッグデータに対する深い意味的情報分析技術の現状について述べる。第4章では、異種・異分野の横断的情報分析技術について述べる。第5章は結論と今後の展望について述べる。

3 テキストビッグデータ処理技術

ビッグデータ分析の意義の一つは一見かけ離れた事象の間の普通の人間には想定し難い関連性を見いだすことであろう。データマイニングがもてはやされ始めた頃によく話題に上がった成功例として、「スーパーの売場でビールと紙おむつの売上に相関があることが分かったため、それらを並べておいたところ売上げが上がった」というようなものがあった。筆者らはこの真偽について確認をしたわけではないが、この例が印象的である理由

は、ビールと紙おむつというかけ離れたアイテムの売上げに相関があるという意外性であり、普通の人々では気が付かない相関をデータから計算機が見付けることができたという衝撃による。ビッグデータの出現はこれまで分析の対象とならなかった多様な事象に関しても、それらの間の意外な関連を特定し、社会の可能な将来像や問題の特定、そうした問題の解決や、施策、プロセスの意外な改善を可能にすると期待される。

NICT ユニバーサルコミュニケーション研究所の研究の目的の一つはビッグデータを取り巻くそうした状況を踏まえて、社会の将来像、問題の所在、改善の可能性などについて、ビッグデータの一部であるテキスト情報、特に Web や SNS 上に存在するテキストを深く分析し、ユーザが容易に理解可能な形で提供したり、気付きを与えることのできる技術を開発することである。近年の自然言語処理の進捗には目覚ましいものがあり、以下で紹介するように、10 年前には想像もしていなかったような分析が数億ページ単位の Web 文書などを対象として可能になりつつあり、分析の内容自体も様々な社会的な問題の検出や解決に直接貢献できるレベルに達しつつある。一方で、こうして検出された問題や、分析された因果関係などは、Web 上で人が書き表したテキストだけを根拠としており、厳密な検証を経ていない「仮説」ということになる。今後、センサデータ等、他種のビッグデータや科学技術文献など信頼のおけるテキストなどを用いて検証可能なものについては自動的な検証を行い、より信頼のおける分析結果を提供できるフレームワークに拡張する予定である。

より具体的に研究の内容について述べると、現在、筆者らは Web 文書を対象とした大規模情報分析システム WISDOM 2013⁽¹⁾を開発している。これは日々 2,000 万件以上の速度で収集・蓄積した Web アーカイブに含まれる億単位から数十億単位の Web ページを対象として、様々な言語処理、情報分析サービスを連携させ、従来にない高度な情報の分析を可能にすることを目的としたシステムである。現在約 200 台のサーバ上で稼働を開始していて、平成 26 年度に一般公開することを予定している。ユーザからの入力としては、自然言語による様々な質問を想定しており、その質問のタイプごとに異なる情報分析サービスが起動される。例えば、「少子化が進行すると経済はどうなる？」といった質問に対しては、筆者らが「未来分析」と呼んでいるサービスが起動され、質問の回答として、少子化の帰結として想定される様々な因果関係の連鎖、すなわちシナリオが木構造として表示される(図1)。現在のシステムの出力には、「少

図1 質問「少子化が進行すると経済はどうなる」による未来分析

子化は生徒の減少を介して、大学の経営を困難にする」、「少人数教育をもたらす一方で、教育の質の低下をもたらす可能性がある」、あるいは「労働力人口の減少を介して、エネルギー使用量の低下と省力化投資の増大をもたらす可能性がある」、更には「少子化と表裏である高齢化に伴う耕作放棄の増大を介して食料自給率の低下をもたらす」など、様々な事象の可能性が示されている。少子化と食料自給率やエネルギー使用量、省力化投資といった事象は多くのユーザにとってはそれぞれかけ離れたものであり、一種新鮮なものと写ると同時に、それらをユーザに提示することは、少子化という問題の全体像、深刻さを把握してもらう上で有効であろう。更に、例えば、少数の新聞記事等を読んだだけでこのような多岐にわたる少子化の帰結を一定程度網羅的に把握することは困難である。筆者らは、特に昨今の複雑化する社会における大問題は、その多様な、帰結も含めて全体像をきちんと把握した上でなければ適切な検討を行うことは難しく、ましてや、適切な解決策を考え出すには自動的な情報分析の活用が不可避であると考えている。

このシステムが出力する因果関係の連鎖はその構成要素である因果関係（例：「労働力人口が減少する」ならば「省力化投資が増える」）を億単位の Web ページから取得し、それらを自動的に組み合わせることで生成されている⁽²⁾。最終的に得られる連鎖全体は入力の中の Web ページにも書かれていないものが一定の割合で含まれており、更にはいまだ誰も想定したことのないシナリオを含んでいる可能性もある。例えば、「森林破壊が進むとどうなる？」という質問の回答には、「地球温暖化が進み、海水温度が上昇して、（海産物を介して食中毒を起こす）腸炎ビブリオを海水中で増大させ、食中毒を増やす」といったシナリオも含まれていた。これ以外にも「地球温暖化が食料不足をもたらす」シナリオが提示されていることも考えると、場合によっては世界的

図2 ファクトイド型質問「ナノテクノロジーによるビジネスは何か」に対する回答

レベルで大量の海産物が食料として使えなくなる可能性を示すシナリオと考えられ、このシナリオが妥当であるとすれば、その潜在的リスクは大きいと考えられる。この腸炎ビブリオのシナリオの各ステップを記載した文書は入力に含まれていたものの、森林破壊から腸炎ビブリオに至るシナリオ全体を記載した文書は入力中に存在せず、ひょっとすると誰も想定したことのないシナリオである可能性すらあった。その一方で、このシナリオの一部である、地球温暖化から腸炎ビブリオの増大に至る部分はその後、筆者らのシステムの出力とは独立に欧米の専門家によって検証され、バルト海において実際現在進行中のプロセスであることが“Nature Climate Change”^{*1} 上において報告された。このことは、このシナリオが、システムがテキスト情報のみから自動生成したものではあるものの、一定の妥当性を持っていることを示唆している。筆者らは少なくとも一般ユーザには想定が困難であるが、妥当であることが科学的に検証されたこうしたシナリオの存在そのものが、こうしたシステムを一般ユーザに広く提供し、考えるヒントとしてもらうことのインパクトを示していると考えている。

WISDOM 2013 には、未来分析以外に「なぜ地球温暖化が進むとプランクトンが減るのか」といった事象の理由、根拠を尋ねる質問に複数の文からなる回答を提示する「Why 型」質問応答⁽³⁾や、「ナノテクノロジーによるビジネスは何か」（図2）「原発廃止は何をもたらすか」「ダイエットで何をを使うか」といった主として名詞数語で回答できる質問である、いわゆるファクトイド型の質問に対する質問応答システム、更には与えられたキーワードに関して、良い評判、悪い評判等を情報の発信者のカテゴリごとに分類して提供する評価表現抽出機能などが実装されている。これらの機能は相互に連携

*1 <http://www.nature.com/nclimate/journal/v3/n1/full/nclimate1628.html>

させつつ使うことも可能であり、例えば、「ダイエットで何を使うか」という質問の複数の回答の評判をサーベイして、そうしたアイテムが引き起こした事故、事件をチェックするといった操作や、未来分析で得られた「地球温暖化が進むとプランクトンが減る」という因果関係、シナリオが理解不能、あるいは妥当でない因果関係と思われた際に、その根拠、理由をシステムに尋ねるといった操作がクリックのみで可能になっている。

また、これらのサービスの組合せは、情報の信頼性を調べることに利用することできる。例えば、「ダイエットで使える」という触れ込みの商品の評判に否定的なものが多ければ、そもそも、その商品がダイエットに使えるというオリジナルの情報の信頼性が疑われるし、未来分析で得られた因果関係やシナリオの根拠や理由として、それらしい情報が得られなければ、そのシナリオの妥当性は薄いと判断することも可能となる。これらの機能を様々な形で連携させて活用することによって、一般ユーザの想像を超える、社会における問題や、その解決策の発見で役立てることができると考えている。

また、今後はこうした機能の更なる高精度化、大規模化を行うと同時に、ある行為の適切な方法を尋ねる「How型」の質問応答を導入し、既に実装済みのファクトイド型質問応答、「Why」型質問応答と合わせて、いわゆる「5W1H」を一とおりカバーする予定である。また、以上に挙げた各種機能によって取得された情報に対して、それらをサポートする情報や矛盾する情報を億単位のWebページから取得し、情報の信頼性を検証する手段も強化していく予定である⁽⁴⁾。分析の対象文書もWISDOM 2013の場合はWebが中心であるが、今後、本章冒頭で述べたように、信頼性が担保されている科学技術文献や政府白書、そこに現れる数値データ、更には各種センサーデータも分析対象に加え、より信頼性の高い情報を提供できるシステムを構築する予定である。更に、大規模災害時に発信される膨大なTwitter情報のリアルタイムな分析を行い、被災者、救援者に重要な情報を提供するシステムとして、対災害情報分析システム^{(5),(6)}の開発も進めており、いずれはこのシステムのリアルタイムでの分析機能とWISDOM 2013との統合も検討している。

4

異種・異分野横断的 ビッグデータ処理技術

4.1 オープンデータの利活用

近年、様々な組織や機関、個人が収集したデータを公共財として自由に利活用できるようにするオープンデー

タの取組みが世界的に広まってきている⁽⁷⁾。代表的な例として、公共データを利活用できるようにする取組みである米国のData.gov⁽⁸⁾が挙げられる。Data.govは、米国政府機関が保有する様々な統計データを公開し、金融や雇用、科学技術、交通、小売、通信、貿易、エネルギーなど46カテゴリー、約38万のデータセットを提供している。更に、同様のポータルサイトが、現在40か国以上に広がっている（英国Data.gov.uk、インドData.gov.inなど）。そのほかにも、科学技術データに関して、世界科学会議（ICSU）の運営するWorld Data System（WDS）⁽⁹⁾では、分野横断型国際データ組織の先駆として、品質管理されたデータの長期保全と提供を目的に、世界50か所以上から集めたデータを公開している。

こうしたオープンデータの利活用で特徴的なことは、これまで個別のプロジェクトやアプリケーションで扱われていたデータを横断的につなぎ合わせ、分野を超えた新たな価値を創造する試みが盛んに行われていることである⁽¹⁰⁾。ビッグデータで重要視されている3V（Volume, Variety, Velocity）において、これまでは主に圧倒的なデータ量（Volume）をリアルタイム性高く（Velocity）扱えるようにすることで従来のアプリケーション性能を大幅に改善させるスケールメリットが重要視されてきたが、オープンデータの広がりにより全体を把握しきれないほど多種多様な（Variety）データが利用できる環境が整いつつあり、これらを横断的に活用しコネクションメリットを発揮する技術へのニーズが高まってきている。筆者らは、こうしたコネクションメリットを支える技術の研究開発に取り組んでいる。

4.2 分野横断相関検索

オープンデータによって様々なデータが自由に活用できるようになったとしても、必要なデータにたどり着けなければ恩恵を得ることができない。そこで、検索が重要な役割を果たす。今日主流となっているのは、従来デジタルライブラリを中心に行われてきたカタログ検索である。データを公開した際の登録情報（名前や作成者等の書誌情報）に対するクエリを指定し、ヒットするデータの一覧を表示する⁽¹¹⁾。こうしたカタログ検索では、クエリを正確に指定する必要があるが、専門家以外には扱いにくいものであった。オープンデータの特徴を生かし分野を超えたコネクションメリットを発揮するには、専門分野に閉じたデータの検索だけでなく、分野を超えてつながる可能性のあるデータを積極的に発見することが大切である。

筆者らが研究開発を行っている分野横断相関検索技術 (Cross-DB Search) ^{(12), (13)} では、オープンデータの中でも特に実世界の状況を反映した様々なセンサデータや科学データを対象に、地理的な相関性、時間的な相関性、概念的な相関、及び引用参照の相関性の四つの観点からデータ間のつながりを評価し、クエリにヒットするデータに対しこれらの相関を複合化し、相互につながるの強いほかのデータ集合を見付け出す。例えば、大気品質の一つの指標であるPM2.5をクエリとして与えると、PM2.5の観測データだけではなく、それらの周辺地域や時期に作成されたデータや、PM2.5と関連性のある分野のデータ、更にはそれらと一緒に引用参照されることが多いデータをいもづる式に検索し、これらの中から相互に高い相関を示すデータ集合を発見する。図3に示した検索結果の例では、1990年代の南極海地域でPM2.5とバイオ分野のデータ集合が高い相関を示している (概念相関には地球環境科学用語体系のSWEET ontology ⁽¹⁴⁾ を使用)。また、同じクエリに対し、これ以外にも複数の相関データ集合の候補が発見される。Cross-DB Searchは、このように従来のカタログ検索の結果から相関の高いデータを追加する一方相関の低いデータを除去することで、気候変動や大気品質など様々な分野にまたがるデータを発見しやすくし検索サービスの利用品質向上に役立っている。Cross-DB Searchでは、現在、World Data Systemに参加している25種類、40分野にわたる科学データや、独自に収集した35種類以上のセンサデータを中心とした114万件のデータを対象に、相関検索に必要なメタデータを抽出し索引付けをしている。

筆者らは、この分野横断検索を、オープンデータを対

図3 分野横断相関検索 (Cross-DB Search)

象としたデータキュレーションに応用している⁽¹⁵⁾。データキュレーションは、ある調査や研究目的のために様々なデータを収集、整理、保存、公開する一連のプロセスであり、従来から大学の図書館などでデータの収集のために行われていた。オープンデータの進展により、個人レベルで膨大な種類のデータが利用できるようになった今日、データ収集の専門家による組織的かつ計画的なキュレーション作業から、個々の科学者が自身の研究の中で個人的かつ発見的な (heuristic) データキュレーションの可能性が生まれつつある。そこでは、オープンデータを使って、相関性を検証するためのデータセットの作成と、データセットを分析し様々な観点から相関性を発見することを繰返し行いながら、相関分析と検証データを同時に絞り込んでいく。こうしたサイクルをいかに迅速かつ効率的に回すかが、特にデータ中心科学⁽¹⁶⁾ などでは重要とされ、分野横断相関検索による研究の効率化が期待される。

4.3 異種データの時空間相関可視化技術

様々な分野のデータの横断的な関係性を把握する上で、情報可視化は有効な手段である。筆者らは、物理センシングや社会センシング (SNS など) によって得られる様々な種類のセンサデータや、様々な分野の観測データを蓄積した科学データアーカイブ、など実世界を反映したオープンデータを中心に、実世界の様々な出来事 of 分野横断的な相関性を視覚的に分析する情報可視化技術の研究開発を行っている。筆者らの STICKER (Spatio-Temporal Information Clustering and Knowledge ExtRaction) ⁽¹⁷⁾ では、地理空間と時間軸から構成される三次元空間上に様々なセンサデータを表示し、様々な可視化手法を用いてそれらの動きや変化の相関性を視覚的に分析する。図4に、STICKERによる可視化の例を示す。図4に示す例では、2013年8月の国内における大気品質 (PM2.5) のデータと、降水量のデータ、渋滞に関連するキーワードを含むTwitterのデータが時空間上に表示されている (平面が地理空間で縦軸が時間)。PM2.5と渋滞Twitterは、8月中旬頃 (中段) に東日本から西日本にかけて広い地域で重なり高い相関が見られる。この部分のデータを調べると、お盆休みで渋滞が多発しPM2.5の濃度増加したことが分かる。一方、8月上旬と下旬 (下段、上段) の降水量が多い時期・地域ではPM2.5との重なりが少なく、つまり相関が低く、降水がPM2.5の濃度を減少させていることが分かる。

このように様々なデータを (時) 空間上で重ね合わせることは、従来からデータマッシュアップなどで盛んに

図4 異種データの時空間相関可視化システム STICKER

行われてきた。筆者らの STICKER の特徴は、単にデータを可視化するだけでなく、様々なデータの中から相関の高いデータの組合せを発見し、かつ高い相関を示す部分のデータを抽出する Visual analytics の機能を備えている事である。Visual analytics は、データの操作と可視化をインタラクティブに行いながら仮説の検証や意思決定を行う分析プロセスである。図4の例では、まず PM2.5 のデータの分布を可視化し、異常に変化している箇所などユーザが関心を持つ部分を指定し、それらと相関の高いほかのデータを検索する。更に検索結果を時空間上に表示する際に、可視化するデータの種類や値域を操作しながら相関が高くなるように調整する。こうしたデータの表示や検索、操作を繰返しながら相関の高いデータの組合せを発見し、後に続く詳細なデータ解析に必要なデータを準備する。これらの作業は、膨大な種類のオープンデータから役立つデータを効率良く絞り込む上で重要である。STICKER では、様々な分野のデータに対し横断的な操作や可視化を可能にすべく、STT セル (Space, Time, Theme) 形式でデータを統一的にモデル化し、各種データから STT セル形式に変換して、様々な形状の可視化オブジェクトの生成とそれらの視覚的操作、STT 属性に基づくデータの選択や集約、異常値検出などの各種データ操作、並びに前節で述べた相関検索を実装している。

4.4 参加型オープンデータ利活用プラットフォーム

オープンデータの流れは、これまで主に政府や研究機関の主導により既存のデータを公開する活動が中心であった。それら初期の取組みの成功を受け、今日ではユーザが自らデータを収集、公開し共有する参加型のオープンデータ活動が広まりつつある。例えば、

図5 参加型オープンデータ利活用プラットフォーム K-L Grid

OpenStreetMap⁽¹⁸⁾ では、ボランティアベースで地理情報を収集しフリーな地図を開発することが行われている。また、人々が協力してローカルな状況をスマートフォンなどを用いて発信し、それらを幅広く収集し共有することで広範囲の状況を網羅性高く把握する参加型センシング (participatory sensing)^{(19), (20)} も、環境問題や社会問題、災害時の対応への有力な手段として行われている。こうした取組みは、これまでのトップダウンのオープンデータの流れから、ユーザが自ら実世界の問題解決に積極的にオープンデータの仕組みを活用しようとするボトムアップの流れに変化しつつあることの表れである。

このような参加型のオープンデータ活動を支える ICT プラットフォームとして、筆者らは K-L Grid (知識・言語グリッド)^{(21), (22)} の開発を行っている (図5)。K-L Grid は、新世代通信網テストベッド JGN-X⁽²³⁾ 上に構築された広域分散グリッドネットワークであり、サービス指向アーキテクチャ (SOA) に基づきデータの収集や分析、配信を行うサービスを開発したり、それらを連携させたアプリケーションを開発する情報サービス基盤である。この上で、データの収集や共有を協調して行い、オープンデータのリポジトリを構築している。実際に、Live-E!⁽²⁴⁾ から全国規模の気象データの収集したり、世界規模の科学データベース World Data System から大規模な科学データのメタデータを収集したり、Twitter など SNS データの収集を行っている。表1に、現在 K-L Grid 上に構築されている情報資産の例を示す。

表 1 K-L Grid 上の情報資産 (2013 年 10 月時点)

分 類	種 類
物理センサデータ	地震情報, 地滑り危険地域情報, 感染症情報, 花粉情報, 環境放射線水準情報, 降雨量情報, 積雪量情報, 気温情報, 風速風向情報, 台風情報 (国内・国外), 犯罪情報, 河川水位・雨量情報, 潮位情報, インフルエンザ情報, 世界災害情報など
社会センサデータ	Twitter アーカイブ, ジオタグ付 Twitter データ, トレンドキーワードで集約した Twitter データ, RSS ニュース, Google ニュース
WISDOM 2013 データ (含 Web アーカイブ)	全文データ, 言語解析済みデータファイル, 発信者データ, 評判データ, 係り受けデータ
科学データ	World Data System メタデータ (Pangaea, ICPSR, DRYAD, ESDS, ADA など 25 種類 40 分野)
地理データ	ランドマーク, 避難所データ
言語データ	EDR 概念辞書, 日本語 WordNet, WordNet
情報分析ツール	意見評価, 一休サービス, 文・フレーズ間の意味的関係 DB 等
翻訳ツール	VoiceTra テキスト翻訳, JServer
テキスト解析ツール	固有名詞抽出, 形態素解析器, 係り受け解析器
地理情報ツール	Google Geocoding, Yahoo Contents Geocoder, ランドマーク名抽出, 郵便番号検索, GeoNLP
音声ツール	VoiceTra 音声サービス (認識, 合成)

現在, K-L Grid は国内 5 拠点 (東京, 北陸, 京都, 岡山, 福岡) のグリッドノードによって構成されている。また, グリッドネットワークの特徴を生かし, ユーザ独自のノードを参加させることで, ユーザが K-L Grid 上で自身のデータを参加ノード経由で公開したり, ほかのデータと組み合わせて利用することを可能にしている。また, 情報サービスの要求に応じてネットワークの構成を自動的に設定する Service-Controlled Networking (SCN) 技術^{(25), (26)}の研究開発も行い, データの収集, 加工 (フィルタリングや分割・統合など), スループット等の要件に基づいてネットワークの発見やパス生成, QoS 制御を動的に行う仕組みを実現することで, 従来に比べより多くの情報サービスをより安定的に移動させることを可能にしている。更に, こうした取組みを普及させるべく, 米国標準技術院 (NIST) と協力し標準化活動にも取り組んでいる⁽²⁷⁾。

5 まとめ

現在, 様々なところでビッグデータというキーワードを耳にするが, ビッグデータは処理技術がなければ, ただの大規模データであり, ごみの山でもある。企業が多くの予算をつぎ込み, 販売データを解析した結果, 「既知のことを再確認しただけであった」という事例も少なくない。過去の蓄積された情報を分析することで価値のある情報が見付け出されることもあるが, ビッグデータ処理技術への期待は, 過去の情報から現在, そして未来の状況を推測することにある。NICT ユニバーサルコミュニケーション研究所では, 他に類を見ない切り口でのビッグデータ処理技術の研究開発を推進しており, そ

の成果を社会に提供することを目指している。現在, 研究開発を進めている情報分析エンジン WISDOM2013 は, 平成 26 年度には一般公開を開始する予定であり, 広く社会で利用されることを目指している。また, その技術を用いた対災害情報システムは災害時に適切な避難・支援情報を提供することを目的として研究開発が進められている。異種・異分野の情報分析技術についても, 研究開発成果を研究者を対象とした情報検索システムに組み込み実用サービスとして公開を予定している。

文献

- (1) M. Tanaka, S. De Saeger, K. Ohtake, C. Hashimoto, M. Hijiya, H. Fujii, and K. Torisawa, "WISDOM2013: A large-scale web information analysis system," In Companion Volume of the Proc. 6th International Joint Conference on Natural Language Processing (IJCNLP 2013) (Demo Track), pp.45-48, Nagoya, Oct. 2013.
- (2) C. Hashimoto, K. Torisawa, S. De Saeger, J.-H. Oh, and J. Kazama, "Excitatory or inhibitory: a new semantic orientation extracts contradiction and causality from the web," Proc. Conference on Empirical Methods in Natural Language Processing and Natural Language Learning (EMNLP-CoNLL 2012), pp.619-630, Jeju, Korea, July 2012.
- (3) J.-H. Oh, K. Torisawa, C. Hashimoto, M. Sano, S. De Saeger, and K. Ohtake, "Why-question answering using intra- and inter-sentential causal relations," Proc. 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013), pp.1733-1743, Sofia, Bulgaria, Aug. 2013.
- (4) J. Kloetzer, S. De Saeger, K. Torisawa, C. Hashimoto, J.-H. Oh, and K. Ohtake, "Two-stage method for large-scale acquisition of contradiction pattern pairs using entailment," Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP 2013), pp.

- 693–703, Seattle, USA, Oct. 2013.
- (5) K. Ohtake, J. Goto, S. De Saeger, K. Torisawa, J. Mizuno, and K. Inui, “NICT disaster information analysis system,” In Companion Volume of the Proc. 6th International Joint Conference on Natural Language Processing (IJCNLP 2013) (Demo Track), pp.29–32, Nagoya, Oct. 2013.
 - (6) I. Varga, M. Sano, K. Torisawa, C. Hashimoto, K. Ohtake, T. Kawai, J.-H. Oh, and S. De Saeger, “Aid is out there: Looking for help from tweets during a large scale disaster,” Proc. 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013), pp.1619–1629, Sofia, Bulgaria, Aug. 2013.
 - (7) Global Open Data Initiative, <http://globalopendatainitiative.org/>
 - (8) Data.gov, <http://www.data.gov/>
 - (9) ICSU World Data System, <http://www.icsu-wds.org/>
 - (10) Linking Open Data, W3C SWEO Community Project, <http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData>
 - (11) CKAN, <http://ckan.org/>
 - (12) E. Gonzales, B.T. Ong, and K. Zettsu, “Searching inter-disciplinary scientific big data based on latent correlation analysis,” Workshop on Big Data and Society (in conjunction with IEEE BigData 2013), 2013.
 - (13) E. Gonzales, B.T. Ong, and K. Zettsu, “Optimized class association rule mining using genetic network programming with automatic termination,” Proc. 2nd International Conference on Advances in Information Mining and Management, pp.102–107, Venice, Italy, Oct. 2012.
 - (14) Semantic Web for Earth and Environmental Terminology (SWEET), <http://sweet.jpl.nasa.gov/>
 - (15) S.K. Sowe and K. Zettsu, “Collaborative development of data curation profiles on a wiki platform: Experience from free and open source software projects and communities,” Proc. 9th Int’l Symposium on Open Collaboration (WikiSym and OpenSym 2013), Article no.16, Hong Kong, China, Aug. 2013.
 - (16) T. Hey, S. Tansley, and K. Tolle, The Fourth Paradigm: Data-Intensive Scientific Discovery, Microsoft Research, 2009.
 - (17) K.-S. Kim, R. Lee, and K. Zettsu, “mTrend: Discovery of topic movements on geo-microblogging messages,” Proc. 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information System, pp.529–532, Chicago, USA, Nov. 2011.
 - (18) OpenStreetMap, <http://www.openstreetmap.org/>
 - (19) J. Goldman, et al., “Participatory sensing: A citizen-powered approach to illuminating the patterns that shape our world,” CENS white paper (2009), available at http://www.wilsoncenter.org/sites/default/files/participatory_sensing.pdf
 - (20) N.D. Lane, E. Miluzzo, H. Lu, and D. Peebles, “A survey of mobile phone sensing,” IEEE Commun. Mag., vol.48, no.9, pp.140–150, Sept. 2010.
 - (21) Y. Murakami, M. Tanaka, A. Bramantoro, and K. Zettsu, “Data-centered service composition for information analysis,” Proc. IEEE 9th International Conference on Service Computing, pp.602–608, Honolulu, USA, June 2012.
 - (22) M. Tanaka, Y. Murakami, and K. Zettsu, “Data-intensive services for large-scale archive access,” Proc. IEEE 9th International Conference on Service Computing, pp.617–624, Honolulu, USA, June 2012.
 - (23) 新世代通信網テストベッド JGN-X, <http://www.jgn.nict.go.jp/>
 - (24) Live EI, <http://www.live-e.org/>
 - (25) K. Zettsu, “Crossdata Platform for Leveraging Big + Diverse Data,” The 2nd Indonesian-Japanese Conference on Knowledge Creation & Intelligent Computing, Keynote, March 2013.
 - (26) 木全 崇, 豊村鉄男, 金 京淑, 是津耕司, “Service-Controlled Networking に基づくネットワークフローの動的制御手法,” 信学技報, SC2012-2, pp.7–12, June 2012.
 - (27) E. Simmon, K.-S. Kim, E. Subrahmanian, R. Lee, F. de Vault, Y. Murakami, K. Zettsu, and R.D. Sriram, “A vision of cyber-physical cloud computing for smart networked systems,” NISTIR 7951 (2013), available at <http://www2.nict.go.jp/univ-com/isp/doc/NIST.IR.7951.pdf>

木俣 豊 (正員)

1990 神戸大学大学院自然科学研究科計測工学専攻了。同年 (株) 神戸製鋼所入社, 1999 神戸大学大学院自然科学研究科情報メディア科学専攻博士後期課程了 (工博), 2001 (独) 通信総合研究所入所, 2006 内閣府へ出向, 2007 (独) 情報通信研究機構 知識処理グループリーダー, 2011 から同機構 ユニバーサルコミュニケーション研究所長。第 54 回情報処理学会全国大会優秀賞, 第 56 回前島賞, 第 43 回市村学術貢献賞受賞, 第 11 回ドコモ・モバイル・サイエンス賞 (奨励賞)。電子情報通信学会データ工学研究専門委員長。

鳥澤健太郎

1992 東大理学部卒。1995 同大学院博士課程中退。同年同大学院助手。1998 科学技術振興事業団さきがけ研究 21 研究員兼任 (2002 まで)。北陸先端大助教授を経て, 現在, 情報通信研究機構情報分析研究室室長。博士 (理学)。自然言語処理の研究に従事。日本学術振興会賞など受賞。

是津耕司 (正員)

1992 東工大・工・情報卒。2005 京大大学院情報学研究科博士課程了。博士 (情報学)。1992 日本 IBM 入社。2003 通信総合研究所 (現 情報通信研究機構) 入所, 2011 から同ユニバーサルコミュニケーション研究所情報活用基盤研究室室長, 現在に至る。データベース, 情報検索に興味を持つ。情報処理学会, 日本データベース学会, ACM 等会員。