「ビッグデータ統合利活用のための次世代基盤技術の創出・体系化」研究領域
領域活動・評価報告書
－平成28年度終了研究課題－

研究総括　喜連川　優
副研究総括　柴山　悦哉

1．研究領域の概要

ICT の社会浸透や、実世界から情報収集するセンサーや計測・観測機器の高度化と普及に伴い、様々な分野で得られるデータは指数関数的に増大し、多様化し続けています。これらのビッグデータの高度な統合利活用により、新しい科学的発見による知的価値の創造や、それらの知識の発展による社会的・経済的価値の創造やサービスの向上・最適化などにつながる科学技術イノベーションが期待されています。

本研究領域では、ビッグデータの複数ドメインに共通する本質的課題を解決し、様々な分野のビッグデータの統合解析を可能にする次世代基盤技術の創出・高度化・体系化を目指します。

具体的には、大規模データを圧縮・転送・保管する大規模管理システムの安定的運用技術や、多種多様な情報を横断して検索・比較・可視化して真に必要となる知識を効率的に取り出す技術、これらを可能にする数理的手法やアルゴリズムなどの開発を推進します。これらの研究の推進にあたり、ビッグデータから社会における価値創造に至るシステム全体の設計を視野に入れ、ICT 以外の分野との積極的な連携・融合によって社会受容性の高い次世代共通基盤技術の創出・高度化・体系化に取り組みます。

また、本研究領域では、関連領域の「科学的発見・社会的課題解決に向けた各分野のビッグデータ利活用推進のための次世代アプリケーション技術の創出・高度化」で得られる次世代アプリケーション基盤技術やデータを共有・活用するなどの連携を推進します。

2．事後評価対象の研究課題・研究者名

件数：　6件

※研究課題名、研究者名は別紙一覧表参照

3．事前評価の選考方針

選考の基本的な考えは下記の通り。

1）選考は、「ビッグデータ統合利活用のための次世代基盤技術の創出・体系化」領域に設けた選考委員（領域アドバイザ）14 名の協力を得て、研究総括が行う。

2）選考方法は、書類選考、面接選考及び総合選考とする。

3）　選考に当たっては、さきがけ共通の選考基準（URL：http://www.jst.go.jp/pr/info/info986/sankou2.html）の他、以下の点を重視した。本研究領域では、今後大きく展開することが予想されるビッグデータ時代の基盤的な研究開発を対象とした。基盤技術とは言え、応用を想定しない技術開発は、その評価軸の設定が必ずしも容易ではないため、ある程度の応用を想定した提案を重視した。

4．事前評価の選考の経緯

応募があった選考対象件数は 100 件あった。一応募課題につき領域アドバイザー12 名が分担して各申請の書類査読（書類審査）を行い、査読結果（評点、コメント）を提出した。次に、書類選考会議においてアドバイザー全員が一堂に会して査読結果を元に議論を行い、12 件を面接対象とした。書類選考会では、査読結果に大きな評価差があったものについては、評価要因を全員で共有して集中的に審議した。その後、面接および総合選考の結果、最終的に採択候補課題 6 件を選定した。

選　考	書類選考	面接選考	採択数
対象数	100件	12件	6件

※本領域においては、5年型、大挑戦型を公募しなかった。

5．研究実施期間

平成 25 年 10 月～平成 29 年 3 月（3年型）

6. 領域の活動状況

1） CREST/さきがけ2領域合同領域会議

・平成 25 年 10 月 18 日　領域キックオフミーティング JST 別館ホール

・平成 26 年 11 月 5 日～6 日　H26 年領域会議　神戸ポートピアホテル

・平成 27 年 10 月 6 日～7 日　H27 年領域会議　TKP ガーデンシティ仙台

2） 公開シンポジウム

・平成 27 年 3 月 19 日　第 77 回情報処理学会イベント企画　京都大学百周年時計台記念館

・平成 28 年 2 月 4 日　情報処理学会ソフトウェアジャパン 2016　一橋講堂

3） JST/NSF 合同国際シンポジウム/NSF PI ミーティング

・平成 28 年 4 月 20 日～22 日　NSF PI ミーティング　米国　ワシントン DC　ジョージタウン大学

・平成 28 年 5 月 11 日～12 日　合同国際シンポジウム　アキバホール　＆　伊藤国際学術研究センター

・平成 28 年 11 月 28 日～30 日　合同国際シンポジウム　ベルサール六本木

・平成 29 年 3 月 15 日～17 日　NSF PI ミーティング　米国　ワシントン DC　バージニア工科大学

4）研究総括（または副研究総括）、領域担当、事務参事の研究実施場所訪問（サイトビジット）：全研究者の研究室訪問と研究者上司への挨拶を実施した。

・生貝　直人　研究者：　JSTにて H27/12/22

・佐藤　一誠　研究者：　東京大学　杉山将教授訪問　H28/1/19

・田部井　靖生　研究者：　東京工業大学　渡辺治教授訪問 H28/1/8

・松谷　宏紀　研究者：　慶應義塾大学　天野英晴教授訪問 H28/2/16

・水野　貴之　研究者：　国立情報学研究所　曽根原登主幹訪問 H28/1/13

・宮尾　祐介　研究者：　国立情報学研究所　佐藤真一主幹訪問 H28/1/13

7. 事後評価の手続き

年2回程度開催した領域会議、公開シンポジウムなどにおいて、研究者が口頭発表やポスター展示を実施し、その場で研究総括、アドバイザーなどから適時・適切なアドバイスを行った。また、半期毎に研究者が作成した研究報告書を研究総括、副総括が確認し、アドバイスやコメントを研究者にフィードバックした。最終年度には、事後評価会を実施し、研究者からの成果発表と総括、アドバイザーによる議論を行った。最後に、研究総括、副研究総括が研究者の終了研究報告書およびこれまでの進捗状況を総合的に判断し、最終的な事後評価結果を作成した。

（事後評価の流れ）

平成 28 年 10 月 28 日　　事後評価会実施

平成 29 年 1 月　　研究報告書提出

平成 29 年 2 月　　研究総括による事後評価

平成 29 年 2 月　　被評価者への結果通知

8. 事後評価項目

（1）研究課題の目標に対する達成度

（2）研究成果（論文、発表、特許など）

（3）研究成果の基礎研究・科学技術イノベーション及び社会・経済への波及効果（今後の見込みを含む）

（4）研究の進め方（他研究者との連携、国内外研究者・産業界等との連携、研究費執行状況など）

9. 評価結果

「ビッグデータ統合利活用のための次世代基盤技術の創出・体系化」領域を今回終了した 6 名の研究者は、ビッグデータ統合利活用という大きな課題に様々な角度から挑戦し、各々独創的な成果をあげた。松谷氏はシステム・アーキテクチャ、佐藤氏と田部井氏は汎用的なアルゴリズムや解析手法、水野氏と宮尾氏はアプリケーション、そして生貝氏は法制度の観点から研究を実施した。

中でも、田部井氏はデータを圧縮したまま検索や機械学習を行う方式を提案し、これを創薬の問題に適用した。また、佐藤氏は統計的潜在意味解析で省メモリかつ高精度の方式を提案し、医用画像の診断に転移学習を適用する研究も行った。両者とも、深みのある理論研究と現実的な応用研究を両立しており、当初予定していた理論研究の枠を超えている。また、松谷氏は、NIC、 FPGA、 GPU を用い、4 種類の NoSQL とバッチ処理及びストリーム処理の省電力化・高速化を行った。FPGA で 4 種類の NoSQL という当初の構想に比べ、は

るかに洗練されたデザインと広がりを達成している。
本領域のシンポジウムは、本領域を含むビッグデータ2領域のCRESTと合同で行う。広範な学術分野の研究者が集まる。さらに、さきがけ研究者による非公式なミーティングも定期的に行われている。さきがけ研究者にとって、異分野の研究者と知り合い、将来に渡って続く関係を築く場となっている。今回終了した6名の中にも既に共同研究を行う関係を築いた者が複数おり、このような関係の構築もさきがけ研究の成果と言えるであろう。

1.生貝　直人　研究者 「ビッグデータ統合利活用のための法制度のあり方に関わる総合的日米欧比較研究による制度設計」
ビッグデータの利活用が大きな価値を生み出すことは論を俟たない。しかし、その価値をデータの利用者が独占し、データの提供者はリスクやコストを負担するのみという構図に陥ると、データの提供が行われなくなり、期待された価値も生み出されなくなる。このような利害の対立を解消あるいは緩和し、社会全体の価値の増大を目指すとともに、不利益を被る関係者が現れないように調停することが、法政策には求められる。
本研究では、パーソナルデータとオープンデータを主な対象に、関連する法制度について日米欧の比較研究を行っている。そして、我が国の法政策に対する示唆も提供している等の研究成果が得られた。パーソナルデータとオープンデータは、ビッグデータ統合利活用の促進が大いに期待されると同時に、単純に市場原理に任せるだけでは停滞が懸念され、さらに、日米欧において今まさに制度や方針が大きく変わろうとしている領域である。したがって、時宜を得た研究と言えるであろう。
法制度に比べ、情報技術の進歩ははるかに早い。時機を逸することなく本研究の成果を実際の政策に反映するための道筋をつけることを期待したい。そして、情報技術の進歩が旧来の法制度の妥当性を脅かすことが予見される領域で、同様の研究を進めることも期待したい。

2.佐藤　一誠　研究者 「統計的潜在意味解析によるデータ駆動インテリジェンスの創発」
ビッグデータを機械学習アルゴリズムの訓練データとして用い、有用な「知能」を人工的に創出する研究が、現在、世界中で活発に行われている。本研究もこのトレンドに沿っているが、「人間のような知能」ではなく、「機械独特の知能」であるデータ駆動インテリジェンスを志向している。そして、人知を持ってしては全体像を掴みきれないビッグデータの分析手法を提案し、医用画像診断などの応用分野も切り拓いている。
論文をトップカンファレンスなどで多数発表しており、優れた学術成果をあげている。特に、統計的潜在意味解析の重要なモデルの一つである Latent Dirichlet Allocation に関しては、ビッグデータのロングテールを捨てることなく、比較的少ないメモリ量で学習可能な方式を提案した。また、民間企業や大学病院と共同で応用研究も行っている。このうち医用画像診断では、学習に使えるデータが少ない病院が、他病院の学習結果を補正利用するために、転移学習の方式を提案し、その有効性を検証している。データの持ち出しが難しい状況でも適用可能な方式として有望である。
要素技術と応用の両面での優れた成果をベースに、今後、体系化や方法論の深化とさらなる応用分野の開拓を行うことを期待したい。

3.田部井　靖生　研究者 「透過的データ圧縮による高速かつ省メモリなビッグデータ活用技術の創出」
ビッグデータの統合利活用のためには大きなメモリを必要とする。その結果、メモリのコストが増加したり、メモリ階層の上位のメモリのヒット率が低下することで、処理性能が劣化したりしやすい。
本研究では、このようなビッグデータの巨大さがもたらす問題の解決策として、データを圧縮した状態で基本演算を高速に実行する方式を提案し、さらに応用分野の実データを用いた実証実験も行っている。主な成果は、類似列が繰り返し現れる場合に有効性が高い文法圧縮技術、圧縮にwavelet木を用いた化合物データベースの類似度検索技術、圧縮したデータ行列を入力とする機械学習技術をあげることができる。これらに関する論文を、トップカンファレンスなどで多数発表しており、優れた学術成果をあげている。これらの成果の中には、バイオインフォマティクスやケモインフォマティクスなどの応用分野の実データを用いた性能評価も含まれる。
今後、ケモインフォマティクスで真に実用に耐える制約の少ない圧縮方式、多様な機械学習アルゴリズムで利用可能な圧縮方式などの研究を継続し、社会的インパクトの大きな成果をあげることを期待したい。

4.松谷　宏紀　研究者 「多様な構造型ストレージ技術を統合可能な再構成可能データベース技術」
ビッグデータの統合利活用のためには、強力なコンピュータやネットワークが不可欠である。一方、ハードウェアを増強する際の制約要因の中で、消費電力の占める割合が、近年、次第に増加しつつある。もちろん、ハ

ードウェア自体のコストも大きな制約条件になる。
本研究は、電力効率とコストパフォーマンスの高いビッグデータ統合利活用を可能とするシステム・アーキテクチャを探求し、その有効性を実証的に示すものである。ビッグデータ取得時に利用されるストリーム処理、蓄積と検索に利用されるNoSQLの処理、解析に利用されるMap-Reduceなどのバッチ処理について、典型的なものを対象に、消費電力削減と性能向上の両立を目指している。そのために、FPGA、 GPU などの既存のアクセラレータを適所適材で組み合わせ、さらに NIC やアクセラレータ内のものを含むメモリ階層の潜在能力を引き出す方式を提案している。そして、これらの分野において、国際会議での最優秀論文賞の受賞や多数の論文の公表など、優れた学術成果をあげている。
本研究の成果である要素技術群と研究の過程で得られた知見が、今後、多様な実アプリケーションのワークロードや要件を反映した統合アーキテクチャの提案として結実することを期待したい。

5.水野　貴之　研究者 「金融ビッグデータによるバブルの早期警戒技術の創出」
現代社会の安定性は経済の安定性に大きく依存している。1990 年代の日本でのバブル崩壊や 2000 年代のリーマンショックを持ち出すまでもなく、いわゆるバブルが社会の安定性に及ぼす影響はきわめて大きい。このバブルは、市場価格が「本来の価格」と乖離することで発生する。今日の主要な市場の中心に、コンピュータで制御された大規模な情報システムが存在することを考えると、バブルの問題に情報学的アプローチで挑戦することには意義がある。
本研究は、経済ビッグデータを解析することにより、バブルの検出、市場における価格変動の引き金の検出、価格変動が伝播する仕組みの解明などを目指すものである。バブルの検出に関しては「本来の価格」を推定する方式、引き金の検出に関しては市場に大きな影響を与える経済ニュースを検出する方式、伝播の仕組みの解明に関してはグローバル・サプライチェーンと企業間の業績連動の分析などを行い、従来「バブルは、はじけて初めてバブルだったとわかる」と言われていた状況から大きな一歩を踏み出すことができた。論文も多数公表されており、優れた学術成果をあげている。
今後は、社会実装を進めることを期待したい。また、本研究では、主に株式市場のバブルを研究対象としていることから、他の市場への展開も期待したい。

6.宮尾　祐介　研究者 「非テキストデータと接続可能なテキスト解析・推論技術の開発」
有史以来、人類が生み出した知識は様々な形態で流通・蓄積されてきた。代表的な形態である書籍や論文は、本文、図、表などで構成されており、電子化されたものだけでも大量に存在する。これらをビッグデータとみなし、本文、図、表などの表現形式の違いを超えて、コンピュータが意味まで読み取って処理を行うことができれば、大きな社会的価値を生み出すであろう。
本研究は、自然言語のテキスト、画像、データベースの表現形式の違いを超えて、統合的に利用するための研究を行うものである。自然言語研究で一般的な意味構造を中心に、テキスト間の含意関係の認識、テキストによるデータベースの問い合わせ、画像からの意味構造の抽出に関する研究を行っている。そして、これらの要素技術を統合して自然言語で問い合わせ可能な画像データベースのプロトタイプも構築している。研究の過程では評価に必要なデータの整備も行っている。これらに関する論文を、トップカンファレンスなどで多数発表しており、優れた学術成果をあげている。
今後、適切な応用分野を設定して、実用化を目指すことを期待したい。また、自然言語処理、画像理解、機械学習などの広範な技術分野に関連する研究テーマであることから、これらの分野の研究者をリードして研究に取り組むことも期待したい。

10. 評価者

研究総括　喜連川　優　国立情報学研究所　所長、東京大学　生産技術研究所　教授
副研究総括　柴山　悦哉　東京大学　情報基盤センター　教授

領域アドバイザー（五十音順。所属、役職は平成 29 年 3 月末現在）

荒川　薫	明治大学　総合数理学部　先端メディアサイエンス学科　教授
石塚　満	東京大学　名誉教授
上田　修功	NTT コミュニケーション科学　基礎研究所　所長
田中　英彦	情報セキュリティ大学院大学　情報セキュリティ研究科　学長 教授
辻井　潤一	産業技術総合研究所　人工知能研究センター　センター長

徳田　英幸　　慶應義塾大学　環境情報学部　教授
徳山　豪　　東北大学大学院　情報科学研究科システム　情報科学専攻　教授
東野　輝夫　　大阪大学大学院　情報科学研究科　情報ネットワーク学専攻　教授
北川　博之　　筑波大学　大学院システム情報工学研究科　教授
山西　健司　　東京大学大学院　情報理工学系研究科　教授
Calton Pu　　Professor、 Georgia Institute of Technology

（参考）
件数はいずれも、平成 29 年 3 月末現在。

（1）外部発表件数

	国　内	国　際	計
論　文	8	38	46
口　頭	60	37	97
その他	14	1	15
合　計	82	76	158

（2）特許出願件数

国　内	国　際	計
1	0	1

（3）受賞等
・佐藤　一誠
日本データベース学会 2014 年度　上林奨励賞（H26）
・松谷　宏紀
・情報処理学会 特選論文（2016）（受賞論文：FPGA NIC 向けノンパラメトリックオンライン外れ値検出機構）（H28）
・Best Paper Award、 The 6th International Symposium on Highly Efficient Accelerators and Reconfigurable Technologies (HEART'15) （受賞論文：A Line Rate Outlier Filtering FPGA NIC using 10GbE Interface）（H27）
・電子情報通信学会 コンピュータシステム研究会 優秀若手講演賞（2014）（受賞論文：カラム指向型データベース向けハードウェアキャッシュ機構の検討）（H26）
・宮尾　祐介
・日本学術振興会 日本学術振興会賞 自然言語の構文解析・意味解析の研究とその応用（H27.2）
・情報処理学会 長尾真記念特別賞 自然言語の深い構文・意味解析の研究とその応用（H26.6）

（4）招待講演
国際　6 件
国内　31 件

別紙

「ビッグデータ統合利活用のための次世代基盤技術の創出・体系化」領域
事後評価実施　研究課題名および研究者氏名

（3年型）

研究者氏名 （参加形態）	研　究　課　題　名 （研究実施場所）	現　職（平成 29 年 3 月末現在） （応募時所属）	研究費 （百万円）
生貝　直人 （兼任）	ビッグデータ統合利活用のための法制度のあり方に関わる総合的日米欧比較研究による制度設計 （東京大学）	東京大学　大学院　情報学環　客員准教授 （情報・システム研究機構新領域融合センター　特任研究員）	8
佐藤　一誠 （兼任）	統計的潜在意味解析によるデータ駆動インテリジェンスの創発 （東京大学）	東京大学　新領域創成科学研究科　講師 （東京大学　情報基盤センター　助教）	40
田部井　靖生 （専任）	透過的データ圧縮による高速かつ省メモリなビッグデータ利活用技術の創出 （東京工業大学）	科学技術振興機構　さきがけ研究者 （東京工業大学　ERATO　湊離散構造処理系プロジェクト　研究員）	29
松谷　宏紀 （兼任）	多様な構造型ストレージ技術を統合可能な再構成可能データベース技術 （慶應義塾大学）	慶應義塾大学　理工学部　情報工学科　講師 （慶應義塾大学　大学院　理工学研究科　講師）	38
水野　貴之 （兼任）	金融ビックデータによるバブルの早期警戒技術の創出 （国立情報学研究所）	国立情報学研究所　情報社会相関研究系　准教授 （同上）	40
宮尾　祐介 （兼任）	非テキストデータと接続可能なテキスト解析・推論技術の開発 （国立情報学研究所）	国立情報学研究所　コンテンツ科学研究系　准教授 （同上）	38

研　究　報　告　書

「ビッグデータ統合利活用のための法制度のあり方に関わる総合的日米欧比較研究による制度設計」

研究タイプ：通常型
研究期間：　平成25年10月～平成29年3月
研　究　者：　生貝　直人

1．　研究のねらい

本研究のねらいは、ビッグデータの利活用を促進するために必要な法制度についての総合的な国際比較研究を行うことにより、今後の我が国における法政策に対する実践的な示唆を見出すことにある。ビッグデータの利活用に関わる法制度は、プライバシーや個人情報の保護、データ自体の知的財産保護、公共セクターの保有するデータの再利用に関わるオープンデータ法制、セキュリティなどをはじめとして、きわめて多岐に渡るものである。

これらのルール全体を、常に変貌と進化を続けるビッグデータの利活用領域において、国家の制定する法制度（政府規制、government-regulation）に詳細に記述することは不可能であり、望ましくもない。ビッグデータの利活用に関わるルールは、第一義的には、イノベーションを主導しそれに関わる詳細な知識を有する、産業界自身の「自主規制（self-regulation）」によって担われるべきである。一方で、個人の尊厳に関わるプライバシーをはじめとする多くの法制度上の問題を、企業や産業界の自主規制に完全に委ねることには、明らかなリスクと限界が存在する。

本研究が重視する「共同規制（co-regulation）」という概念は、このような自主規制の利点を最大限に活かしつつ、そのリスクや限界を政府が法制度等の手段によって補完しようとする、新たなルール形成手法を指す。先に挙げた制度的課題は、いずれを一つ挙げても政府あるいは産業界のみでは解決することができず、共同規制手法を用いた漸進的なルール形成に依らざるを得ないものである。事実ビッグデータの利活用を先導する米国、そしてパーソナルデータ保護の側面を重視するEU、いずれも共同規制手法を念頭に置いたルール形成が進められている。

本研究では、このような公私連携型のルール形成手法を念頭に、「パーソナルデータの保護と利活用」、「オープンデータ政策」という二つの領域に焦点を当て、EU・米国を中心とした諸外国における法制度と公私連携構造を子細に理解することにより、我が国がビッグデータの利活用を促進していくための、柔軟性と確実性を兼ね備えた法制度の設計に関する、実践的な提言を行うための総合的研究を進める。

2．　研究成果

（1）概要

本研究の内容は、大きく「パーソナルデータの保護と利活用」と、「オープンデータ政策」の 2 つのテーマに分けることができる。

「パーソナルデータの保護と利活用」に関しては、2013 年 10 月の研究開始以来、我が国では個人情報保護法の大規模な改正が行われ、またEUでは現行の1995年データ保護指令を全面的に置き換える一般データ保護規則が採択・公布されるなど、パーソナルデータを巡る制度的環境は大きく変化してきている。本研究では特に、我が国を含む世界各国のビッグデ

ータ利活用法政策に与える影響が大きい EU・米国における法政策の正確な理解と、我が国の制度設計に対するインプリケーションを見出すことを主眼とした研究を実施した。具体的には、小テーマとして「共同規制」「データポータビリティ」「プロファイリング」という 3 つの概念に焦点を当て、国内外における制度枠組みの調査と、今後の望ましい制度設計のあり方についての検討を行ってきた。

「オープンデータ政策」に関しては、政府機関や自治体等が保有する膨大なデータを社会・経済的な価値創出につなげるため、我が国においても電子行政オープンデータ戦略の策定等をはじめとして、各種の施策が進められてきているところである。本研究では、EU の公共セクター情報の再利用指令（2003 年成立、2013 年改正）等をはじめとしたオープンデータの基盤となる法政策枠組の他、および我が国ではいまだオープンデータ政策に関わる議論の対象となることが少ない、各種文化施設が保有する文化資源をデジタル化した「デジタルアーカイブ」のオープンデータ化と再利用促進に関わる法政策に焦点を当てた研究を行ってきた。

（2）詳細

研究テーマ A「パーソナルデータの保護と利活用」

主な研究の焦点は次の 3 点である。第一に、近年の各国の法改正における、パーソナルデータに関わる民間事業者が策定する自主規制ルールの位置付けと、それに対する政府関与を行う形の共同規制の制度的枠組についてである。EU では現行のデータ保護指令において、民間団体が各国レベル・EU 全体レベルで策定する自主規制ルール（行動規範）の策定を促し、公的機関が適切性を認めるための規定が置かれ、それに基づいて各国は国内法の整備を行っている。ただし現行制度下においては、それら自主規制ルールに対して法的な確実性を付与するための手続きが明確でないこと、また各国における法制度の相違等を理由として、同規定の活発な利用は行われておらず、また EU レベルで適切性を認められた自主規制ルールは数件に過ぎない。新たに採択された一般データ保護規則においては、EU 全体のデータ保護法制が原則として統一されることに加え、民間の自主規制ルールに対する公的評価の手続きが明確化され、特にその適切性が認められた場合には、委員会が EU 全体での有効性を付与する実施法令を採択できることとし、制度利用のインセンティブを向上させている。また米国においては、従来から連邦取引委員会の関与に基づくパーソナルデータ保護のための自主規制ルール策定が進められてきたが、法的基盤を持たない自主規制ルールの不十分性などを背景として、ルール策定の段階から政府機関が関与する形でプロセスの適正化を図り、また策定されたルールに対して公的な評価を行う、共同規制型アプローチの構築と実践が進められている。本研究ではこれらの制度枠組みについての記述と比較検討を行いつつ、我が国における個人情報保護分野の共同規制枠組である認定個人情報保護団体制度等に対するインプリケーションを得るための研究を行った（論文 1、その他 2）。

第二に、個人の意思に基づくデータの流通や利活用を可能とするデータポータビリティのあり方である。一般データ保護規則は、全体として現行のデータ保護指令の枠組みを踏襲・強化したものだが、新たなアプローチとして 20 条にデータポータビリティの権利（The Right to Data Portability）の権利が導入されている。同権利は、個人が企業等に提供した自らの個人データを、構造化された、一般的に用いられる、機械可読なフォーマットで受け取り他の企業

等に移転することと、ある企業等から別の企業等に直接的に移転することを、個人の権利として認めるものである。欧州委員会は同権利の実現により、個人が自らのデータを簡易にサービス間で移転可能となることで、個人情報を利用するサービス間の競争が加速されプライバシー親和的なサービスが競争優位を得られるようになること、大企業が保有する個人データにスタートアップがアクセスしやすくなることなどを指摘している。本研究では、一般データ保護規則における規定の策定プロセスや理論的背景に加え、英国をはじめとするEU各国、ならびに米国等において先行的に実施されているデータポータビリティ関連制度についての調査を行うことにより、①データポータビリティの対象となるべきデータの範囲や権利行使の条件、②他者の権利（プライバシーや知的財産）に対する悪影響の抑止、③データ標準の策定に関わる公私連携手法のあり方等の論点を中心に、データ利活用と権利保護を両立する望ましい制度設計のあり方について検討を行ってきた。

第三に、プロファイリングに関わる規律のあり方である。個人の行動履歴等を解析するプロファイリングとそれに基づく高度なパーソナライズド・サービスは、ビッグデータの社会的活用のあり方として高いポテンシャルを有するが、一方でプライバシーへの影響や社会的差別といった問題への対応を検討する必要性を提起する。一方でプロファイリングはパーソナルデータ活用に関わる広範な活動に関係するため、法規制のあり方には慎重な検討を要する。本研究では、EU 一般データ保護規則において導入されたプロファイリング規制と、米国において主にデータブローカー事業者が行うプロファイリングに関わる規制政策を主な対象とした調査を行い、データの利活用と個人の権利保護を両立しうるプロファイリング規制のあり方について、①透明性の確保、②プロファイリング行為自体に対する個人の異議申立て、③プロファイリングに基づく決定に対する個人の異議申立て、という3つの観点から検討を行ってきた。

研究テーマB「オープンデータ政策」

オープンデータ政策については、特に公的な文化施設（美術館・博物館・図書館・文書館等）が保有する文化資源をデジタル化して公開する、デジタルアーカイブの再利用促進に焦点を当てた研究を行った。ここ数年来、我が国では国や自治体が保有する各種の統計情報や行政文書等を再利用可能な形で公開するためのオープンデータ政策は急速に進められているものの、文化施設保有データのオープンデータ化は取り上げられることが少なかった。この点 EU においては、欧州全体の文化施設が保有・公開するデジタルアーカイブのポータル・プラットフォームであるヨーロピアナを中核とした、デジタルアーカイブの再利用促進施策を進めてきている。本研究では文献調査や現地調査などを通じて、これまで研究蓄積の少ない当該分野のオープンデータ政策の枠組みを明らかにすることを目的として、特に①2013 年の改正によりオープンデータ原則の対象機関に文化施設を含むこととした公共セクター情報再利用指令の枠組と各国における国内化状況、②参加文化施設とヨーロピアナの間で締結されるデータ交換協定をはじめとした、ヨーロピアナを中心として構築されるデジタルアーカイブの再利用促進枠組、③パブリック・ドメイン（著作権保護期間満了）作品の利活用促進に関わる制度的・実務的課題と解決方法についての詳細な調査と検討を行った（論文 3、4）。これらの研究の中で、デジタルアーカイブのオープンデータ化に関しても、基盤となる制度枠組みを政府レベルで整備しつつ、各文化施設分野や民間団体等が具体的な実施ルールを策定・運用す

る共同規制構造が機能していることを明らかにしてきた。また合わせて、米国全土を対象としたプラットフォーム構築を行う米国デジタル公共図書館との比較研究を行った。

これと並行して、再利用促進の前提条件となる文化資源のデジタル化・公開に必要な著作権制度のあり方について、主に欧州・米国における①文化施設における著作物のデジタル化に関わる権利制限規定、②権利者不明の孤児作品の利用円滑化に関わる制度整備に焦点を当てた研究を行い、我が国の法政策との比較を含む論文等を出版した（論文2、その他5）。

3. 今後の展開

これまでの研究結果を元に、それぞれの研究の理論的深化と、研究・比較対象地域の拡充、そして新たな技術革新に対応した制度設計に向けた研究を進めていく。

研究テーマ A「パーソナルデータの保護と利活用」については、国内外の新制度下における共同規制メカニズムの形成プロセスを継続的に研究対象としていく他、今回の研究で対象としなかった、必ずしもビッグデータ利活用に限られないシェアリングエコノミーにおけるプラットフォーム規制等の新たな規制政策領域での実態を調査し、より分野横断的な研究を進めていく。データポータビリティの研究に関しては、本研究で主として対象としたEU・英国・米国についての研究に加え、個人データのみに限られない広範なデータを対象としたポータビリティ制度の導入を進めつつあるフランス等における議論を研究対象に含めると共に、競争法や消費者保護法の観点からの分析を視野に入れた理論的深化を進めていく。

研究テーマ B「オープンデータ政策」に関しては、本研究で主に対象としたデジタルアーカイブについて、オープンデータ政策という観点に加え、デジタル時代における総合的な知識基盤を構築するための法政策のあり方に射程を広げていく。

4. 評価

（1）自己評価

本研究は、ビッグデータの「利活用」という側面に焦点を当て、我が国において研究蓄積の少ない共同規制やデータポータビリティ、デジタルアーカイブの再利用などの新領域について、現地調査等の手段を通じて制度の全体像を把握し、先駆的な検討を行うことができたという点において、我が国におけるビッグデータ利活用のための制度的基盤構築に一定の貢献を行うことができたと考える。一方で、国内外の制度動向が急変する時期であることなどから、海外制度の記述・理解のための研究に多くのリソースを割くことになり、具体的な制度設計の提案に踏み込む形での研究成果を十分に出すことはできておらず、今後本研究を元にした政策提言等への展開を進めていきたい。

（2）研究総括評価（本研究課題について、研究期間中に実施された、年2回の領域会議での評価フィードバックを踏まえつつ、以下の通り、事後評価を行った）。

ビッグデータの利活用が大きな価値を生み出すことは論を俟たない．しかし，その価値をデータの利用者が独占し，データの提供者はリスクやコストを負担するのみという構図に陥ると，データの提供が行われなくなり，期待された価値も生み出されなくなる．このような利害の対立を解消あるいは緩和し，社会全体の価値の増大を目指すとともに，不利益を被る関係者が現れないよ

うに調停することが，法政策には求められる．

本研究では，パーソナルデータとオープンデータを主な対象に，関連する法制度について日米欧の比較研究を行っている．そして，我が国の法政策に対する示唆も提供している．パーソナルデータとオープンデータは，ビッグデータ統合利活用の促進が大いに期待されると同時に，単純に市場原理に任せるだけでは停滞が懸念され，さらに，日米欧において今まさに制度や方針が大きく変わろうとしている領域である．したがって，時宜を得た研究と言えるであろう．

法制度に比べ，情報技術の進歩ははるかに早い．時機を逸することなく本研究の成果を実際の政策に反映するための道筋をつけることを期待したい．そして，情報技術の進歩が旧来の法制度の妥当性を脅かすことが予見される領域で，同様の研究を進めることも期待したい．

5．　主な研究成果リスト

（1）論文（原著論文）発表

1. 生貝直人．インターネットの自主規制・共同規制．ドイツ憲法判例研究会（編）『憲法の既判力とメディア法』．2015, pp.63-85, 信山社．
2. 生貝直人．文化芸術デジタルアーカイブと著作権―総合芸術アーカイブセンターにおける実践と比較法的観点からの覚書―．東京藝術大学社会連携センター紀要．2015, vol.1, .17-31.
3. 生貝直人．デジタルアーカイブと法政策：統合ポータル、著作権、全文検索．大学図書館研究．2016, no.106, pp.11-18.
4. 生貝直人．ナショナルデジタルアーカイブの条件について．金沢 21 世紀美術館研究紀要．2016, no.6, pp.6-14.

（2）特許出願

なし

（3）その他の成果（主要な学会発表、受賞、著作物、プレスリリース等）

1. 生貝直人．デジタルアーカイブと利用条件．国立国会図書館カレントアウェアネス．2014, no.322, pp.8-12.
2. 生貝直人．共同規制－ルールは誰が作るのか．藤城裕之編『ソーシャルメディア論：つながりを再設計する』．2015, pp.193-204, 青弓社
3. 生貝直人．ウェブサービス・プラットフォームの事例：ヨーロピアナとナショナルデジタルアーカイブ．日本図書館情報学会研究委員会編『メタデータとウェブサービス』．2016, pp.181-195, 勉誠出版
4. 生貝直人．オープンなデジタルアーカイブに向けた日米欧の法政策．慶應義塾大学 DMC 紀要．2016, vol.3, no.11, pp.5-12.
5. 生貝直人．デジタルアーカイブと著作権に関する国内外の動向．図書館界．2016, vol.67, no.6, pp.346-352.

研　究　報　告　書

「統計的潜在意味解析によるデータ駆動インテリジェンスの創発」

研究タイプ：通常型
研究期間：　平成26年10月～平成29年3月
研　究　者：　佐藤　一誠

1．　研究のねらい

個人が日常的に触れる情報量が膨大になり、一個人に関係する情報に限ったとしても、その全容を知ることがすでに不可能な状況に我々は直面している。しかし、創造的な活動は、普段組み合わせて考えない物事の組み合わせや日常注目していなかった物事に目を向けるなど、多様な情報に対して横断的な思考を行うことが重要である。本研究では、科学技術によって人の知の営みを拡張することを目的とする。この目的を達成するアプローチとして、「データ駆動インテリジェンス(英：Data-driven Intelligence)」という概念により1つの見通しを与える。「データ駆動インテリジェンス」は、「人工知能(英: Artificial Intelligence)」や「人間の知能(英:human Intelligence)」に並ぶ第 3 の知のカテゴリとして期待されている。 本研究では、人の知能と計算機による知能の交差する領域ではなく、データが誘発する計算機特有の知能である「データ駆動インテリジェンス」を創発するアルゴリズム開発を目的とした。データ駆動インテリジェンスを創発する基盤技術として統計的潜在意味解析に着目する。統計的潜在意味解析が利用する基本情報は「共起性」である。事象間の共起性を捉えることで、単純な頻度の数え上げでは発見できない事象間の関係性を抽出することができる。特に、データ上実際に共起した表層的なパターンだけではなく、潜在的に共起するパターンも含まれる。この潜在的共起情報が新たな知の創発につながると考えられる。本研究では特に「カウントデータ」および「医用画像データ」という異なる2つの性質を持つデータに着目し研究を進めた。多くのカウントデータで共通に現れる問題として、特定のデータが高頻度で出現し、その他の多様なデータが低頻度であるという、いわゆるロングテール現象を扱うアルゴリズムの開発を行った。医用画像データ解析では、解析対象である病変データ数は正常データ数よりも圧倒的に少ない状況で効率的に学習するためのアルゴリズム開発を行った。

2．　研究成果

（1）概要

研究成果として主に以下の3つのテーマについて説明する。
研究テーマ A 「大規模ロングテールデータにおける LDA の学習アルゴリズム開発」
研究テーマ B 「ロングテールデータにおける Population bias を緩和するアルゴリズム開発」
研究テーマ C 「医用画像読影支援システム開発」
研究テーマ AB はカウントデータにおけるロングテール現象に対する学習アルゴリズムに関する研究であり、研究テーマ C は医用画像解析に関する研究である。共通する問題として興味に対象とする自称の頻度が非常に低いことが挙げられる。したがって、データ全体としては大規模であるものの重用であると考えられるデータは非常に少ない状況下での学習をテーマと

している。

（2）詳細

研究テーマ A 「大規模ロングテールデータにおける LDA の学習アルゴリズム開発」

カウントデータにおいて統計的潜在意味解析を行う最も重用なモデルの1つとして Latent Dirichlet Allocation (LDA)がある。LDA は、文書や閲覧履歴をカウントデータとして扱ったときに、単語やアイテムの共起現象をモデル化したものである。潜在変数と呼ばれる確率変数をモデルに組み込むことで、潜在的な共起を扱うことができる。このような潜在的な共起は、文書の潜在的なトピックやユーザの潜在的な嗜好を表現する情報としてさまざまな応用分野で用いられている。

LDA の大規模データからの学習アルゴリズムとして、データ集合から部分集合をサンプリングし、部分データからの学習を繰り返すことで、実質的にデータ数に依存しない学習手法が注目を集めている。しかし、従来の研究では、部分データを用いたとしても、部分データ毎の学習において、パラメータの更新に必要な計算量が全データの次元に依存してしまうという問題がある。そのため、カウントデータにロングテールの性質がある場合、テール部分に該当する単語やアイテムを削除することで計算量を予め減らし学習するのが主流であった。しかし、このような前処理をしてしまうと、テール部分に該当するデータが分析対象であったときに、その潜在意味解析を行うことができず、有用な情報を得ることができない可能性がある。

本研究では、部分データにおける学習時に、部分データに現れるデータの種類数のみに依存したアルゴリズムを開発した。これによりテール部分を前処理として削除せずに大規模ロングテールデータにおいて LDA を学習することができる。さらに提案手法は、カウントデータにおける統計モデルの汎化（予測）能力を測る指標である Perplexity において、state-of-the-art である学習アルゴリズムにおいて 5 倍程度効率的なアルゴリズムであることも実験的にわかった。具体的には、state-of-the-art のアルゴリズムが達成する Perplexity をを20%程度の学習データで達成することができた。

研究テーマ B 「ロングテールデータにおける Population bias を緩和するアルゴリズム開発」

リクナビ 2015 のユーザの閲覧行動履歴を解析すると、企業に対する閲覧頻度にロングテール現象があり、特に就職活動初期においては、特定の企業のみを閲覧し企業に関する知識が偏っている可能性があることがわかった。その結果として、最終的に就職する企業に行き着くまでに必要以上にコストがかかり、就職活動が長期化する原因の1つと考えられる。実際に就職活動を円滑に進めているユーザは、就職活動初期においても様々な企業の情報を取得することに長けており、そのユーザに合った企業選びができる傾向にある。このような閲覧履歴にみられる閲覧数の偏りは、カウントデータにしばしばみられるロングテール現象であり、Population bias と呼ばれている。このような偏りを解消するシステムが必要である。

就職活動初期において、ユーザに合った企業を推薦する推薦システムが有用であると考えられるが、推薦システムを構築する際にもこのデータの偏りは問題となる。推薦システムを構築するためには、過去の閲覧履歴データから学習する必要があるが、通常推薦システムで使われる学習アルゴリズムを用いるとデータ数の多いアイテムの推薦精度を重要視することになる。したがって、本研究で対象とする閲覧頻度の低いアイテムを推薦する目的には使うこと

ができない。また、通常、購買情報や5つ星評価のようにユーザの興味が明示的に与えられたデータにおける推薦システムとは異なり、閲覧履歴はユーザの嗜好が明示的には得られない。前者は explicit feedback、後者は implicit feedback と一般的に呼ばれている。本研究では、implicit feedback である閲覧履歴を用いて、explicit feedback である将来的なエントリーを予測する問題であり、この問題設定も従来とは異なる。

本研究で開発したアルゴリズムは、ユーザ・企業それぞれの嗜好を表す潜在変数を学習する際に Popularity effect を表現するコスト関数を設計し、コスト付き最適化問題として学習アルゴリズムを導出する。さらに、ユーザ・企業がもつ様々な情報（出身地、大学、学部、心理テスト結果、TOEIC の点数、留学の有無、所在地、企業規模などなど）を事前知識として組合せることで、このような偏りを緩和するモデルを構築した。構築したモデルは、5つのハイパーパラメータを保持することから、閲覧履歴を用いた交差検証をガウス過程によるベイズ的最適化で自動化した。問題となるのは、今回の目標は最終的にはエントリー予測であるが、エントリーデータをシステムの構築段階では使うことができないため交差検証で用いることができない。そこで、閲覧履歴における1回以上の閲覧予測によってベイズ的最適化交差検証を行ったところ、エントリー予測の精度と高い相関があることがわかった。つまり、エントリーデータを用いること無く閲覧履歴のみを用いてシステムを構築すればエントリー予測の精度を高めることができることがわかった。従来手法にくらべ、2倍程度の精度向上が確認された。

研究テーマ C 「医用画像読影支援システム開発」

医用画像診断装置が多くの医療現場に導入され、検査数はここ数年で増加している。

検査で撮影した画像から病変の有無を判断し診断を行う読影は、読影医と呼ばれる医師によって目視により行われ、東大病院では一検査あたり10～30 分で読影を行い、診断を行うことが現状求められている。しかし、医用画像診断装置の技術進展によって、1 検査あたり300～500 枚の画像を撮影できるようになっており、その時間内で読影するのは、時間的制約が非常に厳しくなりつつあるというのが問題として認識されている。また、医用画像診断では、病変あり（陽性）の判断が行われる件数は、病変なし（陰性）の判断が行われる件数に比べ非常に少ないという性質があるため、医師は、画像をより正確に診断し、数少ない重大な病変を見落とすことのないようにすることが求められる。つまり、医師は短時間で、重大な病変を見落とすことなく診断するという、非常に難しい要求を迫られている。

このような背景の下、我々は機械学習を用いた Computer-Assisted Detection （CAD） ソフトウェア と呼ばれる医用画像読影支援システムを開発している。

医用画像読影支援システムでは、読影後にシステムが病変部位の特定を行い、医師へ特定箇所を提示する支援システムである。医師は提示された病変部位を元に、最終的な診断をする。このような支援システムは、医師の負担を軽減するだけでなく、病変を見落とす危険性も低減させることができる。本研究では、東大病院で臨床実用中の医用画像読影支援システムを他の病院で運用する際の制約や問題点を指摘し、その解決策を提示する。

機械学習を用いて支援システムを構築する際に問題となるのが病変データの希少性である。医療の現場で用いられるシステムを作るためには性能のよい学習アルゴリズムが必須であるが、そのための学習データを作るコストが高い。本研究では、日常診療の読影過程で病

変ラベル付きデータを蓄積することができる機能を支援システムに構築した。医師が日常診療で読影する際に、病変を見つけた場合、その部分を簡単にマーキングすることで、病変ラベルの定義が容易にできる機能を構築した。これにより他施設が支援システムを導入することで、そこで新たに病変ラベル付きデータが蓄積され学習に反映される機能を追加することができる。

他の病院が支援システムを導入する場合、導入初期に十分な病変ラベル付きデータを作成するのは時間的にもコスト的にも難しいため、既存のシステムで用いられているデータを学習に利用することが運用上求められる。しかし、病院間における医用画像診断装置の違いによりデータの性質が変化し、そのままデータを用いたのでは性能悪化が起こる可能性がある。また、そもそも医療におけるデータは個人情報であるため病院間でデータを共有することは難しい。

このように学習データの少ない領域で学習する際に、他の利用可能な領域における学習データを用いることで学習効率を上げる手法は転移学習と呼ばれている。

転移学習を読影支援システムで利用する場合、病院間における医用画像診断装置や設定の違いによりデータの性質が変化し、そのまま学習データを用いたのでは性能劣化が起こる可能性がある。このように転移学習させた結果、学習器の性能を劣化させてしまう現象は「負の転移」と呼ばれている。

本研究では、各々の施設での学習器の出力を転移することで病院間でデータの共有をせずに、「負の転移」に関する性能保証も与える学習する手法を開発した。具体的には、読影支援システムの性能評価に用いられる症例ベースの AUC(Area Under Curve)において、転移したことによる性能の劣化の上限を抑えることで「負の転移」に関する性能保証付き転移学習を可能にした。 また、我々の手法は出力を転移することから異なる学習器間の転移も可能である。例えば、現在の読影支援システムでは Adaboost という学習器が動いているが、畳込みニューラルネットを用いた場合でも Adaboost と畳込みニューラルネット間において転移が可能である。

現在の読影支援システムでは、脳動脈瘤の読影支援を中心としており、脳動脈瘤に関する特徴量を専門家の知見をもとに作成し Adaboost を用いて識別を行っている。この性能自体は現在臨床実験中で臨床に耐えうるものだと判断できるが、将来的に異なる症例に関して読影支援システムを構築する際には、症例ごとに特徴量設計することは非現実的である。そこで、畳込みニューラルネットに基づく読影支援システムの開発も行った。畳込みニューラルネットではボクセルデータそのものを入力として学習するため特徴量設計の必要がない。開発したアルゴリズムは、現行の Adaboost の性能を超えるものであることがわかった。畳込みニューラルネットでは Adaboost とは異なり数多くのハイパーパラメータが存在するため、本研究ではベイズ的最適化による自動チューニングシステムの開発も行った。通常の畳込みニューラルネットを用いる問題設定とは異なり病変検知は、病変データと正常データとのデータ数の偏りが大きいため、単純な適用ではうまくいかない。本研究ではデータ拡張、AUC 最適化に基づく学習、ベイズ的最適化を組み合わせることでこのような問題を解決した。

3. 今後の展開

LDA は応用範囲が広く様々な研究分野で使われているモデルである。近年で1細胞解析など全く事なる分野でも用いられるためこの領域の研究者が大規模なデータを解析する際のアルゴリズムとしてコラボレーションすることが考えられる。東大病院と共同開発しているシステムは、「日常診療におけるデータの取得」と「取得されたデータから機械学習によって病変検知を行うシステムの実験自動化（ハイパーパラメータ自動チューニング）」から成り立つが，この枠組みは病理診断にも応用できると考えている．医用画像読影支援システムは、現在脳動脈瘤に特化しているが、症例の種類数を増やすとともに、細胞診断へ広げるプロジェクトを立ち上げている。

4．　評価

（1）自己評価

統計的潜在意味解析を可能にする代表的なモデルである LDA は引用数が15000件を超えるモデルであり、実に様々な応用研究がなされている。今回基礎研究として開発したアルゴリズムは計算量および汎化能力の点でも最も優れたアルゴリズムであり、応用研究への波及効果は高いものと考えられる。また、応用研究として開発した読影支援システムでは、システムを様々な施設で用いる際の問題点を数理的な背景のある手法を提案することで解決することができた。基礎研究としても負の転移に関する理論保証付きのアルゴリズムは無いため、理論面・応用面で大きな貢献ができたと考えられる。今回開発したシステムやアルゴリズムをより多くの症例・多くの施設で展開することで、国内における医療診断の様々な問題の解決につながると考えられる。

（2）研究総括評価（本研究課題について、研究期間中に実施された、年2回の領域会議での評価フィードバックを踏まえつつ、以下の通り、事後評価を行った）。

ビッグデータを機械学習アルゴリズムの訓練データとして用い，有用な「知能」を人工的に創出する研究が，現在，世界中で活発に行われている．本研究もこのトレンドに沿っているが，「人間のような知能」ではなく，「機械独特の知能」であるデータ駆動インテリジェンスを志向している．そして，人知を持ってしては全体像を掴みきれないビッグデータの分析手法を提案し，医用画像診断などの応用分野も切り拓いている．

論文をトップカンファレンスなどで多数発表しており，優れた学術成果をあげている．特に，統計的潜在意味解析の重要なモデルの一つである Latent Dirichlet Allocation に関しては，ビッグデータのロングテールを捨てることなく，比較的少ないメモリ量で学習可能な方式を提案した．また，民間企業や大学病院と共同で応用研究も行っている．このうち医用画像診断では，学習に使えるデータが少ない病院が，他病院の学習結果を補正利用するために，転移学習の方式を提案し，その有効性を検証している．データの持ち出しが難しい状況でも適用可能な方式として有望である．

要素技術と応用の両面での優れた成果をベースに，今後，体系化や方法論の深化とさらなる応用分野の開拓を行うことを期待したい．

5．　主な研究成果リスト

（1）論文（原著論文）発表

1. Issei Sato, Hiroshi Nakagawa. Stochastic Divergence Minimization for Online Collapsed Variational Bayes Zero Inference of Latent Dirichlet Allocation. The 19th ACM International Conference on Knowledge Discovery and Data Mining (KDD2015). 2015, pp.1035-1044.
2. Issei Sato, Hiroshi Nakagawa, Approximation Analysis of Stochastic Gradient Langevin Dynamics by using Fokker-Planck Equation and Ito Process. The 31st International Conference on Machine Learning (ICML 2014). 2014, pp. 982-990.
3. Issei Sato, Hisashi Kashima, Hiroshi Nakagawa. Latent Confusion Analysis by Normalized Gamma Construction. The 31st International Conference on Machine Learning (ICML 2014). 2014. pp. 1116-1124.
4. Masahiro Kazama, Issei Sato, Haruaki Yatabe, Tairiku Ogihara, Tetsuro Onishi, Hiroshi Nakagawa. Company Recommendation for New Graduates via Implicit Feedback Multiple Matrix Factorization with Bayesian Optimization. The 2016 IEEE International Conference on Big Data. 2016.
5. 佐藤 一誠, 野村 行弘, 林 直人. オンライン転移学習と医用画像読影支援への応用. 日本応用数理学会. 2016.

（2）特許出願

研究期間累積件数：1件

1.

発　明　者: 佐藤 一誠、 石黒 勝彦

発明の名称: 収束判定装置、方法、及びプログラム

出　願　人: 東京大学

出　願　日: 2014/02/28

出 願 番 号: 2014-039036

（3）その他の成果（主要な学会発表、受賞、著作物、プレスリリース等）

受賞

日本データベース学会 2014 年度　上林奨励賞

著作物

・佐藤一誠(著)、 奥村学(監修):トピックモデルによる統計的潜在意味解析、 コロナ社、 2015 年 3 月

・佐藤一誠(著):ノンパラメトリックベイズ、講談社、 2016 年 4 月

研　究　報　告　書

「透過的データ圧縮による高速かつ省メモリなビッグデータ活用技術の創出」

研究タイプ：通常型
研究期間：　平成25年10月～平成29年3月
研　究　者：　田部井　靖生

1．　研究のねらい

近年の様々な情報処理分野において，データは高度に大規模化・複雑化している.データは主に文字列，木，グラフで表現される．大規模化に伴いこのようなデータを従来の表現で取り扱うことは困難になりつつあり，データ中に潜む有意義な情報を抽出する手法の研究開発が，現代社会における緊急の課題となっている.

このような状況から，データから統計的な情報をもとにデータの背後に潜む規則を自動的に抽出するデータマイニング及び機械学習の研究が近年盛んに行われている．従来の研究に共通することは，　入力の規模をあまり考慮しないで研究が行われているために，近年の大規模化・複雑化するデータに必ずしも適応可能であるとは限らない.

透過的データ圧縮とは，データを圧縮した状態のまま定数時間でランダムアクセスを可能にする圧縮方式のことで，近年，文字列，木，グラフを含む様々なデータ表現のための透過的データ圧縮法が盛んに研究されている．代表的な応用例として ローマ字として入力されたひらがな単語を漢字に変換する Input Method Editor (IME)があり，IME では各ひらがな単語を漢字に対応付けるために大規模なトライとして表現される辞書が必要となる．　通常トライのポインターを用いた標準的な実装では，トライの各ノード当たり log(n)ビット(n:全ノード数)のメモリが必要となり，全ノードで nlog(n)ビットのメモリが必要で，数億からなる日本語の単語を登録するには適していない．代表的な IME である GoogleIME では，木の代表的な透過的データ圧縮法 LOUDS を用いることにより，トライの各ノードを 2 ビットで表現している．全単語を登録するためのトライのサイズはわずか 2n ビットである．さらにトライをコンパクトに表現することは，キャッシュアクセスによる高速化の役割も担っている．　このように透過的データ圧縮はデータ処理における省メモリ化と高速化の両方において有効であるにもかかわらず，これまでの研究では代表的なデータ構造の操作をいかに圧縮したまま実現するかに焦点があてられており，より高度な処理であるデータ中に潜む有意義な情報を抽出する処理に関しては手法が開発されていないのが現状である.

本研究のねらいは，複雑な構造を持つ膨大な量のデータから，その背後に潜む有益な規則を自動的に見つけ出す高速かつ省メモリなデータ処理手法を透過的データ圧縮法に基づき開発することである．本研究プロジェクトでは，バイオインフォマティクス，ケモインフォマティクスなどの分野に存在するビッグデータを対象として，普遍的なデータ処理課題を抽出し，その課題を数理的な立場から問題を定式化する．そして，ビッグデータ処理に適応可能な高速かつ省メモリなデータ処理手法を開発する.

2. 研究成果

(1)概要

本研究課題における成果は主に 3 つの領域に分けられる. (i) 大規模反復テキスト処理のための文法圧縮法, (ii) 大規模化合物データベースの類似度検索法, (iii) データ圧縮技術によるスケーラブルな機械学習法に分けられる. いずれの成果も最新のデータ圧縮技術に関するものである.

(2)詳細

研究テーマ A「大規模反復テキスト処理のための文法圧縮法」

文法圧縮とは, 与えられたテキストを一意に表現する文脈自由文法を構築することで圧縮する技術である. 文法圧縮は, 繰り返しの多いテキストに対して高い圧縮率を達成するが可能である. そのようなテキストの実例としては, DNA 配列やバージョン管理されたテキストなどがある.特に DNA 配列の個体間での違いは 0.1%ほどと言われており, 文法圧縮は DNA 配列に対して高い圧縮率を達成することが期待できるので, DNA 配列を処理するための有効な手段として期待されている. 本研究プロジェクトでは, Edit Sensitive Parsing (ESP)と呼ばれる文法圧縮アルゴリズムをテキストビッグデータ利活用のために, (i)スケーラブルなオンライン文法圧縮技術, (ii) 文法圧縮されたテキストの類似度検索, (iii) 文法圧縮されたテキスト上での rank, select, access 操作などの様々なテキスト処理アルゴリズムを開発した.

研究テーマ B「大規模化合物データベースの類似度検索技術」

大規模化合物データベースの類似度検索は, 創薬における重要なタスクである. データベース中の各化合物は, 0 または 1 を要素とするバイナリーベクトルで表現される. 化合物のバイナリーベクトル表現はフィンガープリントと呼ばれ, フィンガープリント表現の類似度検索技術も数多く提案されてきた. 近年, ディスクリプターと呼ばれる化合物の整数ベクトル表現が提案され注目を集めている. 代表的なディスクリプターに ringo や KCF-S があるが, ディスクリプター表現された化合物の類似度検索技術はまだないのが現状である. そこで, 本研究課題では, ディスクリプターとして表現された化合物データベースの類似度検索技術を wavelet 木などの透過的データ圧縮技術を応用することで開発した. 本手法により, 米国国立生物工学情報センターの化合物データベース PubChem 中の 4 千万化合物に対して, 高速に類似度検索を行うことを可能にした.

研究テーマ C「データ圧縮技術によるスケーラブルな機械学習法」

解釈可能な統計モデルを学習させることは, 大規模データから有益な情報を抽出するための有効な手段の一つである. 機械学習では, 学習データは特徴ベクトルを要素として持つデータ行列として表現され, データ行列を入力として統計モデルの学習が行われる. しかし, 学習データが大規模になる時, 学習には大量のメモリが必要になり統計モデルの学習が困難になる傾向がある. このような問題に対して, 我々はデータ圧縮されたデータ行列上でのスケーラブルな統計モデル学習法を開発した. 提案手法により約 100GB のデータ行列を 4GB にまで圧縮した状態で統計モデル学習が可能となった. 提案手法の応用として, 創薬におけるバーチャルスクリーニングの応用を行った. バーチャルスクリーニングは, 化合物とタンパク質の

相互作用予測として定式化され，大規模機械学習問題となる傾向にある．提案手法により大規模な化合物とタンパク質のデータからでも効率よくモデル学習が可能となることを示した．

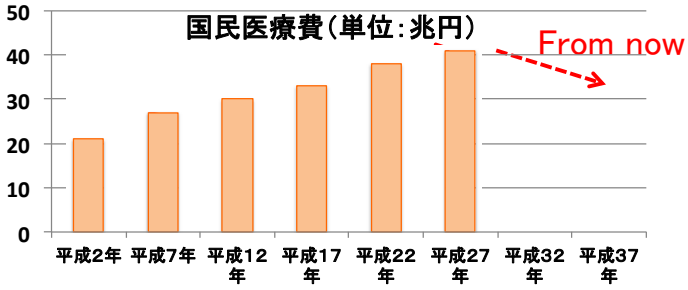
3．　今後の展開

開発したデータ圧縮技術はビッグデータ処理の様々な場面で有効であると考えられる．そこで，考えられる今後の展開は開発した技術を様々なビッグデータ処理に適応していくことである．研究テーマ B の類似度検索に関しては，要求されている制約条件も複雑になってきおり本研究課題で開発した手法だけでは不十分である．そこで，今後はより複雑な制約条件にも適応可能な類似度検索技術を開発する予定である．研究テーマ C の「機械学習応用」に関しては，開発したデータ圧縮技術を様々な機械学習法に適応していくことが考えられる．また，本研究課題では，機械学習アルゴリズムのメモリ削減に焦点を当てて手法を開発したが，機械学習アルゴリズムの速度を向上する研究の方向性も考えられる．機械学習アルゴリズムの高速化はビッグデータからの統計モデル学習において有効な手段と考えられるので，今後，開発して行く予定である．

4．　評価

（1）自己評価

研究当初の目的は概ね達成した．研究の進め方に関しては，研究補助者を見付けられなく，研究を加速させることができなかったことが残念な点である．本研究プロジェクトで開発した様々な技術は新規創薬の場面で有効な技術ある．現在，新規薬の開発コストの増加に伴い薬の価格も増加している．さらに，薬の価格の増加と高齢社会により国民医療費も増加していることが現代社会における重要な問題となっている．本研究課題で開発した技術は創薬におけるバーチャルスクリーニングの効率を上げることが可能であるので，薬の低価格化，さらには，国民医療費を下げることが期待できる．



（2）研究総括評価（本研究課題について、研究期間中に実施された、年2回の領域会議での評価フィードバックを踏まえつつ、以下の通り、事後評価を行った）。

ビッグデータの統合利活用のためには大きなメモリを必要とする．その結果，メモリのコストが増加したり，メモリ階層の上位のメモリのヒット率が低下することで，処理性能が劣化したりしやすい．

本研究では，このようなビッグデータの巨大さがもたらす問題の解決策として，データを圧縮した状態で基本演算を高速に実行する方式を提案し，さらに応用分野の実データを用いた実証実験も行っている．主な成果は，類似列が繰り返し現れる場合に有効性が高

い文法圧縮技術，圧縮に wavelet 木を用いた化合物データベースの類似度検索技術，圧縮したデータ行列を入力とする機械学習技術をあげることができる．これらに関する論文を，トップカンファレンスなどで多数発表しており，優れた学術成果をあげている．これらの成果の中には，バイオインフォマティクスやケモインフォマティクスなどの応用分野の実データを用いた性能評価も含まれる．

今後，ケモインフォマティクスで真に実用に耐える制約の少ない圧縮方式，多様な機械学習アルゴリズムで利用可能な圧縮方式などの研究を継続し，社会的インパクトの大きな成果をあげることを期待したい．

5. 主な研究成果リスト

（1）論文（原著論文）発表

研究テーマA「大規模反復テキスト処理のための文法圧縮法」

1. Djamal Belazzougui, Patrick Coding, Simon J. Puglisi, Yasuo Tabei, Access, Rank and Select in Grammar-compressed strings, In Proceedings of the 23rd European Symposium on Algorithms (ESA), 2015.

2. Shirou Maruyama and Yasuo Tabei,, Fully-online Grammar Compression in Constant Space, In Proceedings of Data Compression Conference, 2014.

3. Yoshimasa Takabatake, Yasuo Tabei, Hiroshi Sakamoto, Online Self-indexed Grammar Compression, In Proceedings of the 22nd edition of the International Symposium on String Processing and Information Retrieval (SPIRE), 2015

4. Djamal Belazzougui, Travis Gagie, Paweł Gawrychowski, Juha Kärkkäinen, Alberto Ordóñez, Simon J. Puglisi, Yasuo Tabei: Queries on LZ-Bounded Encodings, In Proceedings of the Data Compression Conference (DCC), 2015.

5. Yoshimasa Takabatake, Yasuo Tabei, Hiroshi Sakamoto: Improved ESP-index: a practical self-index for highly repetitive texts, 13th International Symposium on Experimental Algorithms (SEA), 2014.

研究テーマ C「データ圧縮技術によるスケーラブルな機械学習法」

1. Yasuo Tabei, Hiroto Saigo, Yoshihiro Yamanishi, Simon J. Puglisi: Scalable partial least squares regression on grammar-compressed data matrices, In Proceedings of the 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2016.

2. Yasuo Tabei, Yoshihiro Yamanishi, Masaaki Kotera, Simultaneous Prediction of Enzyme Orthologs from Chemical Transformation Patterns for De Novo Metabolic Pathway Reconstruction, In Proceedings of the 23rd International Conference on Intelligent Systems for Molecular Biology (ISMB), 2016.

3. Yoshihiro Yamanishi, Yasuo Tabei, Masaaki Kotera: Metabolome-scale de novo pathway reconstruction using regioisomer-sensitive graph alignments, In Proceedings of ISMB/ECCB, 2015.

4. Masaaki Kotera*, Yasuo Tabei*, Yoshihiro Yamanishi*, Ai Muto, Yuki Moriya, Toshiaki Tokimatsu, Susumu Goto: Metabolome-scale prediction of intermediate compounds in

multi-step metabolic pathways with a recursive supervised approach, 22nd Annual International Conference on Intelligent Systems for Molecular Biology (ISMB), 2014.

（2）特許出願

なし

（3）その他の成果（主要な学会発表、受賞、著作物、プレスリリース等）

・ Succinct Data Structure for Scalable Knowledge Discoveries, チュートリアルセッション, The 20th Pacific Asia Conference on Knowledge Discovery and Data Mining (PAKDD), 2016 年 4 月 19 日(火)

・ 文法圧縮の理論と実践, 第 27 回 RAMP シンポジウム, セッション「離散構造とアルゴリズム」, 2015 年 10 月 15 日(木)

・コンパクトなデータ表現による機械学習, 第 14 回情報科学フォーラム(FIT), イベント企画, ビッグデータ解析のための機械学習技術, 2015 年 9 月 17 日(木)

・透過的データ圧縮法による高速かつ省メモリーなビッグデータ活用技術の創出, 第 77 回情報処理学会全国大会, CREST・さきがけ「ビッグデータ」2 領域 成果報告会, 2015 年 3 月 18 日(火)

・ Dictionary based compression for processing massive genome sequences, ゲノムテクノロジー164 委員, 2014 年 12 月 18 日(木)

・ 透過的データ圧縮法による高速かつ省メモリーなビッグデータ活用技術の創出, ビッグデータ時代に向けた革新的アルゴリズム基盤, 京都リサーチパーク, 2014 年 1 月 11 日(土)12 日(日)

研　究　報　告　書

「多様な構造型ストレージ技術を統合可能な再構成可能データベース技術」

研究タイプ：通常型
研究期間：　平成 25 年 10 月～平成 29 年 3 月
研 究 者：　松谷　宏紀

1．　研究のねらい

科学技術分野において日本の国際競争力を高めるには大量データ（ビッグデータ）処理技術と機械学習ベースの AI（Artificial Intelligence）技術の普及と浸透が不可欠である。これらを活かすには本来莫大な量の計算リソースが必要であり、現状では、巨大なデータセンタを管理、運用する限られた企業がそのような計算プラットフォームを握っている。このような状況において、大量データおよび AI 利活用のための敷居を下げ、これらの技術革新の恩恵を最大限享受できるようにするには、計算プラットフォームの低消費電力化および低コスト化が必須である。そこで、本研究では、大量データ処理や機械学習に関するオープンソースプロダクトのフレームワークを維持しつつ、FPGA（Field-Programmable Gate Array）や GPU（Graphics Processing Unit）のような汎用アクセラレータを利活用することに着目した。実際、FPGA や GPU の利活用によって得られた性能向上の分だけ計算リソースを削減できるため、これらのアクセラレーションは大量データ処理や機械学習処理の低消費電力化および低コスト化に大きく寄与する。

アクセラレーション対象の 1 つ目として、大量データの蓄積と検索を司るデータベース技術に着目する。具体的には、キーバリューストア（KVS）型、カラム指向型、ドキュメント指向型、グラフ型データベースなど多様なデータベースに対し、FPGA や GPU のようなアクセラレータをどのように使うべきかを探求する。アクセラレーション対象の 2 つ目はストリーム処理であり、無限に生成される時刻順データに対する計算処理に着目する。具体的には、オンラインの外れ値検出などストリームデータに対する各種機械学習アルゴリズムを対象に FPGA を用いたアクセラレーションを行う。3 つ目はバッチ処理である。10GbE（10Gbit Ethernet）経由で接続された GPU クラスタを用いて既存のバッチ処理フレームワークを高速化する。当初はアクセラレーション対象をデータベース処理に絞っていたが、最終的には既存のラムダアーキテクチャの要素技術を網羅的にカバーできるようアクセラレーション対象を広げて行った。広範囲に渡るアクセラレーション事例を通して、大量データ利活用に向けたアクセラレーション戦略の指針を示すことが本研究のねらいである。

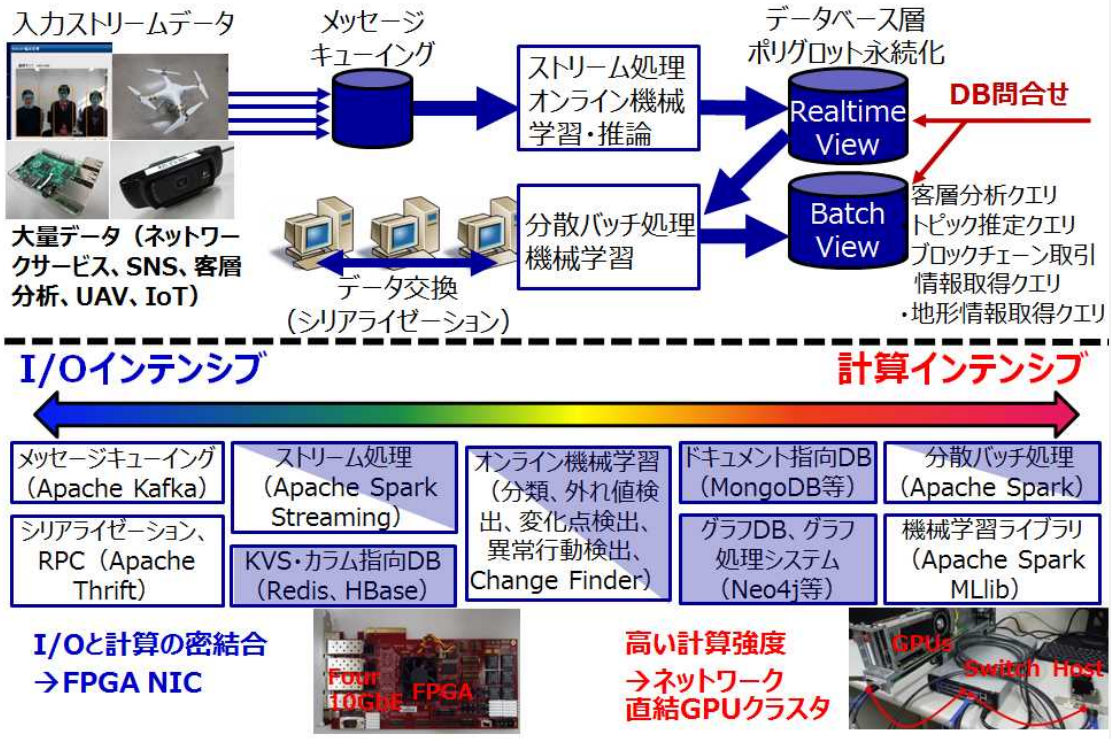
2．　研究成果

（1）概要

大量データ利活用のための統合システムの一例を次ページの図に示す。上図の左端から入力されたストリームデータはストリーム処理やバッチ処理によって加工もしくは集計され、データベースに蓄積されている。①データの収集では、ネットワークサービスや IoT（Internet of Things）デバイスによって生成された大量データを集め、データベース（Realtime View）を更新

する。データの受信と配送にはメッセージキューイングシステム、データベースの更新処理にはストリーム処理フレームワークが利用される。外れ値検出や変化点検出のようなオンライン機械学習も行われる。②データの集約では、1 日に数回など定期的に全データを解析し、集計結果をデータベース(Batch View)に格納する。このためにバッチ処理フレームワークおよび機械学習フレームワークなどが利用される。また、分散処理のために計算機間でデータを交信する際にはシリアライゼーション処理が必要になる。③データの蓄積と検索は、上述の Realtime View と Batch View を格納するデータベースに相当する。Realtime View には時々刻々と更新される直近のデータ、Batch View には定期的に集計される過去の集計済み全データが格納される。ユーザからの問い合わせに対し、Realtime View と Batch View の検索結果を結合したものを返答することで全データ性および即応性の両立を図る。

以上が本研究で想定する大量データ利活用に向けた統合システムである。このためのデータ検索、データ収集、データ集約処理など広範囲に渡るトピックを FPGA や GPU のような汎用アクセラレータを用いていかに高効率化するかという点を探求した。以下にその詳細を示す。



上図：大量データ処理向け統合システムの一例。下図：そのための要素技術と計算強度。

(2)詳細

研究テーマ A「データ検索のアクセラレーション」

KVS は、データをキーとバリューの組として扱う非常にシンプルなデータベースである。計算負荷が低いため、ネットワーク処理がボトルネックになりやすい。そこで、ネットワーク処理と計算を密結合できるデバイスとして FPGA ベースのネットワークインタフェース(FPGA NIC と呼ぶ)を用いた KVS のアクセラレーションを研究した。業績[1]では 10GbE インタフェースを 4 個有する NetFPGA-10G ボードを用いて KVS の一種である Redis のハードウェアキャッシュを実現した。

ただし、FPGA ボードに搭載できる DRAM 容量は大きくはないため、業績[1]では FPGA ベースのハードウェア KVS キャッシュに加え、Linux カーネル内にソフトウェアベースの KVS キャッシュを設けることを提案した。我々は前者を L1 NoSQL キャッシュ、後者を L2 NoSQL キャッシュと呼び、要求されたデータが L1 NoSQL キャッシュに存在しない場合は L2 NoSQL キャッシュをルックアップし、L2 NoSQL キャッシュにもヒットしない場合はアプリケーション層で動作する Redis に問い合わせが行くようにした。

カラム指向型では、各行データは複数個のカラムから構成され、行キーを基にソートされた状態で扱われる。このため HBase では startRow と stopRow を用いた範囲問い合わせも可能である。カラム指向型も KVS 同様、ネットワーク処理がボトルネックとなりやすいため、我々は HBase を対象に FPGA NIC を用いたアクセラレーション手法[23]や Linux カーネル内キャッシュを用いたアクセラレーション手法[9]を提案してきた。

ドキュメント指向型ではデータをドキュメントの集合、グラフ型データベースではデータをグラフとして扱う。前者では正規表現ベースの文字列探索、後者ではグラフ探索を伴う問い合わせが生じる。これらの処理は一般的に計算負荷が高いため、単一 GPU を用いたドキュメント指向型データベース MongoDB の高速化[8]、グラフ型データベース Neo4j の高速化[3]を提案してきた。ただし、データベースで扱うデータに比して GPU のデバイスメモリはあまりにも小さいため、GPU とデータベース間でデータ交信が頻発し、性能の新たなボトルネックとなっていた。そこで、GPU 内にキャッシュできるデータ容量をスケーラブルに増強するために、我々は PCI-Express over 10GbE 技術によってネットワーク接続された GPU クラスタを用いるアプローチを提案している[7]。業績[7]では多数の GPU のデバイスメモリを分散共有メモリとして扱い、分散ハッシュテーブルから着想を得たハッシュ技法によってデータを各 GPU のデバイスメモリに分散させている。

研究テーマ B「データ収集のアクセラレーション」

主要な要素技術はストリーム処理フレームワークとオンライン機械学習である。これらの処理は総じて計算負荷は低めであり、ネットワーク処理と計算を密結合できる FPGA NIC を用いたアプローチが有利であると考えている[2,5,6]。

オンライン機械学習処理の代表例として異常検出が挙げられる。異常検出はさらに外れ値検出、変化点検出、異常行動検出などに分類できる。例えば、ネットワークから流れてくるサンプルデータに対し、通常とは異なる値、傾向、振る舞いなどを検出するために利用できる。我々はこのうち外れ値検出を FPGA NIC で高スループット化する研究を行ってきた[2,6]。業績[2]では、マハラノビス距離を用いて外れ値を検出する手法を FPGA NIC 上に実現し、10GbE ラインレートの 95.8%のスループットを実現した。業績[6]では、k-Nearest Neighbor（kNN）や Local Outlier Factor（LOF）アルゴリズムを FPGA NIC 上に実現する方法を提案している。LOF による外れ値検出では、過去のサンプルデータと入力サンプルデータの比較が必要だが、当然、FPGA NIC の限られた DRAM に過去の全サンプルデータを保持することはできない。そこで、最近アクセスされた過去のサンプルデータクラスタのみを FPGA NIC にキャッシュしておき、一方で、FPGA NIC にキャッシュされている限られた情報だけでは外れ値かどうか判別できないような入力データについてはアプリケーション層にて完全な LOF 処理を行うアプローチを提案した。

ストリーム処理フレームワークに関しては、まず、One-at-a-time 方式と Micro-batch 方式に

大別できる。One-at-a-time 方式ではデータ要素 1 つ 1 つに対し決められた処理を適用するのに対し、Micro-batch 方式では短い時間間隔の間に到着したデータ要素をまとめて 1 つの Micro-batch とし、この Micro-batch 毎に決められた処理を適用する。前者の実例として Storm、後者の実例として SparkStreaming が挙げられる。通常のソフトウェア処理の場合、高機能な処理を One-at-a-time 方式で実現しようとすると計算負荷が高くなり過ぎる。一方、Micro-batch 方式では Micro-batch のサイズに応じて外れ値や変化点などのイベントを検出するまでの遅延が増大してしまう。そこで、我々は One-at-a-time 処理と Micro-batch 処理を組み合わせる 2 段階ストリーム処理を研究している。具体的には、Micro-batch 方式の SparkStreaming に対し、FPGA NIC 上に実現した One-at-a-time 処理を組み込むアプローチを提案した[5]。

研究テーマ C「データ集約のアクセラレーション」

バッチ処理フレームワークとして我々は Hadoop/MapReduce に加え Spark にも注目している。Spark では RDD（Resilient Distributed Dataset）と呼ばれる分散共有メモリ上にデータを保持し、RDD から別の RDD を生成する処理（Transformation）や RDD を集約する処理（Action）が行われる。

バッチ処理は計算負荷が高くなりやすく、GPU によるアクセラレーションが向くことが多い。業績[4]では Spark の RDD に対する Transformation 処理や Action 処理を GPU にオフロードしているが、「研究テーマ A」でも述べたとおり、これだけでは GPU と Spark の間のデータ転送が新たなボトルネックとなる可能性がある。そこで、業績[4]では業績[7]と同じアプローチを採用している。具体的には、PCI-Express over 10GbE 技術によってネットワーク接続された GPU クラスタを前提に、GPU のデバイスメモリに RDD をキャッシュしておく。GPU のネットワークトポロジによっては CPU から近い GPU、遠い GPU が生じてしまうが、業績[4]では RDD の系統グラフを基に頻繁にアクセスされる RDD は近い GPU、それ以外は遠い GPU にキャッシュするアイディアを提案している。

3. 今後の展開

短期的な計画として、既存のラムダアーキテクチャの要素技術のうち本研究によってアクセラレーションできていない処理のアクセラレーション指針をまずは確立したい。具体的には、2 ページ目の図のメッセージキューイングミドルウェアや RPC（Remote Procedure Call）、シリアライゼーションに関する部分である。アクセラレーションのための指針はすでにあり、研究も開始している。これらのアクセラレーションに関する既存研究はほとんど存在しないため、インパクトのある成果になると期待している。また、本研究では 10GbE インタフェースを有する FPGA NIC を用いて研究を行ったが、100GbE 版の FPGA NIC が入手でき次第、100GbE に移行する予定でもある。

本研究では、多数のセンサや IoT デバイス、世界規模のネットワークサービスによって生成される大量データが絶え間なく通過するネットワークに着目し、FPGA NIC 上に各種機械学習アルゴリズムを高性能ハードウェアとして実現した。このアプローチはネットワークを流れるデータからパターンを学習し、知識を得るという用途一般に応用できる。本研究では主として外れ値検出アルゴリズムに焦点を当てたが[2,6]、現在では、いくつかの変化点検出アルゴリズムを対象に研究を開始するなど対象範囲を広げている。NIC やスイッチという特殊、かつ、厳し

いリソース制約のもと、実用的な機械学習アルゴリズムをいかに専用ハードウェアとして実現するかはさらに深く研究すべき課題である。実際、NIC やスイッチへの機械学習機能のオフローディング、さらに、そのような NIC やスイッチによる分散協調学習は、エッジヘビーコンピューティングを実現するうえでのキーテクノロジーに成り得ると考えている。

4. 評価

（1）自己評価

当初計画では、アクセラレーションは対象をデータベースのみで、使用するアクセラレータは FPGA のみとしていた。これでは 2 ページ目の図のごく一部分しかカバーできない。その後、研究を遂行していく過程でアクセラレーション対象はラムダアーキテクチャの要素技術の大半をカバーするまでに広がり、FPGA NIC や FPGA スイッチ[1,2,5,6,9,21,22,23]、GPU [3,4,7,8,12]、PCI-Express over 10GbE [4,7]、Linux カーネル内キャッシュ[9]など様々なアクセラレーション戦略を適材適所で使い分けるようになった。このため、研究期間の途中から「大量データ利活用のための各種要素技術を網羅的にカバーすべくアクセラレーション事例を蓄積し、大量データ利活用に向けたアクセラレーション戦略の指針を示す」ことを目指すべく、研究目標を緩やかに変更していった。上述の「今後の課題」で言及したとおり、まだまだ研究しなければならないアクセラレーション課題は残っているものの、3 年半という研究期間を鑑みるに最善を尽くすことができたものと考えている。今後は、NIC やスイッチへの機械学習機能のオフローディング、さらに、そのような NIC やスイッチによる分散協調学習などエッジヘビーコンピューティングに向けた要素技術群をハードウェア屋の観点から探求して行きたい。

（2）研究総括評価（本研究課題について、研究期間中に実施された、年2回の領域会議での評価フィードバックを踏まえつつ、以下の通り、事後評価を行った）。

ビッグデータの統合利活用のためには，強力なコンピュータやネットワークが不可欠である．一方，ハードウェアを増強する際の制約要因の中で，消費電力の占める割合が，近年，次第に増加しつつある．もちろん，ハードウェア自体のコストも大きな制約条件になる．

本研究は，電力効率とコストパフォーマンスの高いビッグデータ統合利活用を可能とするシステム・アーキテクチャを探求し，その有効性を実証的に示すものである．ビッグデータ取得時に利用されるストリーム処理，蓄積と検索に利用される NoSQL の処理，解析に利用される Map-Reduce などのバッチ処理について，典型的なものを対象に，消費電力削減と性能向上の両立を目指している．そのために，FPGA, GPU などの既存のアクセラレータを適所適材で組み合わせ，さらに NIC やアクセラレータ内のものを含むメモリ階層の潜在能力を引き出す方式を提案している．そして，これらの分野において，国際会議での最優秀論文賞の受賞や多数の論文の公表など，優れた学術成果をあげている．

本研究の成果である要素技術群と研究の過程で得られた知見が，今後，多様な実アプリケーションのワークロードや要件を反映した統合アーキテクチャの提案として結実することを期待したい．

5. 主な研究成果リスト

（1）論文（原著論文）発表

1. Yuta Tokusashi, Hiroki Matsutani, "A Multilevel NOSQL Cache Design Combining In-NIC and In-Kernel Caches", Proc. of the 24th IEEE International Symposium on High Performance Interconnects (Hot Interconnects 24), pp.60-67, Aug 2016.
2. Ami Hayashi, Yuta Tokusashi, Hiroki Matsutani, "A Line Rate Outlier Filtering FPGA NIC using 10GbE Interface", ACM SIGARCH Computer Architecture News (CAN), Vol.43, No.4, pp.22-27, Sep 2015.
3. Shin Morishima, Hiroki Matsutani, "Performance Evaluations of Graph Database using CUDA and OpenMP-Compatible Libraries", ACM SIGARCH Computer Architecture News (CAN), Vol.42, No.4, pp.75-80, Sep 2014.
4. Yasuhiro Ohno, Shin Morishima, Hiroki Matsutani, "Accelerating Spark RDD Operations with Local and Remote GPU Devices", Proc. of the 22nd IEEE International Conference on Parallel and Distributed Systems (ICPADS'16), pp.791-799, Dec 2016.
5. Kohei Nakamura, Ami Hayashi, Hiroki Matsutani, "An FPGA-Based Low-Latency Network Processing for Spark Streaming", Proc. of the 4th IEEE International Conference on Big Data (BigData'16) Workshops, pp.2410-2415, Dec 2016.
6. Ami Hayashi, Hiroki Matsutani, "An FPGA-Based In-NIC Cache Approach for Lazy Learning Outlier Filtering", Proc. of the 25th Euromicro International Conference on Parallel, Distributed, and Network-Based Processing (PDP'17), 8 pages, Mar 2017.
7. Shin Morishima, Hiroki Matsutani, "Distributed In-GPU Data Cache for Document-Oriented Data Store via PCIe over 10Gbit Ethernet", Proc. of the 22nd International European Conference on Parallel and Distributed Computing (Euro-Par'16) Workshops, 12 pages, Aug 2016.
8. Shin Morishima, Hiroki Matsutani, "Performance Evaluations of Document-Oriented Databases using GPU and Cache Structure", Proc. of the 13th IEEE International Symposium on Parallel and Distributed Processing with Applications (ISPA'15), pp.108-115, Aug 2015.
9. Korechika Tamura, Hiroki Matsutani, "An In-Kernel NOSQL Cache for Range Queries Using FPGA NIC", Proc. of the 1st International Conference on FPGA Reconfiguration for General-Purpose Computing (FPGA4GPC'16), pp.13-18, May 2016.

（2）特許出願

なし

（3）その他の成果（主要な学会発表、受賞、著作物、プレスリリース等）

招待講演等

10. 松谷 宏紀, "ビックデータ利活用のための計算基盤", 電子情報通信学会 コンピュータシステム（CPSY）研究会, 招待講演, Dec 2016.
11. 鯉渕 道紘, 松谷 宏紀, 藤原 一毅, "大規模コンピュータ・ネットワークの建築学", 国

立情報学研究所 H28 年度第 2 回 産官学連携塾, Oct 2016.

12. Hiroki Matsutani, "Accelerator Design for Various NOSQL Databases", The 16th International Forum on MPSoC for Software-defined Hardware (MPSoC'16), Invited Talk, Jul 2016.
13. Hiroki Matsutani, "Accelerator Design for Various NOSQL Databases", Big Data French-Japanese Workshop, The Embassy of France in Japan, Invited Talk, Nov 2014.
14. 松谷 宏紀, "ビッグデータ向け計算機アーキテクチャの研究動向と研究事例", インターネットコンファレンス 2014 (IC'14), 招待講演, Nov 2014.
15. 松谷 宏紀, "ポリグロット永続化のための NoSQL アクセラレータ", 情報処理学会 データベースシステム(DBS)研究会, 招待講演, Aug 2014.
16. 松谷 宏紀, "多様な構造型ストレージ(NOSQL)のためのアクセラレータ設計", 日本電気株式会社, 招待講演, Jul 2014.

受賞等

17. 情報処理学会 特選論文 (2016)
    (受賞論文:FPGA NIC 向けノンパラメトリックオンライン外れ値検出機構)
18. Best Paper Award, The 6th International Symposium on Highly Efficient Accelerators and Reconfigurable Technologies (HEART'15)
    (受賞論文:A Line Rate Outlier Filtering FPGA NIC using 10GbE Interface)
19. 電子情報通信学会 コンピュータシステム研究会 優秀若手講演賞 (2014)
    (受賞論文:カラム指向型データベース向けハードウェアキャッシュ機構の検討)

研　究　報　告　書

「金融ビッグデータによるバブルの早期警戒技術の創出」

研究タイプ：通常型
研究期間：　平成25年10月～平成29年3月
研　究　者：　水野　貴之

1．　研究のねらい

日本政府と日本銀行はデフレ脱却と持続的な経済成長を実現するために 2013 年に異次元の金融緩和を実行した結果，金利・日経平均・円が上昇し，バブル崩壊を経験した日本企業は，これらの金融市場の反応が持続的な経済成長の芽ではなく，バブルの芽ではないかという懸念から，設備や人的投資に躊躇した[The wall street journal(May-15, 2013)]．バブルの芽でなければ，このような消極的な行動はデフレ脱却の足枷にしかならない．また，近年は中国バブル，世界的に広がる金融危機が，しばしば発生している．このような背景からも，バブルを察知するシステムを構築することは，社会にとって重要なミッションである．

バブルに対するマクロ経済学における共通認識は，前米 FRB 議長アラン・グリーンスパンが 2002 年に述べた「バブルは、はじけて初めてバブルだったとわかる」という言葉が指し示している通り，バブルがはじける前に，バブルであるか否かを見極める手法が確立していない．

本研究では，ビッグデータを用いて，はじけないとバブルが分からない現状を打開する研究を推し進め，バブルのモニタリングと崩壊を予測する技術を構築することにある．具体的には，金融ビッグデータ（全世界 500 社を越える通信社が発表するニュース記事・主要金融市場の取引注文・先進各国の企業の財務諸表と取引先リスト）を利活用することで，バブルに対して，これまでデータ不足により実証論的な結果から理論を構築するアプローチが出来なかった問題点を解決する．また，横断的なアプローチ（自然言語処理・統計物理学・統計学・機械学習・マクロ経済学・計量経済学）により，各分野が苦手にしている分析を他分野の分析手法で補う．これらにより，一部の企業に投機マネーが集中するバブルの特徴抽出，投資家の金融市場のトレンドを追随する同調行動を強化させるバブルの芽となるニュースの機械的な検出，グローバル・サプライチェーンを通じたバブル崩壊ショックの波及のダイナミクスの解明をおこない，早期にバブル警報を発信する技術を創出する．この技術を搭載したシステムが完成すれば，バブルを引き起こす過剰投機を監視し，金融引締め等のマクロ経済政策をデータドリブンで実行できる．この研究を通じて，日本政府が掲げ，また世界各国が望むバブルのない持続的な経済成長の実現に科学的に貢献する．

2．　研究成果

（1）概要

「金融ビッグデータによるバブルの早期警戒技術の創出」を目指して，はじめに，「（1）経済バブルを検出する（ナウキャスト）技術」を構築した．バブルとは，本来の価値以上に投機により価格が上昇するである．本来の価値を正確に見積もることは困難であるが，ビッグデータを使えば，例えば，不動産市場や株式市場では類似した物件や企業が多数見つけられるため，類似したモノは同じ価値，一物一価の法則により，似た物件間や企業間での価格の違い

の拡大から特定の物件や企業への過剰投機を検出し，バブルのナウキャストがおこなえるようになった[論文 1，主要な学会発表 2，主要な学会発表 3].

次に，大きな株価変動が，しばしばニュースをキッカケとして起こるために，そのようなニュースを機械的に察知する「（2）重要なニュースの検出技術の開発」を進めた．過去のニュース記事と単語の類似が少なければ，ニュースに新規性があり，同時刻における多数のメディアで類似する単語が数多く使われれば，そのニュースは話題性が高いと判断できる．株価の反応は新規性や話題性に強く依存しており，これら新規性と話題性を用いて市場にとって重要なニュースをリアルタイムに検出できる．また，人は株価が下がる出来事ほど，多くの情報を欲しがり，メディアはニュースを供給する傾向があることを見出した．この傾向を応用して，ニュース数の急増から，株価の下降トレンドの発生を検出する技術を開発した．[論文 2，主要な学会発表 1，主要な学会発表 5]

各企業における株価のショックは，企業間の繋がりを通じて全世界に広がるため，「（3）グローバル・サプライチェーンの基本構造の解明」をおこなった．企業は国境を越えて同業種でコミュニティを形成し，そして，そのコミュニティ間を橋渡しする企業がいることで，各企業が世界中の企業と平均6取引先で繋がるというスモールワールド性が成り立っていることを見出した．ショックの世界的な伝搬を防ぐには，ショックを特定の業種内に封じ込める政策が有効であることが明らかになった．この結果は，グローバル・サプライチェーンにおける紛争鉱物や奴隷労働問題等の繋がることによるリスクの対策にも使える．[論文 3，論文 4，論文 5，主要な学会発表 4，著作物 1]

（2）詳細

研究テーマ A「経済バブルの検出技術の確立」

株式市場では，投資家は企業の業績にもとづいて株価を提示する．市場には多くの上場企業が存在し，事業内容や業績の似た企業も存在する．類似した企業では，通常，株価が高い企業の株は売られ，安い方の株が買われる一物一価の法則により，株価は似た値となる．しかし，しばしばマネーゲームが発生して，業績度外視で株価が過度に高値まで釣り上がることがある．この状態がバブル期には発生する．類似企業間での株価（正確には時価総額＝株価×発行済株数）の差を観測することによって，同じ業界で同じ業績なのに，片方だけ株価が高い状態から株式バブルを検出する技術を開発した．

世の中に全く同じ企業は存在しないため，非バブル期においても類似する企業間で株価の差は存在する．そのため，非バブル期における類似企業間の株価の散らばりを分布で観測し，その散らばりを超えるような大きな価格差が頻発しているかどうかで，バブルかどうかを判断した．

企業の業績を表す財務項目は，100 近く存在し，また，しばしば，売上の高い企業は利益も高いように，財務項目間には強い相関が存在する．はじめに，非バブル期である 2004 年の株価に最も反映する財務項目を，ランダムフォレストを用いて探した．次に，財務項目をランダムフォレストから得られる各項目の重要度で重み付けし，企業ごとに財務状況が最も似ている企業とで，株価の対数価格差を計算し，対数価格差の分布を描いた．年度ごとに分布を観測すると，IT バブルが始まる前の 1997 年では，類似する財務状況で株価が 10 倍高くなってい

る企業の割合は全体の 1.5%程度であった．一方で，バブル期の 1999 年では分布の裾野が太り，5%も存在していた．そして，はじけた後の 2004 年には，1997 年と同じ分布に戻った．このように類似する企業間の対数価格差の分布の変化を非バブル期と比較しながら観測することにより，リアルタイムに株式市場のバブルを検出できる．研究室 Web サイトでは，各株式市場のバブル度を日々掲載している．

研究テーマ B「重要なニュースの検出技術の開発」

機関投資家が利用する Thomson Reuters の情報端末 Eikon に全世界 500 社を超える通信社から配信された年間 100 万記事を超えるニュース記事の記事間の関係を使って，リアルタイムにニュースの新規性と話題性を測定する手法を開発し，新規性と話題性の高いニュースには市場が大きく反応することを示した．また，人間が持つネガティブバイアスを利用して，ニュースの記事数から予想より悪い出来事を察知して，今後の株価トレンドが予測できることを発見した．

ニュースの新規性と話題性は，各ニュース記事に含まれている単語の数をベクトルの要素に持つ bag-of-words で表現して，2つの記事の類似度をベクトル間のコサインで数値化することによって定義した．このとき，Stop-words は，文章の内容を表現する上で適切な単語ではないために，IDF 値を使ってベクトルの要素に重み付けをした．ニュースが新規であるかどうかは，過去のニュースとの類似度が低いかどうかで判定できる．一方，ニュースの話題性は，同時刻における別の通信社が配信したニュースとの類似度が高いかどうかで判定できる．重要な出来事ほど，各社が競い合って報道するので，同時刻に同じようなニュースが集中する．

記者によって各上場企業のフラグが付けられたニュースに注目し，その上場企業の株価の反応（出来高，取引数，株価変動幅）との関係を調査した．株価の反応は新規性や話題性に強く依存しており，これら新規性と話題性を用いて市場にとって重要なニュースをリアルタイムに検出できることが明らかになった．

ニュースの記事数の急増から，株価の下降トレンドの発生を察知できることを明らかにした．投資家は株価が下がってしまう出来事には敏感になり，メディアは，そのような出来事についてニュースを多く供給する傾向にあることを証明した．米国株についての日々のニュース数に注目し，ある日に急にニュースが増えた場合には，株価が下がる出来事が発生したと仮定して，次の日の朝に米国の株価指数 S&P500 に収録されている銘柄を売り，その１週間後に買い戻すという仮想取引の損益を調査した．この損益分布は統計的に有意に利益の方に偏っており，ランダムな取引での１σ程度の利益を頻繁にあげていることが見出された．つまり，ニュースの急増から下降トレンド発生の注意報が出せる．

研究テーマ C「グローバル・サプライチェーンの基本構造の解明」

General Motors に関する重要なニュースが発生した直後，General Motors の主な部品供給元である American Axle & Manufacturing Holdings の株も General Motors 株での反応の 40%程度の反応が起きる．つまり，企業間の関係性を通じて，個別企業のショックが他企業に伝搬している．関係性を通じたショックの伝搬予測や，その伝播に対する対策を示すために，グローバルな企業間ネットワークの基本構造と企業間の業績の連動の関係を解明した．

世界中のほぼ全ての上場企業と投資家の関心を集めている未上場企業約 50 万社を対象に，企業間のグローバルな仕入販売，ライセンス契約，資本提携について、それぞれの関係が作るネットワーク構造を調査した．企業間の繋がりは硬直的であり，50%以上の取引関係は 3 年以上継続し，一度取引関係が失われたとしても，約 10%程度が数年以内に復活する．つまり，ネットワークを通じた企業間の業績の相関は持続しやすい．グローバルな企業間ネットワークでは 90%以上の企業ペアがネットワークを通じて繋がっている．そして，その平均経路長は短く，僅か 6 取引先である．平均経路長は短いが，各企業が直接様々な企業と広く取引しているのではなく，企業は企業連合と言うべきコミュニティを，国境を越えて同業種内で構成している．多くの企業はコミュニティ内の企業とのみ密に繋がっており，そして，コミュニティ間を橋渡しするハブ企業が，いくつか存在することによって，企業のグローバリゼーションが形成されている．

上場企業間で日次の株価変動の連動度合いを相互相関関数により計測し，その相互相関係数を，グローバル・サプライチェーンにおける企業間の最短経路長ごとに調査した．サプライチェーンを通じて繋がっている企業同士には，距離が近ければ近いほど，正の相関が強くなっている．この特徴が，各企業で発生したショックが，グローバル・サプライチェーンを通じて次々と他の企業に連鎖していき，株価に影響を与えることを意味することを，ページランクを応用したネットワーク上の流れを記述するモデルから明らかにした．

ネットワークの基本構造を解明し，ネットワーク上の伝搬を再現するモデルを構築したことによって，ショック伝搬は，はじめにショックが起きた企業と同じ業界内に世界的に広まり，その後に，世界各地の他業種に広まるという経路をたどることが分かった．業種間を繋ぐ企業でショックを止めれば，他国も含み他業種へのショックの伝搬がほとんど防げることが明らかになった．

これらの結果は，グローバル・サプライチェーンにおける紛争鉱物の流通問題にも応用できる．紛争地で採掘された鉱物は，まずは，世界的に商社や鉱物精錬企業で流れるために，これらの業種を対象にした紛争鉱物の対策が効果的である．試算すると G8 各国における上場企業で対策をした場合，コンゴ産紛争鉱物の除去は G8 内で3割程度に過ぎないが，上場企業と同数の鉱物精錬企業や商社，採掘企業での対策をすれば9割以上が除去できる．しかも，世界が業種で繋がっているため，G8 地域以外の紛争鉱物も減らすことができる．

3． 今後の展開

株式市場のバブルの検出手法を，住宅地や商業地の不動産市場のバブル検出に順次応用していき，同時に社会実装も進めていく．本研究では，1つの株式市場内で生じている企業間の株価の歪みからバブルの検出をおこなったが，市場内の全ての株が過剰投機状態（全ての株価が歪んでいる状態）にあるとき，バブルの検出精度が悪くなる欠点がある．この問題点は，複数市場での企業間の株価の歪みを観測することにより解決できる．この解決は現在取り組んでいる研究課題で解決される．重要なニュースの検出技術については，既存のニュースに対する自然言語処理の技術に，そのまま組み込むことで精度向上が可能であるために実装を進めている．グローバル・サプライチェーンの研究成果については，関連する業界団体や NGO と社会実装に向けて連携を模索している．

4. 評価

（1）自己評価

研究目的の達成状況

「金融ビッグデータによるバブルの早期警戒技術の創出」という研究の当初の目的は，金融ビッグデータを用いて株式市場や不動産市場のバブルをナウキャストする技術を構築できたことから達成できている．早期警戒技術の創出の次は，バブルの予測や崩壊の危険性の推定であるが，これらについても，ニュースの研究やサプライチェーンの研究で十分な成果が出ている．

研究の進め方（研究実施体制及び研究費執行状況）

金融ビッグデータも，他のビッグデータと同じく，統一感のないフォーマットにイレギュラーなデータ，重複するデータに一部スパースなデータなど，スマートデータ化をしなければ実証分析に利用できない．このスマートデータ化には，手間と時間がかかり，この部分を研究補助者に担当してもらえたことが研究を進める上での鍵になった．十分な研究実施体制を築けた．研究費の執行状況は，当初の計画通りに進めた．

研究成果の科学技術及び社会・経済への波及効果（今後の見込みを含む）

はじけないとバブルが分からない経済学の現状を，金融ビッグデータを用いることで打開することができた．後は，社会実装に向けてバブルの検出精度の向上と，成果のビジュアル化と使いやすいインターフェイス構築である．精度向上については見通しも立って既に研究も軌道に乗っている．ビジュアル化とインターフェイスについては，β版を研究室の Web ページで公開し，実務家からフィードバックを受けながらアップデートを重ねている．社会実装に向けて関係機関と連携を取っている．現在，金融当局は，市場と対話する（金融政策の反応をリアルタイムに観測し，反応に合わせた次の一手をおこなう）技術に強い関心を寄せている．本成果は，その流れに沿うものである．

その他領域独自の評価項目

さきがけ期間中にフィンテックが脚光を集め，本成果もこの流れに乗って関連各所で成果を社会に広めることができた．また，本成果のスピンオフとして，紛争鉱物や奴隷労働問題等のサプライチェーンリスクに関する研究を新規に立ち上げた．これらは，国際関係学で扱われるテーマであり，当該分野ではビッグデータにもとづいた研究は数えるほどしかおこなわれていなかった．ビッグデータ応用分野の新しい研究領域を開拓したことも評価に値する．

（2）研究総括評価（本研究課題について、研究期間中に実施された、年2回の領域会議での評価フィードバックを踏まえつつ、以下の通り、事後評価を行った）。

現代社会の安定性は経済の安定性に大きく依存している．1990 年代の日本でのバブル崩壊や 2000 年代のリーマンショックを持ち出すまでもなく，いわゆるバブルが社会の安定性に及ぼす影響はきわめて大きい．このバブルは，市場価格が「本来の価格」と乖離することで発生する．今日の主要な市場の中心に，コンピュータで制御された大規模な情報

システムが存在することを考えると，バブルの問題に情報学的アプローチで挑戦することには意義がある．

本研究は，経済ビッグデータを解析することにより，バブルの検出，市場における価格変動の引き金の検出，価格変動が伝播する仕組みの解明などを目指すものである．バブルの検出に関しては「本来の価格」を推定する方式，引き金の検出に関しては市場に大きな影響を与える経済ニュースを検出する方式，伝播の仕組みの解明に関してはグローバル・サプライチェーンと企業間の業績連動の分析などを行い，従来「バブルは，はじけて初めてバブルだったとわかる」と言われていた状況から大きな一歩を踏み出すことができた．論文も多数公表されており，優れた学術成果をあげている．

今後は，社会実装を進めることを期待したい．また，本研究では，主に株式市場のバブルを研究対象としていることから，他の市場への展開も期待したい．

5. 主な研究成果リスト

(1)論文(原著論文)発表

1. Takayuki Mizuno, Takaaki Ohnishi and Tsutomu Watanabe. Power laws in market capitalization during the Dot-com and Shanghai bubble periods. Evolutionary and Institutional Economics Review. (in press)
2. Yoshifumi Tahira, Takayuki Mizuno. Trading strategy of a stock index based on the frequency of news releases for listed companies. Evolutionary and Institutional Economics Review. (in press)
3. Takayuki Mizuno, Takaaki Ohnishi and Tsutomu Watanabe. Structure of global buyer-supplier networks and its implications for conflict minerals regulations. EPJ Data Science. 2016, 5, 2 (15 pages)
4. Takayuki Mizuno, Takaaki Ohnishi, Tsutomu Watanabe. The Structure of Global Inter-firm Networks. Social Informatics Lecture Notes in Computer Science. 2015, 8852, pp. 334-338.
5. Takayuki Mizuno, Wataru Souma, Tsutomu Watanabe. The Structure and Evolution of Buyer-Supplier Networks. PLoS ONE. 2014, 9(7), e100712

(2)特許出願

なし

(3)その他の成果(主要な学会発表、受賞、著作物、プレスリリース等)

主要な学会発表

1. Takayuki Mizuno, Takaaki Ohnishi, Tsutomu Watanabe. Financial market reactions to exogenous shocks. ECONOPHYS-KOLKATA VIII. India. Mar. 14-17, 2014. (招待講演)
2. Takayuki Mizuno. Statistically detecting stock bubbles before they burst. ECONOPHYS-2015. New Delhi, India. Nov. 27 – Dec. 1, 2015. (招待講演)
3. Takayuki Mizuno. Pareto Indexes of Market Capitalization, PER, and PBR during Bubble Periods. WEHIA 2016 (The 21st annual Workshop on the Economic Science with Heterogeneous Interacting Agents). Castelló de la Plana, Spain. June 20-21, 2016.

4. Takayuki Mizuno, Takaaki Ohnishi, Hiroshi Iyetomi, Yuichi Ikeda and Tsutomu Watanabe. Structure of global buyer-supplier networks and its implications for conflict minerals regulations. CCS'15 (Conference on Complex Systems). Tempe, Arizona, USA. Sep. 28 – Oct. 2, 2015.
5. Takayuki Mizuno, Takaaki Ohnishi, Tsutomu Watanabe. Exogenous shocks in financial markets: Novelty and topicality detection in business news. Sigma Phi 2014. Rhodes, Greece. July 10, 2014.

著作物

1. Takayuki Mizuno, Wataru Souma, Tsutomu Watanabe. Buyer-Supplier Networks and Aggregate Volatility. The Economics of Interfirm Networks (Springer). pp. 15-38. 2015.

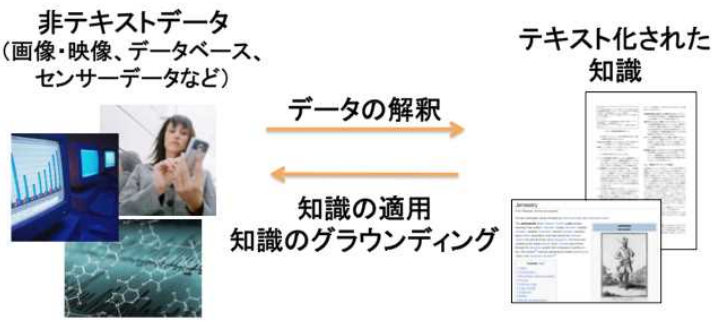
研　究　報　告　書

「非テキストデータと接続可能なテキスト解析・推論技術の開発」

研究タイプ：通常型
研究期間：　平成 25 年 10 月～平成 29 年 3 月
研　究　者：　宮尾　祐介

1.　研究のねらい

本研究は、自然言語テキストで書かれた情報と、非言語データ（画像・映像、データベース、時系列数値データなど）とを統合的に理解して推論を行うフレームワークを想定し、そのために必要となる自然言語の意味表現と、テキスト・非テキストデータの意味解析・推論技術を構築することを目指す。ビッグデータの効果的な利活用のためには、単に個々のデータを解析・検索・可視化するだけでなく、現在までに得られている知識を適用・応用して新たなデータや知識を創出する必要がある。後者のプロセスは、現在のところ研究者やデータサイエンティストなどの人間に全面的に頼らざるをえず、ビッグデータを利用した研究開発のボトルネックの一つと考えられる。様々なデータから得られる知識は自然言語を用いて表出・伝達・蓄積されるため、テキスト化された知識と非テキストデータに埋もれた情報を統合的に利用するためには、これらをつなぐ枠組みが必要である。



そこで、自然言語処理で用いられる自然言語の意味表現を出発点として、この意味表現を介してテキストデータや非テキストデータを接続することを可能にする枠組みについて研究を行う。これまでの研究で、テキスト間の意味的関係や推論関係（含意、矛盾、パラフレーズなど）を計算するための意味表現として、依存構造に基づく構成的意味表現と、それを利用した推論手法について研究を進めてきた。これを非テキストデータに適用することができれば、自然言語テキストと同様の推論を、テキスト・非テキストデータをあわせて行うことができる。

このような技術を実現するためには、自然言語の意味表現と解析技術、および非テキストデータに対して意味表現を求める技術の 2 つが必要である。前者は、テキスト間含意関係認識というタスクとして、自然言語処理において研究が進められている。ただし、非テキストデータと接続するために適した意味表現の設計や、認識精度の向上が課題として残っている。一方、後者の問題については、今のところ研究分野として確立されていない。例えば、「アジアの多くの都市では、モバイルネットワークの利用は夜間に集中している」というテキストがあった時、これを実世界データ、すなわち各都市におけるモバイル通信の利用統計と自動的に対応付けることは自明ではない。逆に、各都市におけるモバイル通信の利用統計があった時、そこからこのような言明を自動生成することも既存技術では不可能である。ここでの問題は、「アジ

ア」「都市」「モバイルネットワーク」「利用」「夜間」といった各単語、および単語間の意味的関係をどのように利用統計データと対応づけるか、という点にある。本研究では、以上の 2 つの問題について、自然言語テキスト、画像、データベース、時系列データを対象に、基盤技術の研究を行う。

2. 研究成果

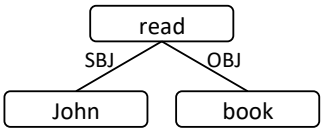
（1）概要

本研究では、以下の 4 項目について研究を進めた。研究項目 A. 自然言語テキスト間の含意関係認識の研究では、依存構造に基づく意味表現を利用して、頑健かつ高精度なテキスト間含意関係認識を行う手法を提案した（論文 5）。また、英語や日本語において、テキスト間含意関係認識、意味構造解析、およびこれらを利用する応用タスクの研究を推進するために、共有リソースの構築およびタスク提案を行った（論文 3, 4）。研究項目 B. 記号化された非テキストデータに対する解析技術の研究では、自然言語の質問文を大規模データベース（Freebase や DBPedia などのリンクトデータ）に対するクエリに自動変換することで、質問応答を行う技術について研究を行った（論文 2）。これは、自然言語の構文・意味構造をクエリという形式言語に変換することで、自然言語とデータベースをつなぐ技術である。また、日本語においてこのような研究を行った例がないため、既存の英語のデータセットを翻訳することで、日本語の評価用データセットを構築した。研究項目 C. 記号化されていないデータに対する解析技術の研究では、画像に対して依存構造に基づく意味表現を計算する手法を提案し、映像検索タスクにおいて有効性を示した。また、この意味解析技術を株式市場データに適用し、同様の意味解析が可能であることを示した。さらに、映像データに対する意味解析の研究を進めるために、自然言語による説明文を付与した映像データの開発を行った。研究項目 D. これらの技術を応用したプロトタイプシステムの開発では、上記の技術の応用例として、画像に対して自然言語により検索を行うシステムを実装し、本研究で開発した技術の有効性を示した。

（2）詳細

研究項目 A. 「自然言語テキスト間の含意関係認識の研究」

本研究項目では、依存構造に基づく構成的意味表現を利用して頑健かつ高精度なテキスト間含意関係認識を行う手法について研究を行った。依存構造に基づく構成的意味表現とは、右図（“John reads books”の意味表現）に示すように、単語とその間の構文的・意味的関係（依存関係）の木あるいはグラフで文の意味を表現するものである。述語論理などに基づく意味表現と比べて、直感的にわかりやすく、かつ解析がしやすいという長所がある。一方で、他の形式論理と比べて記述力が弱いという短所があるが、自然言語の文の意味や推論関係を表すには、ほとんどの場合この表現で十分であることを示した。さらに、この意味表現は構文構造との対応関係が明確であるため、構文構造を用いたパラフレーズ認識手法などを組み合わせることができる。英語におけるテキスト間含意関係認識の標準データセットを用いた実験では、人手で構築されたシソーラスである WordNet や、統計的手法で類義フレーズを認識することができる word embedding を用いることにより、高精度を達成することを示した（論文 5）。本手法は、日本語のテキスト間含意関係認識や法



律文書におけるテキスト間含意関係認識にも適用され、有効性を示した。また、本手法をさらに拡張し、一階述語論理では記述することが難しい一般化量化子に関して高精度かつ高効率な推論が可能であることを示した。

また、テキスト間含意関係認識は英語での研究がほとんどであったため、日本語における研究を推進するため、日本語の評価用データセットを構築し、評価型タスクを運営した(論文4)。本タスクには、国内外から多くのチームが参加し、日本語のテキスト間含意関係認識の研究の標準データとなっている。また、欧州を中心とした評価型ワークショップ CLEF や、自然言語の意味解析に関する評価型ワークショップ SemEval において、意味解析やその応用に関する評価用データを開発し、評価タスクを運営した(論文 3)。

研究項目 B.「記号化された非テキストデータに対する解析技術の研究」

本研究項目では、関係データベースやリンクトオープンデータなど記号化されているデータベースに対し、自然言語テキストとの対応関係を計算する手法について研究を行った。特に、大規模リンクトデータ(Freebase や DBPedia など)を知識源として用いて、自然言語の質問文で質問応答を行うタスクにフォーカスして研究を進めた。本タスクで代表的な評価型タスクである Question Answering over Linked Data (QALD) で提供されているデータから抜粋した例を下に示す。

質問文: Which river does the Brooklyn Bridge cross?

答え: http://dbpedia.org/resource/East River

QALD では、質問文は自然言語文として与えられ、答えとして Uniform Resource Identifier (URI) を返すことが求められる。このタスクは、与えられたリンクトデータ(この例の場合は DBPedia)に答えが存在することを前提としている。すると、データベースに対する検索クエリを作ることができれば、答えを得ることができる。上記の例については、以下のようなクエリを作ればよい。

```
SELECT DISTINCT ?uri WHERE {
   res:Brooklyn Bridge dbo:crosses ?uri
}
```

詳細は割愛するが、これは下図のような部分グラフを検索することに相当する。

したがって、本タスクは、自然言語文を上図のようなグラフに変換することが目的となる。この例から明らかなように、本タスクは、自然言語の単語やフレーズ(Brooklyn Bridge や does … cross など)を、データベースのノードやリンクに変換し、それらを適切に組み合わせる、という問題に帰着される。

本研究では、このタスクに対し、木構造を変換するフレームワークである Tree Transducer を応用する手法について研究を行った(論文 2)。Tree Transducer は、入力の木構造を有限の変換規則を用いて出力の木構造に変換する手法であり、これまで機械翻訳などに応用されている。そこで、自然言語の構造を表す木構造を入力とし、データベースクエリを表す木構造を出力する手法を提案した。機械翻訳と比べると、本タスクでは利用できる学習データが圧倒的に小さい。そこで、単語やフレーズをノードやリンクに変換する部分には教師なし学習を、変換ルールを学習する部分には教師あり学習を用い、双方の組み合わせを最適化すること

で変換規則の学習・適用を行う手法を提案した。これにより、少ない学習データからでも高精度な変換規則が得られることを示した。

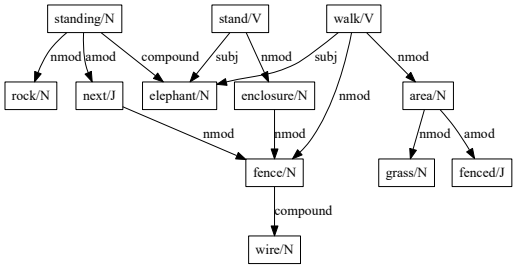
また、上記のタスクはこれまで主に英語で研究が行われてきたため、既存の標準データセットの英語データを翻訳することで、日本語データセットを構築した。このデータセットは、日本語の質問文をデータベースクエリに変換する研究などで利用されている。

また、上述のタスクとは逆に、自然言語テキストからデータベースを自動構築する研究について、共同研究を行った。情報抽出技術を利用することで、大規模テキストから高精度で関係知識を獲得することができ、質問応答において有効性を示した。

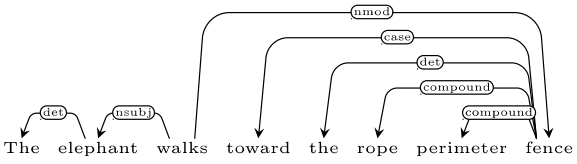
研究項目 C.「記号化されていないデータに対する解析技術の研究」

ビッグデータの多くは、画像・映像データ、センサーデータなどの時系列データなど、記号化されていないデータである。本研究項目では、主に画像データに着目し、上記研究項目で用いている依存構造に基づく意味表現を計算する手法について研究を進めた。

本研究では、右上図のような画像データを入力とし、その画像の意味内容を示す右下図のようなグラフ構造を出力する手法を提案した（論文 1）。このグラフ構造は、上述の研究で用いた依存構造を元としており、ノードが単語、エッジが単語間の関係を表している。このような形式で画像の意味表現が得られれば、上記のような意味解析・推論技術を組み合わせることができ、様々な応用が期待できる。



多様な画像に対して高精度でこのような出力を認識するモデルを学習するためには、大規模な学習データが必要である。しかし、このようなグラフ構造を、大量の画像に対して人手で構築するのは不可能である。そこで提案手法では、画像と説明文がペアとなった既存の大規模リソースを利用し、画像と意味表現のペアのデータを構築した。まず、説明文に対して自然言語の構文解析器を適用することで、構文・意味構造が得られる。たとえば、上記の画像に対する説明文 “The elephant walks towards the rope perimeter fence.” という文に対しては、下図のような構造が得られる。



一つの画像に対し、複数の説明文が与えられているため、すべての説明文に対して構文解析を行い、その結果を前処理・マージすることで、上記のような意味表現を得ることができる。

このような学習データが得られれば、入力画像に対してグラフ構造を予測するモデルを学習することができる。提案手法では、画像処理において複数の物体や概念を高精度で認識することができる深層学習手法を応用し、入力画像に対して各エッジを予測するモデルを提案し

た。これにより、人間が記述した説明文から得られる意味表現の一致率と同程度の精度で、意味表現に自動認識が実現できることが示された。また、この技術を応用することで、映像検索タスクにおいて有意な精度向上が見られることを示した。

また、本研究項目では、時系列データとして株式市場データや、映像データに対して同様の技術の研究開発を進めるため、研究用データの構築を行った。これらのデータを用いた予備実験では、画像に対する意味解析と同様に、意味表現の自動認識が可能であることを示している。

研究項目 D.「応用プロトタイプシステムの開発」

これまでに述べた基盤技術の応用可能性を示すため、画像データに対して自然言語で検索を行うプロトタイプシステムを実装した。本システムでは、大量の画像データに対してあらかじめ依存構造を計算し、検索エンジンでインデックスを作成しておく。この検索エンジンは、自然言語テキストの検索で広く用いられているものである。つまり、画像データに対し、あたかもテキストデータかのように検索を行うことができる。本システムは意味表現を用いた検索を実現しているため、同じ単語でも異なる意味関係を区別するような検索や、否定などの論理関係を用いた検索が可能であるなど、今後の様々な応用が期待される。

3. 今後の展開

本さきがけ研究では、自然言語と非テキストデータを意味表現を介してつなぐ技術について研究を行い、特に自然言語間の意味的推論、自然言語とデータベース、および自然言語と画像を対象とした研究において、大きな成果を挙げてきた。これまでの研究成果では、自然言語の意味表現と、データベース・画像それぞれをつなぐ技術の可能性について示すことができた。しかし、これらの技術を実社会の応用につなげるには、精度の大幅な向上が必要である。このためには、画像やデータベースをそのまま用いるだけでなく、自然言語テキストやシソーラスなど、外部知識を有効活用することが必要であると考えられる。また、本研究成果のさらなる可能性として、意味表現を介して複数の非テキストデータをつなぐ技術も考えられる。たとえば、データベースに蓄積された情報と、画像・映像データに埋もれた情報をつなぐことができれば、これまでにない応用技術が生まれる可能性がある。

具体的な応用アプリケーションとしては、画像やデータベースを利用した自然言語検索、質問応答システム、対話システムなどが考えられる。たとえば、観光案内の対話システムを考えると、あらかじめ想定している質問だけでなく、ユーザから与えられる画像・映像データと気象や交通データの情報を統合して、観光情報を提示したり新たな情報を蓄積するといった応用が考えられる。現時点では、このような知的システムの実現はまだ遠いと言えるが、本研究で開発した技術が重要な基盤となると期待される。

4. 評価

(1)自己評価

本研究課題は、自然言語処理だけでなく、データベース検索や画像処理など、複数分野にまたがった知見や技術が必要とされる、きわめてチャレンジングなものであった。本研究を開始する前は、自然言語処理の研究経験しかなく、他分野に関する知識がほとんどない状況であり、研究の推進は困難であったと言える。しかし、本研究がさきがけ研究者をはじめとして多

くの研究者と接点をもつ機会となり、現在ではさまざまな分野の研究者やエンジニアと共同研究を行い、研究論文を発表するまでに至っている。たとえば、画像に対して意味構造を認識する技術については、自然言語処理と画像処理の最先端技術を組み合わせる必要があり、当初の計画以上の成果が得られたと評価できる。

研究費は、主にデータ作成、研究補助員の雇用、および実験のためのクラウドサービスに用いた。データ作成は、新たな研究領域を開拓する本研究では必要不可欠であり、研究費が有効活用できたと考えている。研究補助員は、主にデータ作成とソフトウェア開発の補助にあたり、研究開発の効率的な推進に大きく寄与した。また、提案手法の評価実験のためには大きな計算リソースが必要であるが、クラウドサービスを活用することにより、管理コストや導入コストを大幅に削減することができた。

今後の展開で述べたように、本研究の成果はまだ基礎研究の段階であるが、自然言語とその他多様なデータを統合的に利用する場面は社会の様々な活動で現れるため、将来的な波及効果は大きい。画像・映像データの意味理解は、上で述べた観光案内をはじめとして様々な質問応答システムや対話システムに応用できるほか、視覚情報を利用する知的ロボットにも必須の技術である。また、株式データなどの時系列・数値データの意味理解は、企業における戦略判断、政策の妥当性や効果の検証など、自然言語を主とした知的活動において様々なデータをエビデンスとして必要とする場面に応用できる可能性がある。

（2）研究総括評価（本研究課題について、研究期間中に実施された、年2回の領域会議での評価フィードバックを踏まえつつ、以下の通り、事後評価を行った）。

有史以来，人類が生み出した知識は様々な形態で流通・蓄積されてきた．代表的な形態である書籍や論文は，本文，図，表などで構成されており，電子化されたものだけでも大量に存在する．これらをビッグデータとみなし，本文，図，表などの表現形式の違いを超えて，コンピュータが意味まで読み取って処理を行うことができれば，大きな社会的価値を生み出すであろう．

本研究は，自然言語のテキスト，画像，データベースの表現形式の違いを超えて，統合的に利用するための研究を行うものである．自然言語研究で一般的な意味構造を中心に，テキスト間の含意関係の認識，テキストによるデータベースの問い合わせ，画像からの意味構造の抽出に関する研究を行っている．そして，これらの要素技術を統合して自然言語で問い合わせ可能な画像データベースのプロトタイプも構築している．研究の過程では評価に必要なデータの整備も行っている．これらに関する論文を，トップカンファレンスなどで多数発表しており，優れた学術成果をあげている．

今後，適切な応用分野を設定して，実用化を目指すことを期待したい．また，自然言語処理，画像理解，機械学習などの広範な技術分野に関連する研究テーマであることから，これらの分野の研究者をリードして研究に取り組むことも期待したい．

5．　主な研究成果リスト

（1）論文（原著論文）発表

1. Sang Phan, Yusuke Miyao, Duy-Dinh Le, Shin'ichi Satoh. Video Event Detection by Exploiting Word Dependencies from Image Captions. Proceedings of COLING 2016. 2016. pp. 3318-3327.
2. Pascual Martínez-Gómez and Yusuke Miyao. Rule Extraction for Tree-to-Tree Transducers by Cost Minimization. Proceedings of EMNLP 2016. 2016. pp. 12-22.
3. Stephan Oepen, Marco Kuhlmann, Yusuke Miyao, Daniel Zeman, Silvie Cinková, Dan Flickinger, Jan Hajič, Angelina Ivanova, and Zdeňka Urešová. Towards Comparability of Linguistic Graph Banks for Semantic Parsing. Proceedings of LREC 2016. 2016. pp. 3991-3995.
4. Suguru Matsuyoshi, Yusuke Miyao, Tomohide Shibata, Chuan-Jie Lin, Cheng-Wei Shih, Yotaro Watanabe and Teruko Mitamura. Overview of the NTCIR-11 Recognizing Inference in TExt and Validation (RITE-VAL) Task. Proceedings of NTCIR-11. 2014. pp. 223-232.
5. Ran Tian, Yusuke Miyao, Takuya Matsuzaki. Logical Inference on Dependency-based Compositional Semantics. Proceedings ACL 2014. 2014. pp. 79-89.

（2）特許出願

なし

（3）その他の成果（主要な学会発表、受賞、著作物、プレスリリース等）

受賞

日本学術振興会 日本学術振興会賞 自然言語の構文解析・意味解析の研究とその応用 2015 年 2 月

情報処理学会 長尾真記念特別賞 自然言語の深い構文・意味解析の研究とその応用 2014 年 6 月