別紙１３

「知の創生と情報社会」研究領域 領域活動・評価報告書
－平成23年度終了研究課題－

研究総括 中島 秀之

1．研究領域の概要

本研究領域は多様もしくは大規模なデータから、有用な情報である「知識」を生産し、社会で活用するための基盤的技術となる研究を対象としている。

具体的には、大規模データを処理するための革新的な技術、統計数理科学を応用した分析・モデル化技術、あるいは実社会から得られる多様なデータを構造化・分析して知識を抽出する技術、センサによる情報取得やシミュレーション結果等の複数のリソースから新たな知識を創出する技術などの基盤技術に加えて、獲得した知識を実社会に適用するために必要とされる、シミュレーション、データの可視化、新しい情報社会の仕組みを支える応用技術などに関する研究を含んでいる。

2．研究課題・研究者名

別紙一覧表参照

3．選考方針

選考の基本的な考えは下記の通り。

1） 選考は、「知の創生と情報社会」領域のアドバイザー9 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。

・書類選考において 1 提案につき 3 名の選考委員が査読評価を行なう。

・選考委員の所属機関と応募者の所属機関が異なるよう配慮し、書類選考は利害関係者を査読対象とせず、面接選考において利害関係者は席を外して実施する。

・査読結果に基づき、3 年型と 5 年型に分けて、事前に総括と事務局とで順位付けを施す。

・面接選考では可能な限り多くの研究提案を直接聴取し、質疑応答する。特に 5 年型については、初年度の 5 年型に相応しい提案か否か(研究構想が本領域の趣旨に合っているだけでなく、研究期間の後半あるいは期間終了後において、実社会での応用がしっかりと考えられているかどうか)について質疑する。また、制度として、5 年型を 3 年型に回すことはできなかったため、再度挑戦の価値がある提案については、不採択理由にコメントをつけて次回以降に期待することにした。

3） 基本的に、3 年型は「知の創生」の基盤技術や理論を開発するもの、5 年型は実社会での適用や実運用のためのアプリケーションの開発など、「情報社会」での応用を目指すものを求めた。

4．選考の経緯

一応募課題につき領域アドバイザー3 名が書類審査し、書類選考会議において面接選考の対象者を選考した。続いて、面接選考および総合選考により、採用候補課題を選定した。

	選 考	書類選考	面接選考	採択数
合計	104 件	39 件	22 件	10 件※

※採択数 10 件の内、1 件は 5 年型

備考:

1）平成 20 年度採択課題のうち、以下 2 件は今年度事後評価を実施しない。

・大羽成征研究者

研究期間が5年で、今年度終了しないため

・島野美保子研究者

研究の一時中断により、今年度終了しないため。

5．研究実施期間

平成 20 年 10 月～平成 24 年 3 月

6. 領域の活動状況

1）領域会議：7回

うち1回はさきがけ「情報環境と人」研究領域と合同。

2）研究報告会：1回 （日本科学未来館)実施(2011/12/16)

「しみわたる情報技術・にじみ出す知 ― いま、収穫の時！」をテーマに実施。

3）研究総括の研究実施場所訪問、他

・研究総括は、公立はこだて未来大学内(函館)、技術参事および事務参事は平成 20 年 6 月始～平成 23 年 11 月末まで科学技術振興機構三番町ビル内、平成 23 年 11 月末以降は科学技術振興機構東京本部別館内にて、さきがけ初のバーチャル領域事務所態勢で業務を実施。

・研究場所訪問は、「サイトビジット」とも呼び、研究環境が十分であるかどうかの確認と、総括が上司の方にご挨拶することが目的。平成 21 年 1 月から 2 月にかけ、東京、名古屋、京都、大阪、仙台、札幌、函館に 10 名の研究者と面談。研究期間中に異動となった研究者 2 名については、平成 21 年 5 月に新しい研究場所の訪問を実施。

・当領域では、研究者間のコラボレーションを重視しており、その一環として、サイトビジットの他、研究者の希望や状況把握のため、領域事務所単独での研究者訪問(レクチャービジット)を実施、昨年 6 月以降、研究者が自発的に「オフ会」という交流会を開催し、積極的なコラボレーションを図っている。

4）他研究領域とのコラボレーション

・「情報環境と人」研究領域と合同で、アウトリーチ活動の一環として、情報処理学会第 72 回全国大会(創立 50 周年記念大会)にてさきがけセッションを開催(2010/3/11)、2011 年春にも情報通信学会全国大会でさきがけセッション開催予定であったが、東日本大震災のため中止。

7. 評価の手続き

研究者の作成した研究報告書および自己評価を基に、年 2 回の領域会議における経過報告および討議内容、領域アドバイザーの意見、さらに成果報告会(公開)での発表を参考にして研究総括が総合評価を行なった。

（評価の流れ）

平成 23 年 7 月　第 6 回領域会議（総括・アドバイザーによる進捗評価とアドバイス）
平成 23 年 12 月　研究報告会開催（一般参加者および総括・アドバイザーによる評価）
平成 24 年 2 月　第 7 回領域会議（総括・アドバイザーによる進捗評価とアドバイス）
平成 24 年 3 月　研究期間終了
平成 24 年 3 月　研究報告書提出
平成 24 年 3 月　研究総括による評価

8. 評価項目

（1） 当初の研究計画から見た進捗状況や達成度
（2） 当初計画で想定されていなかった新たな展開が生じているか否か
（3） 成果の科学的・技術的インパクト、国内外からの類似研究と比較したレベルや重要度

9. 研究結果

当領域では、実社会への応用を見据えた新しい基盤技術の研究開発提案や、すでに得られている大規模情報を対象とするだけでなく、情報を現実世界から取り込むための手法などの課題が採択されている。例えば、ネットワークに漂っているデータから、構造や機構を推定したり、情報を読み取ったり、実社会に読み取った情報を発信したりといった、社会との関わりのある課題である。現在、情報処理技術のキーワードとなっている「ビッグデータ」を扱うための基礎技術とその社会応用をテーマとした課題である。

「さきがけ」の先進性を理解した、できる限りチャレンジングであり、そのためにも現在の社会ニーズにとらわれず、それらを超える新しいニーズを創出する技術シーズを示した、今後の研究方向を創り出す核となれるような課題が集まったと考える。

〇猪口　明博　研究者

「大規模グラフ系列からの知識体系化と理解支援手法の開発」

大規模データからグラフ構造を、その変化に着目して抽出するグラフマイニングアルゴリズムのメインストリームである手堅い研究である。社会的な応用を意識しながら、データマイニングの先端的な技術を追求していくことを期待していた。

この 3 年半のさきがけ研究の研究期間で、構造が変化するグラフデータを対象として、特徴的な構造の変化をマイニングする手法を開発し、その有用性を示している。開発した手法は、後続の研究により開発された手法よりも汎用なデータに適用できる手法であり、様々な分野のデータに適用可能であることを評価する。
今後、データストリーム処理の概念を取り入れ、リアルタイム分析が可能な処理基盤を開発し、さらに本分野の研究を推進して行くことを期待する。

〇大野　和則　研究者
「ロボットの視覚・触覚を用いた環境情報獲得手法の開発」

ロボットが能動的に実環境に働きかけてさまざまな情報を持つ 3 次元地図を獲得する手法の研究である。未知物体が存在する空間で、触覚と視覚情報とを用いて物体情報を獲得する知能の開発を行い、レスキューロボットや家庭用ロボットの実現を通じた貢献を期待していた。

途中、東日本大震災によるブランクや原発建屋調査への協力等、（後者は技術の社会応用という意味で好ましいことであるが本領域にとっては）研究進捗阻害要因もあったが、当初予定の課題をほぼ達成しており、今後、ロボットを通した情報の獲得と知識の構築技術の開発に期待している。また、獲得する情報をネット上の情報と組み合わせる等、ロボットを通した情報の獲得と知識の構築を掘り下げて欲しい。

実世界の全ての情報を電子化するという目標は道半ばであり、今後もこの目標の達成を目指していってほしい。

〇岸本　章　宏研究者
「大規模並列化によるハイパフォーマンス人工知能技術」

囲碁等のゲームを素材とした大規模探索というテーマである。探索は大規模データからの知識獲得のための基本方式であるため、今後、学習や知識発見との関わりを付けて、知識創出に関係する研究の方向が出てくることを期待していた。

当初の計画よりも早いペースで、ゲームとプランニングの分野で、効率の良い大規模並列化アルゴリズムの開発を進めることができ、予想以上の大きな研究成果を出すことができている。また、当初の計画よりも研究が順調に進んだため、計画にはなかった逐次探索アルゴリズムの性能改良と BDD 構築の並列化も行っており、高く評価したい。

一方、アルゴリズム開発・改良では成果を上げているが、情報社会への応用については十分な成果を得られていない。今回の研究で駆使した考え方である「データ駆動型スケジューリング」に基づく並列プログラミング言語や並列ライブラリを開発できれば社会で実際に利用される機会が広がると考えられ、今後のテーマにしていってほしい。

〇寺沢　憲吾　研究者
「擬似コード変換と統計解析による文書画像からの知識抽出」

文字認識不能な文書画像を画像のまま部分検索などできるようにする重要な基盤技術の提案である。実績は高く、技術としての有用性が高い課題であった。

研究成果を社会に還元する、研究のアウトリーチ活動は当初計画を超えた進展を得ている。函館市中央図書館との連携、人文学の研究者との協力関係を構築できている。さらにウェブサービスの公開により、一般ユーザの利用にもつながっていることは高く評価できる。情報検索の分野に大きく貢献したと考える。

この研究のねらいの「画像データとテキストデータの間に橋を架けること」に関しては達成している。これまでに存在しなかった研究分野を開拓した功績は大きい。人文科学における新しい道具ができたので、それを活用した研究成果も期待できる。今後、更に効率の良いアルゴリズムの開発を中心とする理論的研究を含め、この活動が個人を超えて発展することを期待する。

〇ナイジェル・コリアー　研究者
「健康被害を監視するための多言語ウェブサーベイランスシステム」

インターネット上に流れている健康被害に関するニューステキストを分析して健康被害情報に関するアラームを発信するシステムの提案であり、社会的意義が高く、大規模な応用分野をカバーしている課題である。技術的には、イベント系列の解析により情報の重要度を推測し、アラートを出せるようになることを期待していた。

この課題で、公衆衛生に関する脅威を、オンラインのメディアソースを用いて検出するシステムの性能向上のために、アルゴリズムとリソースを開発しており、本来の目的の重要な部分は達成できたものと評価する。また、新たなアルゴリズムの開発だけでなく、地理−時間的情報の自動注釈のための新たなスキーマの作成、報告事象間の関連性を理解するための知識モデルの構築といった本来の目的の一部はすでに達成されていることを評価する。自然言語処理および変化点検出アルゴリズムを併用するという、警報に対する新たなアプローチの開拓が可能となり、開発した多言語の公衆衛生オントロジーは、これまでに、世界中で 350 を超えるグループによりダウンロードされ、利用されている。課題提案時より暫定的に動いていたシステムを中心にした研究開発であったため、完成度が何より重要である。この期待に応える利用実績を示し、この分野に大きく貢献したと考える。

〇福田　健介　研究者
「時空間解析に基づくインターネット異常トラフィックの検出とそのデータベース化」
インターネットトラフィックの時系列を時空間パターンにして解析し、大量のデータに埋もれた少量の検出対象を見つけるという課題である。具体性があり、研究を評価するためのデータの準備も整っており、インターネットに限らず、汎用性のある画像認識によるネットワークダイナミックの異常を検知する技術、また、トラフィック解析のみにとどまらず、幅広い場面での適用が可能な技術となることを期待していた。
当初の提案では、画像処理アプローチに基づく異常検出器をメインターゲットとして研究を進める予定であり、実際、精度の高い、自動パラメータ設定可能な検出器を提案できたことから、当初の目的を達成できていると評価する。さらに、現在のインターネットトラフィック異常検出におけるさまざまな問題点(共通したデータセットの欠如、正解データの欠如、異常検出器間の性能比較方法の欠如)を解決するよう、研究トピックを追加した。その結果、提案時にはなかった、複数異常検出器の比較ベンチマークアーキテクチャおよび共通データベースの精度向上に関しての研究が進み、トップ国際会議への採択に至っている。また、複数検出器出力に基づいた、インターネットトラフィック異常データベースを公開したことで、他研究者からベンチマークに使用される標準データベースとなりつつある。この分野に大きく貢献したと考える。
一方、異常検出器の実ネットワーク(リアルタイム)への適用が遅れており、今後の研究に期待する。

〇星野　崇宏　研究者
「マルチソースデータ高度利用のための統計的データ融合」
マルチソースデータからシングルソースデータをシミュレートし補完する手法であり、断片的な大量データから各種の統計的推定を安定に行うために不可欠な技術である。必ず開発していかなければならない技術であり、コア的研究といえる。研究成果を具体的な調査活動につなげることを意識して研究を進めることを期待していた。
数理的な方法論の開発と言う観点では、「既存のパラメトリックな手法をセミパラメトリックモデルで代替する」という当初目標にとどまらず、「既存の手法の仮定を大幅に緩和することが可能である」ことを発見できたため、非常に大きな進展があったと評価する。また、疑似パネルデータ解析やマルチレベル分析において、今回開発した方法を応用することで既存の方法論よりも頑健かつ効率の良い因果効果推定の方法を開発することができたという点で社会科学への応用研究と言う点で当初目標を達成していると評価する。この分野に大きく貢献していると考える。
実務に応用できるパッケージ化という観点では、事前の共変量選択に対する示唆を得ることが出来なかったという点が残念であるが、今後、様々な応用例の積み重ねを行うことで一般則を見出していくことを期待する。

〇松尾　豊　研究者
「ネットワーク理論と機械学習を用いたウェブ情報の構造化・知識化」
ネットワーク理論と機械学習を用いたウェブ情報の構造化・知識化についての研究である。ウェブマイニングにおける「エンティティ」発見をネットワーク構造上の機械学習の観点から捉えていて独創的であり、ウェブマイニングの核となるアルゴリズムを発見・構築したいという意気込みに期待していた。
提案する手法に関して議論を深め、手法の大枠を示すことができ、また、ウェブマイニングに関して社会的にもインパクトのある応用例を実現し、公表している。これらは彼ならではの優れた成果と考える。また、当初の目標を達成しており、この点も高く評価したい。一方で、構想自体は大きなものであるため新しい今後の研究課題も多く出ている。アルゴリズムとしての精錬や評価の点では、今後、更なる成果につながっていくことを期待する。

10. 評価者

研究総括 中島 秀之 公立はこだて未来大学・学長

領域アドバイザー(五十音順。所属、役職は平成 24 年 3 月末現在)

麻生 英樹 産業技術総合研究所知能システム研究部門・主任研究員
有村 博紀 北海道大学大学院情報科学研究科・教授
高野 明彦 国立情報学研究所連想情報学研究開発センター・センター長／教授
林 晋 京都大学大学院文学研究科・教授
林 幸雄 北陸先端科学技術大学院大学知識科学研究科・准教授
樋口 知之 統計数理研究所・所長
堀 浩一 東京大学大学院工学系研究科・教授
安田 雪 関西大学社会学部・教授
鷲尾 隆 大阪大学産業科学研究所・教授

(参考)

(1)外部発表件数

	国 内	国 際	計
論 文	10	37	47
口 頭	53	51	104
その他	2		2
合 計	65	88	153

※平成 24 年 3 月現在

(2)特許出願件数

国 内	国 際	計
1	0	1

(3)受賞等

岸本章宏研究者
- International Conference on Automated Planning and Scheduling "best paper award"(2009.9)
- 人工知能学会 全国大会優秀賞(H21.9)
- Advances in Computer Games "best paper award"(2011.11)

寺沢憲吾研究者
- 電子情報通信学会 論文賞(H21.5)

星野崇宏研究者
- 日本行動計量学会 出版賞(H23.9)

(4)招待講演

大野和則研究者
- 国内 2 件

岸本章宏研究者
- 国内 1 件

ナイジェル・コリアー研究者
- 国際 10 件 国内 2 件

星野崇宏研究者
- 国際 1 件

松尾豊研究者
- 国際 1 件

別紙

「知の創生と情報社会」領域　終了評価実施　研究課題名および研究者氏名

研究者氏名 （参加形態）	研　究　課　題　名 （研究実施場所）	現　職（平成 24 年 3 月末現在） （応募時所属）	研究費 （百万円）
猪口 明博 （　兼任　）	大規模グラフ系列からの知識体系化と理解支援手法の開発 （大阪大学）	大阪大学産業科学研究所・助教 （同上）	35
大野 和則 （専任）	ロボットの視覚・触覚を用いた環境情報獲得手法の開発 （東北大学）	JST さきがけ研究者 （東北大学大学院情報科学研究所・助教）	39
岸本 章宏 （　兼任　）	大規模並列化によるハイパフォーマンス人工知能技術 （公立はこだて未来大学、東京工業大学）	東京工業大学大学院情報理工学研究科・助教 （公立はこだて未来大学システム情報科学部情報アーキテクチャ学科・助教）	32
寺沢 憲吾 （　兼任　）	擬似コード変換と統計解析による文書画像からの知識抽出 （北海道大学、公立はこだて未来大学）	公立はこだて未来大学システム情報科学部・助教 （北海道大学大学院情報科学研究科知識メディアラボラトリー(VBL)・PD 研究員）	36
ナイジェル・コリアー （　兼任　）	健康被害を監視するための多言語ウェブサーベイランスシステム （国立情報学研究所）	国立情報学研究所情報学プリンシプル研究系・准教授 （同上）	33
福田 健介 （　兼任　）	時空間解析に基づくインターネット異常トラフィックの検出とそのデータベース化 （国立情報学研究所）	国立情報学研究所 情報学アーキテクチャ科学研究系・准教授 （同上）	32
星野 崇宏 （　兼任　）	マルチソースデータ高度利用のための統計的データ融合 （名古屋大学、シカゴ大学）	名古屋大学大学院経済学研究科・准教授 （同上）	41
松尾 豊 （　兼任　）	ネットワーク理論と機械学習を用いたウェブ情報の構造化・知識化 （東京大学）	東京大学大学院工学系研究科・准教授 （同上）	40

※本領域では、研究場所について複数記載した。

研 究 報 告 書

「大規模グラフ系列からの知識体系化と理解支援手法の開発」

研究期間: 平成 20 年 10 月～平成 24 年 3 月
研 究 者: 猪口 明博

1, 研究のねらい

計算機技術やネットワーク技術の発展により人間の処理能力を超えるほどの, 多様で, 大規模なデータの生成, 蓄積が可能となった. 多様で, 大規模なデータを分析するための手法を確立することで, 従来手法では獲得できなかった有用な情報を獲得でき, 社会活動を向上させることが可能である. 本研究では, 構造が変化するグラフを分析対象とする. 例えば, 人間関係ネットワークの人間をグラフの頂点, 人間関係をグラフの辺とすると, ある時点での人間関係はグラフで表現することができる. 人間関係は常に一定ではなく, 時間とともに変化する. すなわち変化するグラフで表現することができる. 人間関係に限らず, 遺伝子が頂点, 相互関係が辺である遺伝子ネットワークは進化の過程で遺伝子を新規獲得, 欠落, 突然変異する変化するグラフにより表現可能である. このようなデータを対象として, 共通する変化を見出すことができれば, 将来の構造変化の予測に役に立つと考えられる. 本研究の目的は, 構造が変化するグラフを分析対象として, 変化するグラフに特徴的に現れる共通の変化をマイニングするための手法を確立することである.

2, 研究成果

(1) 変化するグラフから共通の変化をマイニングする手法 GTRACE

本研究の目的は, 構造が変化するグラフを分析対象として, 変化するグラフに特徴的に表れる共通の変化をマイニングする手法を確立することである. 対象とするグラフ系列は以下の通りである.

- グラフ系列において, 頂点数や辺数は増減する.
- グラフ系列において, 頂点ラベルや辺ラベルは変化する.
- グラフ系列において, 連続する 2 つのグラフの構造は大きくは変化しない.

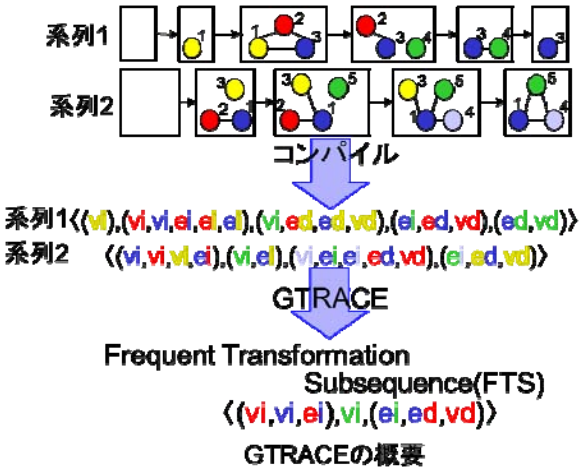
グラフ系列中において, 連続する 2 つのグラフの構造変化は小さいという仮定に基づいて, グラフの変化を表現するために, グラフの変換規則を提案した. この場合グラフ系列の各グラフ全体の情報を保持するのは冗長である. 従って, 連続する 2 つのグラフの差分のみを保持すれば十分であり, 解析手法の領域量を大きく削減することができ, それにより計算量を大きく下げることが可能となる. この変換規則の系列により任意のグラフ系列を表現することが可能である. また, 任意のグラフ系列を頂点数と系列長に比例する計算量で変換規則の系列にコンパイルできる.

さらに変化するグラフに現れる共通の変化をマイニングするために, 変換規則の系列から頻出する部分系列をマイニングする手法 GTRACE (Graph TRAnsformation sequenCE mining) を開発した. GTRACE は, 入力として与えられた n 本の変換規則の系列から k 本以上の変換系列に頻繁に表れる頻出変換部分系列(FTS: Frequent Transformation Subsequence)を効率良くマイニングする手法であり, 頻出変換部分系列(FTS)は変化するグラフに特徴的に表れる

共通の変化である.

提案手法の概要は図に示す通りである. 入力として, グラフ系列の集合と閾値 k が与えられる. 各グラフ系列を変換規則の系列にコンパイルする. コンパイルにより得られた変換規則の系列から GTRACE により頻出変換部分系列(FTS)をマイニングする. さらに, GTRACE に逆探索の概念を取り込むことで, GTRACE を効率化した手法 GTRACE-RS を提案した.

GTRACE を提案して以降, 変化するグラフから特徴的なパターンをマイニングする手法が提案されているが, 辺のみが変化し, 頂点数は一定である手法や頂点の増加のみを許容する手法などグラフ系列に制約を課したものが多い. このような制約を課すことで, グラフ系列のより単純な構造であるグラフに変換でき, 2000 年より知られる既存のグラフマイニング手法で解くことが可能となる. GTRACE は, 他の手法が適用できるグラフ系列に比べ汎用なクラスのグラフ系列の適用可能な手法である.



GTRACEの概要

(2) 変化するグラフから共通の構造をマイニングする手法 FRISSMiner

成果 1 において, 構造が変化するグラフを分析対象として, 変化するグラフに特徴的に表れる共通の変化をマイニングする手法 GTRACE を確立したことを述べた. GTRACE では, グラフ系列中の連続する 2 つのグラフで, その構造は大きく変化しないことを仮定している. しかし, グラフ系列を観測する際(グラフデータを収集する際)に, 時間分解能が低い場合, 観測されたグラフ系列の連続する 2 つのグラフの間で, グラフの大部分が変化する可能性がある. 従って, このようなデータに対する解析に, GTRACE は適さない. そこで, このような課題を克服するために, 我々は, FRISSMiner(Frequent, Relevant, and Induced Subgraph Subsequence Miner)を開発した. FRSSMiner は, グラフ系列の集合と閾値 k を入力として受け取り, グラフ系列中に頻繁に, かつ共通して表れる構造をパターンとしてマイニングする手法である. FRISSMiner では人間に理解容易で, 可読な共通パターンをマイニングするために, 得られるパターンにおける頂点の関連性(relevancy)を連結グラフにより定義し, パターンと入力データとの頂点の関係の再現性を頂点誘導部分グラフ(vertex-induced subgraph)により定義している. FRISSMiner が出力する頻出パターンを関連があるグラフ系列であり, かつ誘導部分グラフ系列として含まれるものに限定したため, FRISSMiner がマイニングするパターンは理解容易であり, それらを効率良く探索することができる.

以上, 成果 1 と 2 により, 構造が変化するグラフを分析対象として, 変化するグラフに特徴的に表れる共通の変化をマイニングする手法を確立した.

(3) グラフ系列マイニング手法の係り受け解析への応用

近年, 係り受け解析は, 情報抽出, 機械翻訳, テキスト含意認識, 質問応答, オントロジー導出などに様々に応用される自然言語処理における基礎技術として注目を浴びている. 係り

受け解析手法は，状態遷移系に基づく手法，グラフ理論に基づく方法，文法に基づく方法に大別することができる．状態遷移に基づく方法は，その内部状態をグラフ，単語（あるいは文節）を頂点，係り受け関係を辺とするグラフで表すことができる．さらに初期状態から最終状態への過程において，各状態を表すグラフは変化するために，状態遷移の系列はグラフの系列により表現することが可能である．状態遷移に基づく係り受け解析器が，出力結果を誤る状態遷移系列の集合から共通する変化（状態遷移の共通性）を見出すことができれば，係り受け解析器が誤る原因の究明，新たな係り受け解析アルゴリズムのデザインなどに役に立つと考えられる．前述のグラフ系列マイニング手法により得られる共通パターンは，人間に可読であり，共通性を見出すという目的に合致する．我々は，状態遷移に基づく係り受け解析器の 1 つである，arc standard shift reduce 型の係り受け解析器について，日本語の係り受け解析の検証実験を行った．この検証実験では，新聞記事からなる京都大学テキストコーパスを分析対象とした．arc standard shift reduce 型の係り受け解析器を用いて係り受け解析を行い，係り受け解析器内の状態遷移の系列をデータ化した．さらに，グラフ系列マイニング手法を適用し，arc standard shift reduce 型の係り受け解析器が解析を誤る典型的な状態の遷移をパターンとしてマイニングした．検証実験の結果，日本語の文法上，妥当なパターンが得られた．また，これらのパターンに基づいて，グラフを書き換えることで，係り受け正答率を改善することができた．以上により，前述のグラフ系列マイニング手法の有用性を確認できた．

さらに，適用手法により得られるパターンが文法上，妥当であった場合，そのパターンに基づいて係り受け関係を修正することで誤った係り受け構造が得られる文は，人手で作成されたコーパスデータが誤りの可能性がある．従ってコーパスの改善にも役に立つと考えられる．

3，今後の展開

成果3に示した手法の原理は，日本語に限らず，また arc standard shift reduce 型の係り受け解析器に限らず適用できる手法であるので，他言語，あるいは arc standard shift reduce 型ではない状態遷移に基づく係り受け解析器へ適用可能である．さらに，これの手法は，係り受け解析に限らず，状態がグラフで表現される状態遷移系全般に適用可能であるため，さらに広い範囲の応用分野に適用可能であると考えられる．

また，さきがけ研究開始時点では，目的の 1 つとして掲げてはいなかったが，データストリームに対するリアルタイム分析も重要な課題の 1 つであると考えられる．Facebook におけるコメントや Titter におけるリツイートなど，人間関係ネットワークにおいて生成されるデータは，データストリームとして大量にインターネット上を流れている．これらの過去のデータによって構成される人間関係ネットワークの構造と現在のデータによって構成される人間関係ネットワークの構造が必ずしも一致するとは限らない．リアルタイムに起こっている構造の変化を検知，把握する分析を行うためには，本研究で開発したバッチ処理による分析手法ではなく，ストリーム処理による分析手法，それらのハイブリッド型の分析手法が必要である．今後，本研究での技術をさらに発展させ，多様で，大規模なデータの分析を，さらに“リアルタイム”に実行する計算基盤技術の確立が重要であると考える．

4，自己評価

この 3 年半のさきがけ研究の研究期間で，構造が変化するグラフデータを対象として，特徴

的な構造の変化をマイニングする手法を開発し，その有用性を示した．開発した手法は，後続の研究により開発された手法よりも汎用なデータに適用できる手法であり，様々な分野のデータに適用可能である．今後，データストリーム処理の概念を取り入れ，リアルタイム分析が可能な処理基盤を開発し，さらに本分野の研究を推進したい．

5　研究総括の見解

大規模データからグラフ構造を、その変化に着目して抽出するグラフマイニングアルゴリズムのメインストリームである手堅い研究である。社会的な応用を意識しながら、データマイニングの先端的な技術を追求していくことを期待していた。

この 3 年半のさきがけ研究の研究期間で、構造が変化するグラフデータを対象として、特徴的な構造の変化をマイニングする手法を開発し、その有用性を示している。開発した手法は、後続の研究により開発された手法よりも汎用なデータに適用できる手法であり、様々な分野のデータに適用可能であることを評価する。

今後、データストリーム処理の概念を取り入れ、リアルタイム分析が可能な処理基盤を開発し、さらに本分野の研究を推進して行くことを期待する。

6，主な研究成果リスト

（1）論文（原著論文）発表

1. Akihiro Inokuchi and Takashi Washio. A Fast Method to Mine Frequent Subsequences from Graph Sequence Data. Proc. of the 8th IEEE International Conference on Data Mining (ICDM 2008), pp 303-312, 2008.
2. Akihiro Inokuchi and Takashi Washio: GTRACE2: Improving Performance Using Labeled Union Graphs. Proc. of 14th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2010), pp. 178-188, 2010.
3. Akihiro Inokuchi and Takashi Washio. Mining Frequent Graph Sequence Patterns Induced by Vertices. Proc. of SIAM International Conference on Data Mining (SDM 2010), pp. 466-477, 2010.
4. Akihiro Inokuchi and Takashi Washio: GTRACE: Mining Frequent Subsequences from Graph Sequences. IEICE Transactions 93-D(10): pp. 2792-2804, 2010
5. 猪口　明博，頻出パターンマイニングのグラフ系列への適用，人工知能学会学会誌，掲載予定

（2）特許出願

該当なし

（3）その他の成果（主要な学会発表、受賞、著作物等）

- 猪口　明博，鷲尾隆，頂点により誘導される頻出グラフ系列パターンのマイニング，第 12 回 人工知能学会 データマイニングと統計数理研究会，2010
- Nguyen Duy Vinh, Akihiro Inokuchi, and Takashi Washio, Graph Classification Based on

Optimizing Graph Spectra. Proc. of International Conference on Discovery Science (DS2010), pp. 205-220, 2010.

- 生田 泰章, 猪口 明博, 鷲尾 隆, 逆探索法によるグラフ系列マイニングの高速化, 第 3 回データ工学と情報マネジメントに関するフォーラム, B10-2, 2011
- 猪口 明博, グラフ系列マイニング, 第 2 回 Latent Dynamics Workshop (招待講演)
- Akihiro Inokuchi, Hiroaki Ikuta, and Takashi Washio: GTRACE-RS: Efficient Graph Sequence Mining using Reverse Search CoRR arXiv: 1110.3879, 2011
- 猪口明博, 山岡 歩, 鷲尾 隆, 松本裕治, 浅原正幸, 岩立将和, 賀沢秀人, 係り受け解析における状態書き換え規則のマイニング, 第 1 回 データ指向構成マイニングとシミュレーション研究会 予稿集, pp.2.33-2.41, 2011

研 究 報 告 書

「ロボットの視覚・触覚を用いた環境情報獲得手法の開発」

研究期間： 平成20年10月～平成24年3月
研 究 者： 大野 和則

1，研究のねらい

3次元空間内の未知の物体をロボットが触って動かすことで分割、モデリングを行う方法の開発を行う．従来研究ではモデルベースの物体認識を用いることで，空間中から知識にある物体を見つけることができる．一方、実際の世界には電子化されていない情報がまだまだ多く，知識にない物体は障害物として認識され，ロボットが未知物体を認識・操作することができない．この問題を解決するため，空間の詳細な3次元計測を行い，そこから，未知の物体を触って動かすことで切り出し，切り出した物体の詳細な形状をモデリングする方法を開発する．

2，研究成果

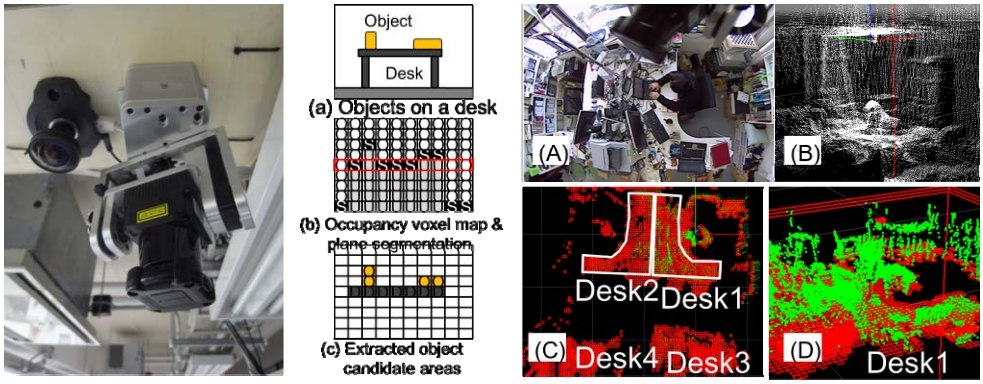
本研究の主な成果は下記の4つである．

（1）3次元点群計測と認識処理

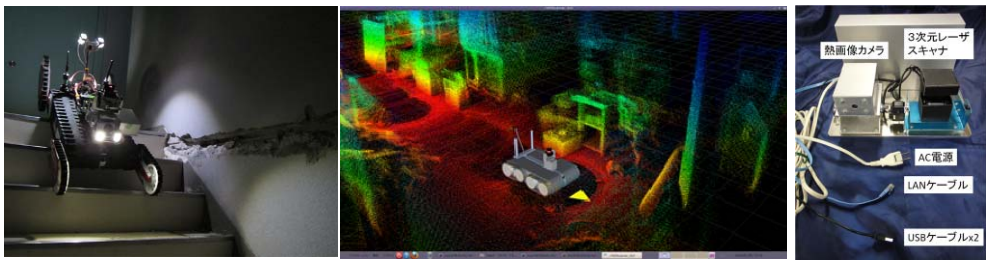
広範囲の密な 3 次元点群の計測が可能な独自のレティキュレートスキャン方式の3次元レーザスキャナを開発した．天井に取り付けられる軽量なセンサ(TKScanner)と，移動ロボットに搭載可能な防塵・防水のスキャナ(HDScanner)の 2 種類を開発した．図1に3次元レーザスキャナと3次元計測の成果を示す．

天井に TKScanner を取り付けて計測した3次元点群から，机のような大きな平面を検出し，平面上にある物体を検出する方法を開発した．占有度グリッドマップと平面検出を併用し物体の大きさを予測可能な方法を開発した【参考文献＊】．

HDScnaner は災害現場の3次元計測を行うため2011年4月に Quince に搭載し移動計測が可能な状態に整備した．また，TKScanner の技術を応用して、日本原子力機構（JAEA）と共同開発したロボット操縦車両 TEAM NIPPON に搭載する小型3次元レーザスキャナを開発した【参考文献】．



（a）3次元レーザスキャナを用いた机上の物体の発見



(b)　3次元レーザスキャナを搭載した移動ロボットによる被災建物の3次元計測

図1．3次元レーザスキャナと移動ロボットによる空間の3次元計測

（2）押し動作をもちいた未知物体のモデリング

未知物体を触って動かすことで分割・モデリングを行う方法を開発している．距離センサとカメラ映像で発見した未知物体を押し動作により分割し，さらに押し動作で回転させることで全周囲のモデリングを行った．図2に押し動作を用いて構築した円柱物体の 3 次元モデルを示す．円柱や立方体などの物体を押し動作を通してモデリング可能なことを確認した【参考文献】

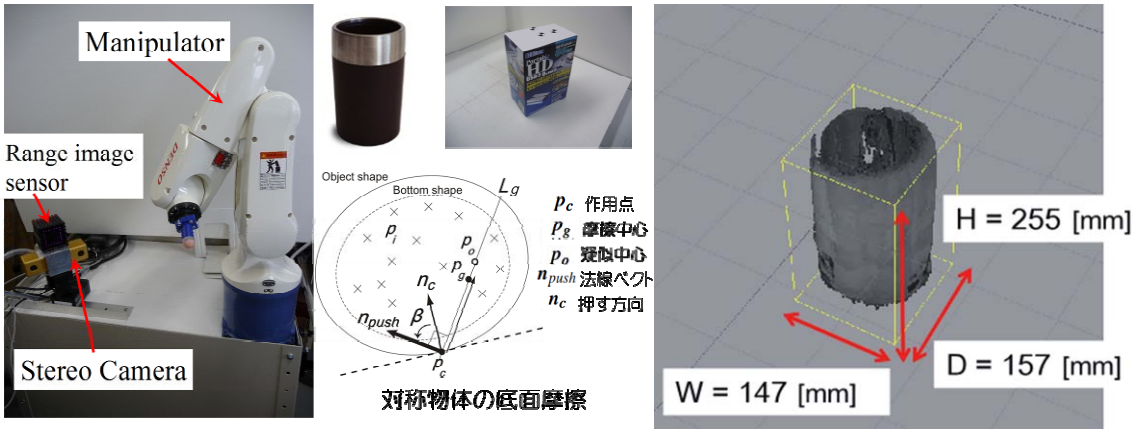


図2．押し動作を用いた未知物体のモデリング

（3）透明物体の認識

透明な物体を発見する方法を開発した。カメラとレーザを併用した方法を新たに提案した。レーザ光が透明物体を通過する際に反射光強度が減少するという現象を利用し，ペットボトルやカップなどを発見する方法を開発した。カラーカメラから得られるハイライトスポットの情報を利用して，透明物体が存在する領域を特定し，反射強度の映像からグラブカットを用いて透明な物体を抽出した。

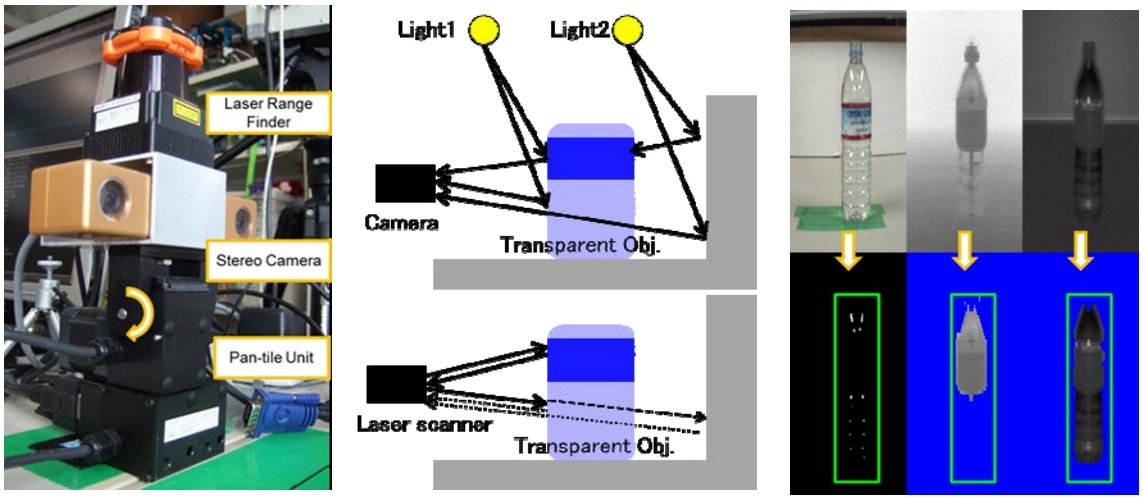


図3．カメラとレーザ距離計を用いた透明物体の検出

（4）移動台車の整備と高精度位置推定

レーザ光を反射するマーカとオドメトリを融合した高精度な位置・姿勢を推定する方法を開発。図4に示す移動ロボットを用いて研究室内で実証試験を行った．

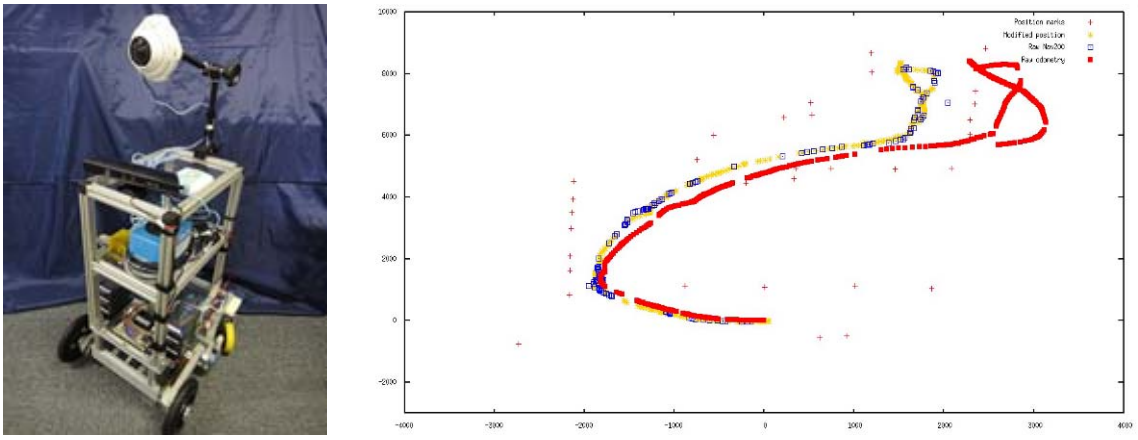


図4．移動ロボットと高精度位置推定手法の開発
（赤：オドメトリで推定した位置,黄：融合後の位置）

3，今後の展開

開発した各技術を統合し，レスキューロボット，福祉ロボットなどが実世界で未知物体を発見・把持するアプリケーションに応用する．

4，自己評価

さきがけ研究では，透明な物を含む未知物体の計測，動きを用いた物体の分割とモデリング方法の研究開発に取り組んだ．開発期間を通して個々の要素技術を作ることができ，実験を通して課題も分かってきた．実世界の全ての情報を電子化するという目標は道半ばであり，今後もさきがけで培った技術を基盤に目標の達成を目指す．

5，研究総括の見解

ロボットが能動的に実環境に働きかけてさまざまな情報を持つ 3 次元地図を獲得する手法の研究である。未知物体が存在する空間で、触覚と視覚情報とを用いて物体情報を獲得する知能の開発を行い、レスキューロボットや家庭用ロボットの実現を通じた貢献を期待していた。

途中、東日本大震災によるブランクや原発建屋調査への協力等、（後者は技術の社会応用という意味で好ましいことであるが本領域にとっては）研究進捗阻害要因もあったが、当初予定の課題をほぼ達成しており、今後、ロボットを通した情報の獲得と知識の構築技術の開発に期待している。また、獲得する情報をネット上の情報と組み合わせる等、ロボットを通した情報の獲得と知識の構築を掘り下げて欲しい。

実世界の全ての情報を電子化するという目標は道半ばであり、今後もこの目標の達成を目指していってほしい。

6，主な研究成果リスト

（1）論文（原著論文）発表

1. Kazunori Ohno, Peter Andersson, Zhong Lei, Eijiro, Takeuchi, Satoshi Tadokoro, "Multi-Object Recognition on the Basis of Vision and Manipulation,", Proc. of 2010 IEEE/SICE International Symposium on System Integration, A1-4, 2010.

2. Kazunori Ohno, Satoshi Tadokoro, Keiji Nagatani, Eiji Koyanagi, Tomoaki Yoshida, "Trials of 3-D Map Construction Using the Tele-operated Tracked Vehicle Kenaf at Disaster City," Proc. IEEE International Conference on Robotics and Automation, pp.2864-2870, 2010.

3. Kazunori Ohno, Shinji Kawatsuma, Eijiro Takeuchi, Kazuyuki Higashi, Satoshi Tadokoro and Takashi Okada,"Robotic Control Vehicle for Measuring Radiation in Fukushima Daiichi Nuclear Power Plant," Proc. of IEEE International Workshop on Safety, Security, and Rescue Robotics (SSRR2011), pp. 38-43, 2011.

4. Zhong Lei, Kazunori OHNO, Masanobu TSUBOTA, Eijiro TAKEUCHI, Satoshi TADOKORO, "Transparent Object Detection Using Color Image and Laser Reflectance Image for Mobile Manipulator," 2011 IEEE International Conference on Robotics and Biomimetics (ROBIO2011)pp. 1-7, 2011.

5. Kazunori Ohno, Kurose Kensuke, Eijiro Takeuchi, Lei Zhong, Masanobu Tsubota and Satoshi Tadokoro, "Unknown Object Modeling on the Basis of Vision and Pushing

Manipulation," 2011 IEEE International Conference on Robotics and Biomimetics (ROBIO2011),pp. 1942-1948 ,2011.

（2）特許出願

研究期間累積件数：0件

（3）その他の成果（主要な学会発表、受賞、著作物等）

1. 2011年11月 SSRR2011 Best paper finalist 受賞
(Kazunori Ohno, Shinji Kawatsuma, Eijiro Takeuchi, Kazuyuki Higashi, Satoshi Tadokoro and Takashi Okada,"Robotic Control Vehicle for Measuring Radiation in Fukushima Daiichi Nuclear Power Plant," Proc. of IEEE International Workshop on Safety, Security, and Rescue Robotics (SSRR2011), pp. 38-43, 2011 に対して)

2. 2011年12月 福島第一原発の Quince の活躍に対して東京電力から感謝状

3. 2012年1月 第7回競基弘賞特別賞を受賞（Quince 開発チーム千葉工業大学 小柳栄次教授、吉田智章研究員、西村健志氏、東北大学 田所諭教授、永谷圭司准教授、桐林星河氏、岡田佳都氏、大竹一樹氏、大野和則客員准教授、竹内栄二朗助教、東和幸氏、工学院大学 羽田靖史准教授に対して）

4. ロボティクス, 機械学会出版 3 章, 5 章分担, 2011（ISBN 978-4-88898-208-5 C3053 ）.

5. ロボットテクノロジー, 日本ロボット学会, レーザ距離計について分担執筆, 2011(ISBN 978-4-274-21072-3).

研 究 報 告 書

「大規模並列化によるハイパフォーマンス人工知能システム」

研究期間: 平成20年10月～平成24年3月
研 究 者: 岸本 章宏

1，研究のねらい

本研究では、大多数の計算機を利用して、人工知能技術の超高速化の研究を遂行する。本研究の最終目標は、数千 CPU コアを利用した計算機環境でもアルゴリズムの高速化が達成できる並列化手法を開発することである。並列化の対象には、人工知能分野の基盤技術である探索アルゴリズムを主に利用する。多くのアプリケーションで用いられる探索アルゴリズムは、組み合わせ爆発の生じる大規模な空間を高速に探索し、実時間で解を求める必要があるので、大規模並列化による高速化は、アプリケーションの性能改善のための重要なアプローチである。また、本研究は、探索空間を大規模な情報とし、解を有益な情報とみなせば、本領域の「大規模な情報から知識を生産・活用するための基盤技術の創出」という目標に合致する。

2，研究成果

本研究の主な成果は次の通りである。

（1）プランニングにおける大規模並列探索アルゴリズムの開発

プランニング・システム(プランナー)では、ユーザーが解きたい問題を PDDL や STRIPS などのプランニング言語で記述すれば、ユーザー自身が解法を考えなくてもプランナーが自動的に解を求めてくれるので、ユーザーにとっては様々なドメインに簡単に適用できるシステムである。このため、汎用プランナーの産業界からの需要は、将来高まっていくことが予想される。

本研究では、高性能なプランナーであり、研究者間で最も広く用いられている Fast Downward を並列化した。具体的には、Fast Downward では、人工知能分野の基盤アルゴリズムである A*アルゴリズムが用いられており、この A*アルゴリズムの大規模分散並列版である HDA* (Hash Distributed A*)を開発した。

分散並列 A*の問題点は、どのようにして効率よくプロセッサ間で探索結果を保持し、その探索結果を再利用することで無駄な並列探索を省略できるかにある。逐次 A*では、探索結果の参照の際にローカルのメモリをアクセスすればよいだけであり、この問題自体が生じないのに対し、分散並列 A*では、ある計算ノードが他の計算ノード上にある探索結果にアクセスするためには、通信や同期オーバーヘッドが生じる。これが分散並列 A*の性能低下の原因である。

本研究の HDA*では、データ駆動型スケジューリングという考え方を提案し、データ（ハッシュ表やテーブルのエントリー）のあるところに必ず仕事（計算）を移動させることによって、探索結果の効率的な再利用を実現しつつ、非同期並列計算を駆使できるようにした。その結果として、HDA*は、並列アルゴリズムの性能を低下させる大きな原因である通信遅延を

隠ぺいできるようになり、高い高速化率を達成できるようになった。

HDA*の性能評価は、プランニング分野で最もスタンダードな問題集を解かせることによって行った。その結果、HDA*は 1024 コアで最大 600 倍の高速化を達成しただけでなく、分散環境の莫大なメモリ(1-2TB)を利用することによって、これまでの高性能なプランナーで解けなかった問題のいくつかを高速に解くことに初めて成功した。

（2）大規模並列ゲーム木探索アルゴリズムの開発

ゲームは、社会で利用されている実アプリケーションに比べ、ルールと結果が単純であるのにもかかわらず、計算量の組み合わせ爆発が生じるので、難しい問題である。このため、人工知能分野における理想的な題材として、約60年にわたって研究されてきた。

UCT アルゴリズムは、最近特にコンピュータ囲碁の分野で注目されているゲーム木探索アルゴリズムである。ルート並列化は、UCT アルゴリズムの並列化法であり、簡単でかつ広く用いられている。

本研究では、まずルート並列化を改良(合議制ルート並列化と呼ぶ)し、ルート並列化との性能を囲碁プログラムで比較した。性能評価実験では、両手法を取り入れた囲碁プログラムで、64CPU コアを利用して、プログラムの強さを調べた。合議制 Root 並列化は、Root 並列化よりも性能が良かったが、CPU コア数を増加させれば、両手法には大きなボトルネックがあることが分かった。 このため、合議制ルート並列化やルート並列化では、非常に限定的な性能向上しか達成できず、64 コア以上の CPU を利用しても性能が改善しないことが分かった(64 コアでも 2-3 倍程度の高速化)。性能が向上しない原因としては、UCT の構築する探索木が大きければアルゴリズムの性能が向上するのにもかかわらず、ルート並列化では、この探索木をコア間で共有しないために、ルート並列化が構築する探索木の実質的な大きさは、逐次 UCT とほぼ同じであることが挙げられる。

本研究では、以上の通り、ルート並列化の限界を示した後、プランナーの大規模分散並列化でも利用した考え方である、データ駆動型スケジューリングを駆使し、探索木を CPU コア間で効率良く共有する並列 UCT アルゴリズムを開発した。この新しい並列 UCT では、データ駆動型スケジューリングの特長である計算の非同期化と、ルート並列化では行えなかった、メモリの効率的な利用による探索木のサイズの巨大化が実現できたのであるが、特定の CPU コアが過剰のメッセージを受け取ってしまうという問題が生じた。そこで、UCT と同等の性能を持つ深さ優先探索アルゴリズム DFUCT を並列化の対象に変更し、データ駆動型スケジューリングに基づいて、大規模分散並列化を行った。この大規模分散並列 DFUCT では、スタンダードなテスト問題である pgame において、4800CPU コアで最大で約 1800 倍という高い台数効果が得られた。

（3）逐次探索アルゴリズムの改良

本研究に着手してから、A*アルゴリズムを代表とする多くの逐次探索アルゴリズムでは、そもそもメモリを大量に利用するため、大規模並列化によってメモリサイズを大幅に増やしたとしても、メモリ不足のせいで解を求められないことが多いことが判明した。このため、先行研究のアルゴリズムよりもメモリの利用量が大幅に少ない逐次探索アルゴリ

ズムをいくつか開発し、プランナーやゲーム・プログラムの能力を向上させることに成功した。これらの新しい探索アルゴリズムの並列化は、研究成果(1)(2)で駆使したデータ駆動型スケジューリングを利用すれば、効率良く行えると期待できる。
さらに、15 年以上前に発表され、基礎技術として広く用いられている探索アルゴリズム EIDA* (Enhanced IDA*)アルゴリズムに間違いがあることを発見した。本研究では、この EIDA*の誤りを正しく修正し、理論的に常に正しい結果を返せることを保証した。

（4）Binary Decision Diagram (BDD)構築の並列化

データ駆動型スケジューリングを用いた分散並列化の3つ目の対象アルゴリズムとして、BDD 構築を選んだ。BDD は、二値論理関数のコンパクトな表現方法であり、ハードウェア回路検証やデータ・マイニング、機械学習やパズルなど、様々な応用分野がある。
本研究では、北海道大学の湊真一研究室で開発されている BDD パッケージの逐次ソース・プログラムを入手し、並列化を行った。テストドメインには、N-Queen 問題の解を列挙する問題を利用した。本研究で行った並列化のアプローチは次の二つである。
まず、BDD 構築アルゴリズムの性質を調べるため、8 コアを持つ共有メモリ計算機上でアルゴリズムの並列化を行った。この並列化では、3.5 倍程度の高速化を達成できた。
次に、データ駆動型スケジューリングを駆使し、分散環境で BDD 構築の大規模並列化を行ったが、600 コアを利用した場合に約 20 倍程度遅くなってしまった。この性能低下は、BDD 構築アルゴリズムがデータ駆動型スケジューリングを利用したとしても、ゲームやプランニングよりも通信遅延を隠ぺいしにくい手法であることに起因した。この問題の解決策は、今後の課題にしたい。

3, 今後の展開

本研究で成功を収めたアルゴリズムである HDA*や大規模分散並列 DFUCT は、当然プランニングやゲームにそのまま利用できる。しかし、本研究で開発したアルゴリズムで共通して利用する並列化法のアイデア（つまり、データ駆動型スケジューリング）をどの程度まで一般化できるかを考えれば、他のアルゴリズムの並列化にも(例えば、分散並列ダイナミックプログラミング)利用できる可能性が生じるので、本研究の適用範囲が自然に広がる。

HDA*と大規模分散並列 DFUCT では、データ駆動型スケジューリングを利用する点が同じであるのにもかかわらず、現状では、対象とするアプリケーションを並列化する際には、それぞれのアプリケーションに合わせて、データ駆動型スケジューリングのソースコードをスクラッチから記述している。

データ駆動型スケジューリングを一般的に適用できるようにするためには、このような職人芸的な開発スタイルから脱却し、他の研究者やプログラマーがもっと手軽にデータ駆動型スケジューリングに基づく並列プログラムを開発できる環境を整備しなければならない。例えば、並列ライブラリや並列プログラミング言語などを設計・公開することによって、大規模分散並列アルゴリズムの開発を支援できれば理想的である。今後の研究の展開としては、プログラミング言語研究者と協力しながら、データ駆動型スケジューリングに介在する基本性質を解析し、その知見に基づいて、データ駆動型スケジューリング用の基本プリミティブを設計することが挙げられる。

4，自己評価

本研究の当初の予定は、探索アルゴリズムの代表的なアプリケーションであるゲームとプランニングを研究題材として、数千コアを用いた大規模並列化を行うことであった。実際に、我が国を代表とするスーパーコンピュータである東京工業大学の TSUBAME 上で1000コア以上用いた実験を行えたこともあり、当初の計画よりも早いペースでこれらの分野で効率の良いアルゴリズムの開発を進めることができただけでなく、予想以上の大きな研究成果を出すことができた。特に、プランニングでの成果について執筆した論文は、当該分野の最高峰の国際会議 ICAPS'09 で最優秀論文賞を受賞（ICAPS で日本人が最優秀論文賞を受賞するのはおそらく初めての快挙である）し、本研究のインパクトの高さを国際的に示せたと考えている。

当初の計画よりも研究が順調に進んだため、さきがけ開始当初には考えていなかった研究である、逐次探索アルゴリズムの性能改良と BDD 構築の並列化を行った。逐次探索アルゴリズムについては、15年以上にわたり利用されているアルゴリズムの理論的な欠陥を修正したり、プランナーやゲーム・プログラムの能力を大幅に向上させたりするなど、重要な成果をあげることができた。開発したこれらのアルゴリズムは、ドイツやカナダの大学院のセミナー等でも取り上げられ、これらのアルゴリズム改良に関する研究が国際的に進み始めている。BDD 構築の並列化に関しては、8コア程度の共有メモリ型計算機上ではある程度の成功を収めたものの、数百コアを利用する大規模分散並列 BDD 構築アルゴリズムの開発には成功しなかった。この BDD 構築の分散並列化に関しては、新たな手法を考えて、再挑戦したい。

人工知能分野での代表的なアルゴリズム開発・改良では成功を収めたものの、本研究領域での目標の一つである情報社会への応用については、社会で実際に利用されているシステムに組み込むという段階までには、残念ながら到達できなかった。この原因の一つとして、最も適用範囲の広いアルゴリズムである BDD 構築の大規模並列化が成功しなかったことが挙げられる。しかし、今回の研究で駆使した考え方である「データ駆動型スケジューリング」自体は一般的な考え方であり、このスケジューリングに基づく並列プログラミング言語や並列ライブラリを開発できれば、さきがけで得られた知見が社会で実際に利用される機会は自然と広がると考えられる。この点については、5-10 年にわたる長期的な計画であるが、今後の大きなテーマにしていきたい。

5，研究総括の見解

囲碁等のゲームを素材とした大規模探索というテーマである。探索は大規模データからの知識獲得のための基本方式であるため、今後、学習や知識発見との関わりを付けて、知識創出に関係する研究の方向が出てくることを期待していた。

当初の計画よりも早いペースで、ゲームとプランニングの分野で、効率の良い大規模並列化アルゴリズムの開発を進めることができ、予想以上の大きな研究成果を出すことができている。また、当初の計画よりも研究が順調に進んだため、計画にはなかった逐次探索アルゴリズムの性能改良と BDD 構築の並列化も行っており、高く評価したい。

一方、アルゴリズム開発・改良では成果を上げているが、情報社会への応用については十分な成果を得られていない。今回の研究で駆使した考え方である「データ駆動型スケジューリング」に基づく並列プログラミング言語や並列ライブラリを開発できれば社会で実際に利用される機会

が広がると考えられ、今後のテーマにしていってほしい。

6, 主な研究成果リスト

(1)論文(原著論文)発表

1. Akihiro Kishimoto, Alex Fukunaga, and Adi Botea. "Scalable, Parallel Best-First Search for Optimal Sequential Planning", In Proceedings of the 19th International Conference on Automated Planning and Scheduling (ICAPS'09), pages 201-208, 2009.
2. Yusuke Soejima, Akihiro Kishimoto and Osamu Watanabe. "Evaluating Root Parallelization in Go", IEEE Transactions on Computational Intelligence and AI in Games, Volume 2, Number 4, pages 278-287, 2010
3. Yuima Akagi, Akihiro Kishimoto and Alex Fukunaga. On Transposition Tables for Single-Agent Search and Planning: Summary of Results. In Proceedings of the 3rd Symposium on Combinatorial Search (SoCS'10), pages 2-9, 2010,
4. Tatsuya Imai and Akihiro Kishimoto. "A Novel Technique for Avoiding Plateaus of Greedy Best-First Search in Satisficing Planning", In Proceedings of the 25th AAAI Conference on Artificial Intelligence (AAAI-11), pages 985-991, 2011
5. Kazuki Yoshizoe, Akihiro Kishimoto, Tomoyuki Kaneko, Haruhiro Yoshimoto and Yutaka Ishikawa. "Scalable Distributed Monte-Carlo Tree Search", In Proceedings of the 4th Symposium on Combinatorial Search (SoCS'11), pages 180-187, 2011

(2)特許出願

研究期間累積件数:0 件

(3)その他の成果(主要な学会発表、受賞、著作物等)

招待講演

1. 岸本章宏, Alex Fukunaga, Adi Botea.「最適解を求めるプランニング・アルゴリズムの大規模並列化について」, 第22回回路と軽井沢システムワークショップ, 2009.

著書

1. 小谷善行、岸本章宏、柴原一友、鈴木豪.「コンピュータ数学シリーズ 7 ゲーム計算メカニズム - 将棋・囲碁・オセロ・チェスのプログラムはどう動く -」, コロナ社, 2010.

受賞

1. Best Paper Award, 19th International Conference on Automated Planning and Scheduling (ICAPS'09), 2009.
2. 第23回人工知能学会全国大会優秀賞, 2009.
3. Best Paper Award, 13th Advances in Computer Games, 2011.

研 究 報 告 書

「擬似コード変換と統計解析による文書画像からの知識抽出」

研究期間: 平成 20 年 10 月～平成 24 年 3 月
研 究 者: 寺沢 憲吾

1，研究のねらい

ディジタル技術の普及と発展に伴い、多くの文献・史料がディジタル画像化され公開されている。これにより従来は図書館や博物館の奥深くに所蔵されていた貴重な文献・史料を、一般のユーザがネットワークを介して容易に入手・閲覧することが可能となってきている。

このようにディジタル技術が文献・史料という知識の共有・流通に対して大きな成果を上げている一方で、これらのディジタル化された文書画像を活用するための情報処理技術は未だ発展途上段階である。文書画像に対して全文検索やテキスト解析・マイニングといった情報技術を適用するには、画像を正確にテキストデータに変換することが望ましいが、OCR（光学文字認識）技術を用いて文書画像をテキストデータ化する方法は歴史的文書や手書き文書を対象とする場合に常に適用可能であるとは限らない。OCR は言語や書体に依存した技術であるため、時代が古いものや特殊な文字あるいは用法を含むもの、また十分なサンプルが確保できないものには適用できない。また手書き文字の場合は精度が低下するため正確なテキストデータの作成のためには専門家の手による修正作業が不可欠であり、あらゆる文書画像をこの方法でテキストデータ化するのはコストの問題で現実的ではない。

本研究のテーマは、画像をテキスト化してから解析するのではなく、画像を画像のままで解析する手法を確立することである。具体的には、文書画像を画像特徴量による擬似コード表現に変換することにより、文書画像データに対する高速な全文検索法を開発するとともに、統計解析による知識抽出のための技術として、頻出語句の抽出、語句の頻度分析による文書の特徴づけ、特徴に基づく文書間の関連性の記述、共起関係等による語句間の関連性の記述などの方法を開発する。こうした一連の手法を確立させることにより、従来「取り扱いにくい」データであった画像データを、「取り扱いやすい」テキストデータと同様に知識創出のための情報資源として役立てることが可能となる。いわば、画像データとテキストデータとの間の架け橋である。この取り組みを通して、ディジタル文書画像という知識の宝庫を、共有・流通という第一段階から、それを基礎とした知識の抽出、さらには知識の創出という次の段階へ進めるということが本研究のねらいである。

2，研究成果

本研究は大きく分けて、文書画像を擬似コードに変換すること、擬似コードを用いた文書解析アルゴリズムを構築すること、それらを用いて実際に利用可能なアプリケーションを構築し、文献研究などの用途に役立てることの3つの柱からなる。以下にそれぞれの成果について述べる。

（1） 文書画像の擬似コードへの変換

本研究の核となる擬似コード LSPC は、文字の形を記述する画像特徴量（高次元の実数ベクトル）を、その記述性能を大きく損なうことなく、比較的低次元の自然数の組として表現する手法で

ある。この変換には、近傍探索問題の解法の1つである LSH のインデックスを用いるが、ここで、ノルムが1に正規化されているベクトルの集合に対して従来の LSH より有効な SLSH（Spherical LSH）を用いることで、擬似コードの性能をさらに高めている。この手法自体はさきがけ研究以前に開発したものだが、さきがけ研究期間中に、SLSH を実際に大規模データに対して適用した場合についての検証評価を行い、有用性を確認した。この研究は論文「Approximate Nearest Neighbor Search for a Dataset of Normalized Vectors」として発表し、電子情報通信学会論文賞を受賞し、高く評価された。

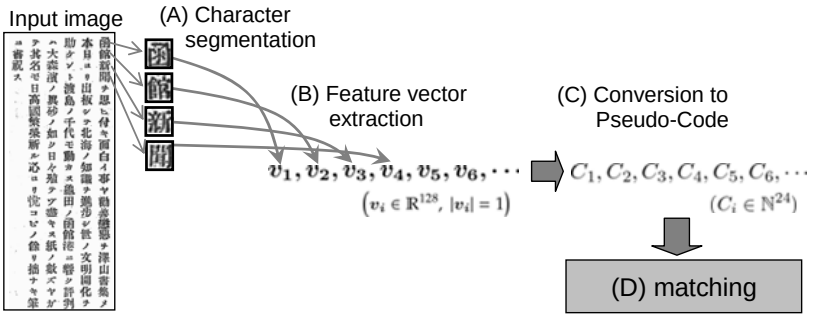
（2） 擬似コードを用いた高速検索手法の確立

擬似コードを用いることで、通常の画像ベクトルに基づく検索よりもきわめて高速な検索が可能であることを示した。「函館新聞」（北海道で最初の民間新聞・明治 11 年～明治 31 年発行）のうち明治 11～17 年の 7 年分、859 万文字から 5～10 文字の文字列を検索するのに、所要時間が1秒未満（線形探索を用いた場合）という高速検索を可能にした。また、Ex-BMH 法というアルゴリズムを開発し、文字列長が 9 文字以上の場合に、線形探索よりも高速な検索が可能であることを示した（図 1，図2，表1）。

（図1）「函館新聞」明治 14 年

（表1）「函館新聞」に対する全文検索時間

検索文字列長	線形探索[秒]	Ex-BMH[秒]
5 文字	0.90	1.23
6 文字	0.81	1.08
7 文字	0.90	1.01
9 文字	0.91	0.86
10 文字	0.87	0.81



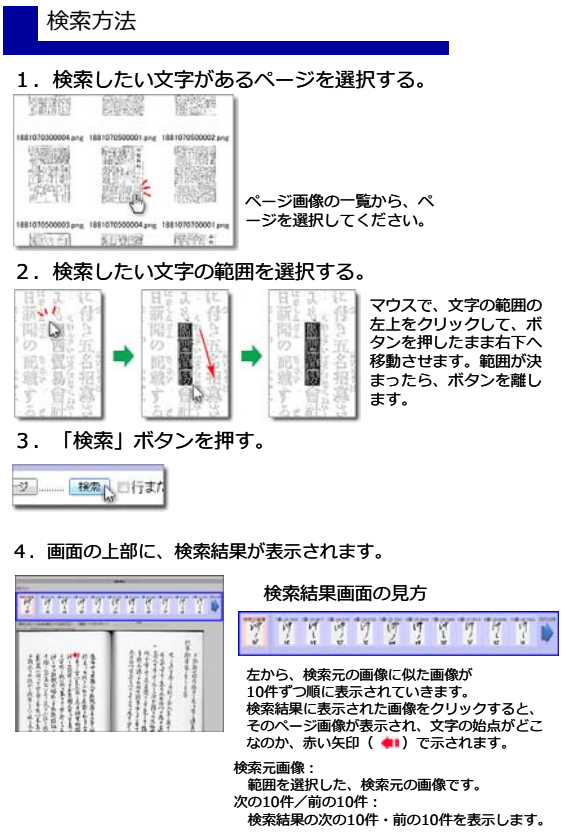
（図2）開発した手法の概略図

（3） 実用アプリケーションの構築

本研究で開発した全文検索システムをインターネットを通して一般のユーザが使用できるよう、ウェブシステムを構築し、平成 23 年 5 月に第1段階として、函館市中央図書館の協力を得て、同図書館所蔵の文献の中から、「亜国来使記」（1854 年 4 月 ペリーが箱館訪問した際の松前藩の応接記録）および「函館新聞」（北海道で最初の民間新聞・明治 11 年～明治 31 年発行）のうち明治 14 年の1年分を公開した。html サーバは専用のものを用意し、本学に設置した（http://records.c.fun.ac.jp/）（図3，図4）。また、函館市中央図書館デジタル資料館のページ（http://lib-hkd.jp/rein/）からもリンクを設定し、誘導した。さらに、平成 24 年 1 月には、京都大学文学研究科と連携し、京都学派アーカイブ（http://kyoto-gakuha.info/）で公開されているコンテンツの一部に対し、全文検索を可能にして公開した。

実用アプリケーションへ検索エンジンを提供する活動は上記以外にも複数行っている。一例として、京都大学文学部の林晋教授が開発している歴史学、文献学などの人文学におけるテキスト研究用のツールである SMART-GS（http://sourceforge.jp/projects/smart-gs/）へ検索エンジンを提供している。SMART-GS は検索の他にリンクやマークアップ、コメントの添付などの機能を持つ多機能なツールであるが、提供した検索エンジンは独立した部品として動作するよう意識して設計したため、SMART-GS の各機能の開発と検索エンジンの開発とは独立に行うことが可能であった。また別の例として、民間企業から産学連携の打診があり、この検索システムを活用した新しいウェブサービスの開発に着手している。

（図3）実際に公開されている「文書画像検索システム」のタイトル画面.
URL: http://records.c.fun.ac.jp/



（図4）オンライン検索システムの操作方法

3，今後の展開

擬似コードに対する解析アルゴリズムの開発に関しては、依然研究開発の余地が残っている。研究のために必要なデータや、基礎的な知見は既に得られているので、今後はこれをさらに進展させ、まだ未開発の部分の開発を進めていきたい。擬似コードに対して解析を行うことで文書の解析がある程度まで可能であることを自らの手で立証し、こうした研究に対するフォロワーを生み、研究手法の一つの潮流を確立させることが将来的な大目標である。

これまでに引き続き、人文系研究者と連携して、人文学の研究のために情報技術が提供できることを模索していく。情報システムのアシスト無しには見いだせなかった人文学上の新たな知見を継続して生み出せるようになれば大きな収穫である。

検索エンジンをウェブサービスとして公開させたので、すでに利用者からの意見や要望が寄せられつつある。今後はこうしたフィードバックをふまえ、システムの改良を図っていきたい。それとともに、コンテンツを増強させて研究成果のアピール力を高め、ディジタルライブラリーの利活用手法の一つとして定着するところまでを目指していきたい。

研究期間中に、産学連携で民間企業と共同研究を行うための交渉を進めることができたので、今後もこの活動を継続させていく。研究者自身の専門である画像処理とアルゴリズムの研究にプラスして、連携先企業の得意とするウェブサービスを融合させることで、新しい情報技術・ウェブサービスを実際に稼働させ、新たな知の創生の苗床とすることを目指していく。

4，自己評価

当初計画と比べ、研究成果を社会に還元する、研究のアウトリーチ活動は想定以上の進展を得た。これには、イノベーション・ジャパン 2009 での展示発表において多くの企業の方に本研究内容を紹介することができ、ウェブサービスを得意とする企業とのマッチングが生じ、産学連携を見据えた協力関係を構築することができたことが大きく貢献している。また、研究中に研究者自身が公立はこだて未来大学へ着任した後、函館市中央図書館との連携が円滑に進んだことも要因の一つである。さらには、研究領域アドバイザーの支援の下、人文学の研究者との協力関係を構築することもできた。また、当初計画より早期にウェブサービスを公開することができたため、一般ユーザの利用履歴データの解析にも着手することができた。

一方で、アルゴリズムの開発を中心とする理論的研究に関しては、部分的には成果が得られたものの、当初目標として掲げた計画を期間内にすべて達成するには至らなかった。ただし、この研究活動は今後も継続していく予定であり、今後当初計画にある成果を得るための足がかりとなる部分は確立できたのではないかと考えている。

そうした点を総括すると、この研究のねらいとして掲げた「画像データとテキストデータの間に橋を架けること」に関しては、研究期間内に、橋を架けることはできたと評価してよいのではないかと考えている。橋の上の交通量をさらに増やしていくための研究活動は今後も引き続き継続していく予定である。

5，研究総括の見解

文字認識不能な文書画像を画像のまま部分検索などできるようにする重要な基盤技術の提案である。実績は高く、技術としての有用性が高い課題であった。

研究成果を社会に還元する、研究のアウトリーチ活動は当初計画を超えた進展を得ている。函館市中央図書館との連携、人文学の研究者との協力関係を構築できている。さらにウェブサービスの公開により、一般ユーザの利用にもつながっていることは高く評価できる。情報検索の分野に大きく貢献したと考える。

この研究のねらいの「画像データとテキストデータの間に橋を架けること」に関しては達成している。これまでに存在しなかった研究分野を開拓した功績は大きい。人文科学における新しい道具ができたので、それを活用した研究成果も期待できる。今後、更に効率の良いアルゴリズムの開発を中心とする理論的研究を含め、この活動が個人を超えて発展することを期待する。

6, 主な研究成果リスト

（1）論文（原著論文）発表

1. K. Terasawa and Y. Tanaka, "Approximate Nearest Neighbor Search for a Dataset of Normalized Vectors," IEICE Transactions on Information and Systems, vol.E92-D, no.9, pp.1609-1619, 2009. （平成 21 年度（第 66 回）電子情報通信学会論文賞受賞）

（2）特許出願

なし

（3）その他の成果（主要な学会発表、受賞、著作物等）

【学会発表】
1. K. Terasawa, T. Kawashima, Y. Tanaka, "The Extended Boyer-Moore-Horspool Algorithm for Locality-Sensitive Pseudo-Code," VISAPP 2011, International Conference on Computer Vision Theory and Applications (Part of VISIGRAPP 2011), pp.437-441, Algarve, Portugal, Mar. 5-7, 2011.
2. K. Terasawa, T. Shima, T. Kawashima, "A Fast Appearance-Based Full-Text Search Method for Historical Newspaper Images," ICDAR2011, 11th International Conference on Document Analysis and Recognition, pp.1379-1383, Beijing, China, Sept. 18-21, 2011.
3. 寺沢憲吾, 川嶋稔夫, "文書画像からの全文検索のオンラインサービス", 人文科学とコンピュータシンポジウム「じんもんこん 2011」, pp.329-334, 京都, 2011 年 12 月.
【公開したシステム】
1. 文書画像検索システム http://records.c.fun.ac.jp/

研 究 報 告 書

「健康被害を監視するための多言語ウェブサーベイランスシステム」

研究期間： 平成 20 年 10 月～平成24年3月
研 究 者： 国立情報学研究所・情報学プリンシプル研究系・准教授ナイジェル コリアー

1，研究のねらい

マスコミなどの公的でないデジタルソースによる、疾患の集団発生に関する情報の収集を行うシステムは今や国内ならびに海外の公衆衛生機関において、重要視されるものとなっている。富裕国では、市販や一般開業医のネットワークといった高度な情報源に溢れているとはいえ、すべての国々にこういったシステムを実施、あるいは維持するソースがあるわけではない。A(H5N1)といった新興疾患の急激なまん延への懸念により、流行性疾患情報システムへの関心が高まってきている。そのシステムは、世界規模での事象を検出し、指標ネットワークを補い、ソースに密接した状態での疾患の対処を可能にするものである。BioCaster BORN(生化学・放射性物質・核)と呼ばれる、この研究プロジェクトの最終目的は、広範囲の感染疾患、ならびに化学、放射性、核による病原物質を早い段階で警戒するため、テキストマイニング技術を基にした、完全稼動状態のウェブ上知的サーベイランスシステムを開発することである。この研究は公衆および動物衛生コミュニティーの研究者らによる緊密な協力のもと、世界規模での健康の向上のため、調査結果の利益が、できる限り広範囲に広がるよう実施されたものである。

2，研究成果

この研究は、いくつか重要な点に関して、我々のこれまでの知識を拡大するものである。(1)歴史的背景に反しての、発生事象の重要性を理解できるよう変化点検出の調査を行った。(2) 化学、放射性、核による食物、水および環境の汚染によって起こる疾患の発生といった広い範囲での健康への脅威に対する知識モデルを製作した。知識モデルには 12 言語での専門用語が含まれる。(3)地理的および時間的類似点のほか、個々の公衆衛生事象の関連性に関する知識を用い、健康への脅威に関する情報を融合した調査を実施した。このワークパッケージに関しては、個別に述べることとする。

この研究を経ての最終的な BioCaster BORN システムは、公共のウェブサーバ(http://born.nii.ac.jp)に組み込む専用の高性能コンピューティング・クラスター上で作働する、モジュール化したテキストマイニングパイプライン(図 1)で構成されている。 システム内のモジュールは言語探知、機械翻訳、文献分類に効率的な自然言語処理アルゴリズムのほか、用語とその関連性を特定し、事象を特定した場合に、そういった事象の危険を世界中の公衆衛生コミュニティーに警告する専用のモジュール類で構成されている。こういったさまざまなモジュールは、疾患、動物種、症状、病原体などに対して語彙の分類および個々の関連性を定めるドメインの高度な知識モデルと統合されているのである。

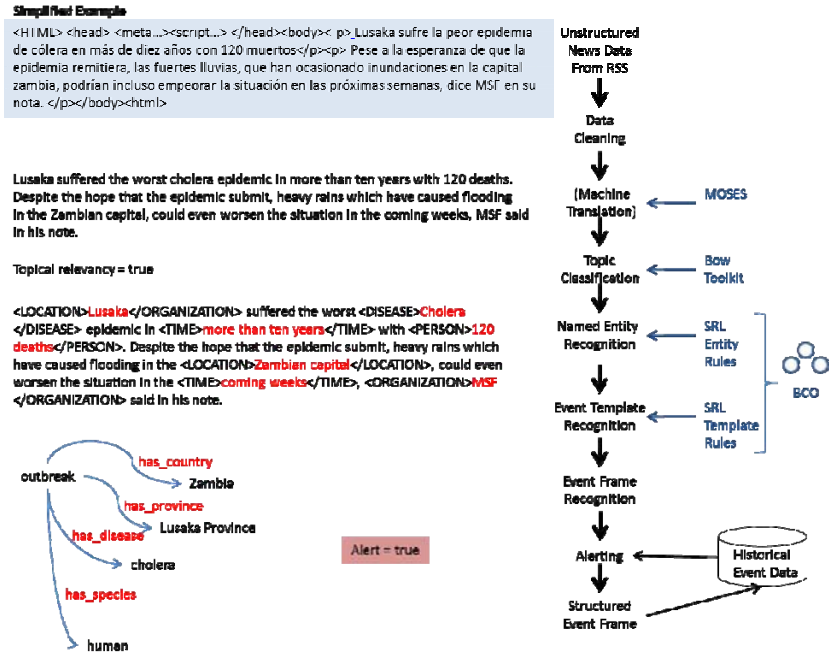


図 1 BioCaster BORN テキストマイニングパイプライン

Google ニュースのほか、ProMed メール、香港特別行政区伝染病注意項目リスト、国際獣疫局警戒項目リスト、ヨーロッパメディアモニター警戒事項リスト、国際獣疫機構(OIE)警戒リスト、 ヨーロッパメディア監視警告事象やアラートネットといった公共または NPO ソースから、通常 1 日約 27,000 のニュース項目が BioCaster BORN によって監理されている。さらに、プロジェクト期間中(2008.10～2012.3)、我々は、eltWater という民間のニュースアグリゲーション会社から新たに 80,000 のニュースソースの使用許諾を受けた。

［変化点検出を用いたニュースの事象警報］

このワークパッケージでは、私は、オンライン衛生関連ニュースの事象を用いて、日々の警報に関して変化点検出アルゴリズムを序列的に評価するうえでの問題点に取り組んだ。毎日の各国の疾患数は BioCaster BORN を用いて、実際の世界中のデータをテキストマイニングした。18 の BioCaster BORN による英語のニュースデータ 18 セットに対して、公衆衛生ドメインで広く用いられる 5 種の異常検出アルゴリズム(EARS C2、C3、W2、F 統計量、EWMA)の、専門家が管理する ProMED メール投稿のシルバースタンダードに対する性能を比較した。ProMED メールは、国際感染症学会の公的プログラムとして、専門家ボランティアが世界規模で、メディア報道、その他のソースを、動物、植物に影響を及ぼす生物的ならびに科学的災害に関する情報についてモニタリングを行っている、毎日 24 時間稼動するボランティアのヒューマンネットワークである。

14 ヶ国における 366 日以上の期間での 18 の発症例に関する 287 件の ProMED メール投稿に関して、システムの感受性、特異性、陽性的中率(PPV)、陰性的中率(NPV)、100 日間の平均警報、および F1 は 95%の信頼区間(CI)で報告された。 F1 に重点を置き、誤警報率および 100 日あたりの平均警報数を公衆衛生分析における最重要基準とした。結果は、W2 に最良の F1 が認められ、C2 に比べてわずかに週の曜日効果を示した。これは、おそらく W2 によって用いた層別サンプリング法によって週の曜日効果が相殺されたためであると考えられた。ドリルダウン分析で

は、国の水準モデルのグラニュラ選択によって起こる問題点ならびに週の曜日効果および報道の偏りによる急激なレポート量の減少について示唆した。選択,その他の研究報告と同様に、私の研究結果は、最終的な検出能力に悪影響を及ぼす事象 1 例に対するニュースの報道が長期にわたる傾向にあることを示すものであった。ニュース報道量に急激な増加がみられた例として、イラクでのコレラ、イギリスでの麻疹があるが、こういった例は、段階的にニュースが増加した米国での A(H1N1)インフルエンザのような事象と比較すると、容易に警戒態勢を取れる傾向にあった。この結果を基に、私は、多言語でのニュースレポートを用いて生じる利益について検討した。一連の新たな実験において、5 種の同じ時間的異常検出アルゴリズムを用い、16 件の疾患発生例の進化を追跡した。さらに、ProMED の報告例 をシルバースタンダードとして用いたところ、129 日以上にわたる試験期間での 13 言語に対する新たなデータの比較分析の結果、さまざまな言語での事象を用いたモデルの多くに、感受性ならびに適時性の向上が認められた。
この結果は、多言語テキストマイニングを用いた自動健康サーベイランスには、インフォームド・チョイスを用いてモデル選択、データソースの管理を行う場合、低価値の情報を警戒事象に変換する可能性があることを示すものであった。 BioCaster BORN ポータル上では多言語でのニュースを用いた C2 警報アルゴリズムの実行が可能となった。

［公衆衛生脅威モデリング］

もうひとつのワークパッケージにおける重要成果は BioCaster オントロジー (BCO)である。これは、疾患報告における一般人向けの言語統一のために考案され無料で提供された公衆衛生アプリケーションの 12 言語のオントロジーである。 (http://code.google.com/p/biocaster-ontology)

BCO の目的

- ニュースでの公衆衛生事象の検出およびリスク評価に必要な、用語ならびにその関連性の説明
- （多言語での） 生体医学上のグレー・リテラチャーと従来の基準のギャップを補充
- 言語間の内容統一を仲介
- 無料提供

BioCaster BORN にとって BCO は主要な知識ソースであり、疾患、薬品、症状、症候群、動物種といったドメイン用語のほかに諸症状を引き起こす疾患、または特定の宿主動物種に作用する薬品といったドメイン感受性の関連事項を含んでおり、これによってテキストマイニングシステムがドメイン内での主要概念や関連性を認識し、ニュースでは明確にされない箇所を補うことができるのである。BCO は、公開メディアの言語での疾患発生サーベイランスに関心のあるシステム開発者に対して、多言語でのサポートを無料で提供しており、我々が知る限り、アプリケーションオントロジーとしてはユニークなものである。
BCO は現在、ヒト、動物の 300 種以上の疾患を、国連が公用語とする 12 ヶ国言語（アラビア語(968 語)、英語(4113 語)、フランス語(1281 語)、インドネシア語(1081 語)、日本語(2077 語)、韓国語(1176 語)、マレーシア語 (1001 語)、ロシア語(1187 語)、スペイン語(1171 語)、タイ語(1485 語)、ベトナム語(1297 語)、中国語(1142 語)）で網羅している。 BCO は、ウェブ上のオントロジー言語(OWL)ならびに単一知識機構システム(SKOS)フォーマットにて、ダウンロードによる入手が可能である。 これは BCO のウェブサイトで無料提供されており 現在までで、各国 350 以上のグループがオントロジーのダウンロードを行っている。

[地理−時間的理解による事象の融合]

このプロジェクトで始めた最大の難課題のひとつがコンピュータに、状況に関してさらなる知識を与えるという試みである。最初のワークパッケージでの作業は、一文章で報告された事実の把握に基づくものであった。さらなる難題として、ある状況の中で、事実が意味を成さなければならない。この問題に取り組むため、私は共同研究者らの協力を得て、解決策となりうる事象オントロジーならびに地理−時間的アノテーション・スキーマを作成した。'入院する'、'検疫'、'治療を受ける' といった事象が疾患の発生における中心的な概念とはいえ、疾患、症状、発生場所といった観察対象物と比較すると、その理解はより困難である。DOLCE（言語と認知工学のための記述的オントロジー）から適用した方法を用いて、BioCaster 事象オントロジー(BCEO)が感染疾患の公衆衛生ドメインで、事象の正式な定義ならびに疾患関連事象を示す表現を提供する。事象オントロジーの初期のバージョンでは、40 例事象の種類に対して、同義語ならびに事象論理上の正式な記述が含まれたものが公開された。

地理−時間の認識は、コンピュータが正確なニュースレポート処理を行うために理解が必須である、また別の複雑な課題である。私は共同研究者とともにアノテーション・スキーマ、ならびに事象発生の場所と期間をニュース文書内より特定するメソッドを作成した。その技法によって、言語の特性による分類に基づき、特殊事象を一般的あるいは仮説上の事象から分類するとともに事象の場所を詳細に特定する手段が与えられた。自動的に対象物およびその関連性(同一性、重複、原因など)に関する記述をニューステキストより検出、統合する機械学習ツールのトレーニング用に、BCEO および地理−時間的標識法に基づき、大規模な、事象に関する注釈つきコーパスの構成を開始した。

3, 今後の展開

3 種のワークパッケージにおける研究から得た見識を基に、今後私は以下の疑問点に答えていこうと考えている。 (a)取り込んだ多変量の事象特性に対する警報の利点は何か。(b)極めて稀な事象に対する警報方法の開発はどのように行えばよいか。(c)'未分類インフルエンザ'のような一般疾患に関する報告などの不特定報告への警報のサポートにはどういった意味的特性が警報最良であるか。(d)発生場所の自動検出アルゴリズムの質はどのように向上させるのか。近年行われたソーシャルメディア分析での調査結果を踏まえると、事象の検出の適時性および達成範囲の向上のためには、今後はおそらく、ニュースによる事象や、ツイッターのようなサイト上での個人的な報告による情報を融合し、取り込んでいく必要があると考えられるであろう。このような方法でのエビデンス結合は今後の研究における大きな課題となってくるであろう。

4, 自己評価

このプロジェクトでは、公衆衛生に関する脅威を、オンラインのメディアソースを用いて検出するシステムの性能向上のために、アルゴリズムとリソースが開発された。私は、本来の目的の重要な部分は達成できたものと考えている。 (a)公衆衛生事象の自動警告に向けての新たな方法の開発への挑戦、(b)コンピュータに公衆衛生を認識させる手助けとなる知識リソースの開発、(c)さまざまなソースからの、エビデンスを融合させる方法の発見 (a)においては、まず最初に、信頼性のある外部基準に対する、自動システム性能を評価する方法を見つける必要があった。研究開

始当初は、基準というものが存在しておらず、ProMED メールの利用によって、自動警告の性能について現実的な見識を得ることができたのである。確立した評価基準によって、自然言語処理および変化点検出アルゴリズムを併用するという、警報に対する新たなアプローチの開拓が可能となった。私自身の、また公衆衛生アナリストらによる独自の研究を踏まえると、私は、この方法によって、分析者は時間と労力を費やすことなく最高水準の結果を得ることができると考えている。この利点は、BioCaster BORN より日本や世界保健機関、米国疾病監理予防センター(CDC)、欧州疾病監理予防センター(ECDC)といった海外の公衆衛生機関のアナリストに送信される警報に見ることができる。私は、(b) において共同研究者とともに、世界初の自由にダウンロードが可能な多言語の公衆衛生オントロジーの開発に成功したのである。このリソース内の用語は公的公衆衛生機関の主導についてアナリストと協議することによって誘導されてきており、今後も改良、拡大が見込まれる。これまでに、世界中で 350 を超えるグループがオントロジーのダウンロードを行っている。最後になるが、(c)では、ニュースレポート内の情報融合を達成するまでには、当初の予測よりはるかに長い時間を要することとなり、また、新たなアルゴリズムの開発だけでなく、知識リソースの産出も必要となった。地理−時間的情報の自動注釈のための新たなスキーマの作成、ならびに報告事象間の関連性を理解するための知識モデルの構築といった本来の目的の一部はすでに達成されている。ひとつの機械学習モデル内にこういったすべてのリソースをまとめることが、今後の研究の当面の主要目標となるであろう。

この研究における最終的な成果は、リアルタイムでの健康関連事象の生物−地理マップや、過去 3 年の新たな事象のデータベースといった、公的に入手可能な BioCaster BORN プラットフォーム上で見ることができる。この研究の開始当初には、データベース化は想定されていなかったのだが、公衆衛生専門家との協議過程で、感染疾患の拡大を調査する上で、その必要性を認識するに至った。プロジェクトを通じてすべての技術ならびにリソースの実現が、世界規模の公衆衛生の保護に利益をもたらすことを願うものである。

5, 研究総括の見解

インターネット上に流れている健康被害に関するニューステキストを分析して健康被害情報に関するアラームを発信するシステムの提案であり、社会的意義が高く、大規模な応用分野をカバーしている課題である。技術的には、イベント系列の解析により情報の重要度を推測し、アラートを出せるようになることを期待していた。

この課題で、公衆衛生に関する脅威を、オンラインのメディアソースを用いて検出するシステムの性能向上のために、アルゴリズムとリソースを開発しており、本来の目的の重要な部分は達成できたものと評価する。また、新たなアルゴリズムの開発だけでなく、地理−時間的情報の自動注釈のための新たなスキーマの作成、報告事象間の関連性を理解するための知識モデルの構築といった本来の目的の一部はすでに達成されていることを評価する。自然言語処理および変化点検出アルゴリズムを併用するという、警報に対する新たなアプローチの開拓が可能となり、開発した多言語の公衆衛生オントロジーは、これまでに、世界中で 350 を超えるグループによりダウンロードされ、利用されている。課題提案時より暫定的に動いていたシステムを中心にした研究開発であったため、完成度が何より重要である。この期待に応える利用実績を示し、この分野に大きく貢献したと考える。

6，主な研究成果リスト

（1）論文（原著論文）発表

1. Collier N. (2011), “Towards cross-lingual alerting for bursty epidemic events”, *BMC Biomedical Semantics*, 2 Supp 5: S10.

2. Collier N *et al.* (2010), “An オントロジー-driven system for detecting global health events”, Proc. COLING 2010, Beijing, China, pp. 215-222.

3. Collier N. (2010), “What’s unusual in disease outbreak news?”, *BMC Biomedical Semantics,* 1(1).

4. Chanlekha H and Collier N. (2010), “A framework for enhanced spatial and temporal granularity in report-based health surveillance systems”, *Medical Informatics and Decision Making,* 2010, 10(1).

5. Collier N, Doan S, Kawazoe A, Goodwin RM, Conway M, Tateno Y, Ngo QH, Dien D, Kawtrakul A, Takeuchi K, Shigematsu S, Taniguchi K. (2008), “BioCaster: detecting public health rumors with a Web-based text mining system”, *Bioinformatics*, 24(24): 2940-2941.

（2）特許出願

（3）その他の成果（主要な学会発表、受賞、著作物等）

[1] Collier N. “GENI-DB: A database of Web-based global event disease reports”, *Bioinformatics* (under review).

[2] Collier, N. “An overview of text mining for epidemic intelligence”, *Global Public Health* (under review).

[3] “Text mining in action: early alerting of disease outbreaks from online media”, Talk given at the American Association for the Advancement of Science Annual Meeting, Vancouver, special track on Web Surveillance: Fighting Terrorism and Infectious Diseases, Canada (2012.2).

[4] “BioCaster: Web sensing for real time disaster detection and tracking”, Invited talk given at the Workshop on the Politics of Disease Surveillance: how unofficial reporting is changing official behaviour, Brisbane, Australia (2011.7).

[5] “Analysis of the grey literature including news events and user generated content”, Invited talk given at the EMBL-EBI Industry Programme Workshop on Literature Services, Cambridge, UK (2011.6).

[6] “Web sensing for real time disaster detection and tracking”, Invited talk given at the University of Manchester, School of Computer Science, UK. (2011.6).

[7] “Web sensing for real time disaster detection and tracking”, Invited talk given at the University of Tokyo Institute of Science and Technology, Department of Computer Science, Japan (2011.6).

[8]Collier N *et al.* (2010), “Navigating the Information Storm: Web-based Global Health Surveillance”, in BioSurveillance: Methods and Case Studies, Kass-Hout, T. and Zhang, X. (eds), Chapman and Hall.

[9] “Text mining in action: Global disease surveillance and alerting from online news”, Invited talk at the University of Zurich, Department of Informatics, Switzerland (2010.10)
[10] “Text mining in action: Global disease surveillance and alerting from online news”, Invited talk at Cambridge University, Computer Laboratory, UK (2010.10).
[11] “Web signals and sensors: an overview of public health alerting in BioCaster”, Invited talk at Oxford University, Department of Zoology, UK (2010.10).
[12] “Online text analysis for early alerting of disease outbreaks”, Invited talk at the National Institute of Public Health, Japan (2010.10).
[13] Doan, S., Conway, M. and Collier, N. (2010), “An Empirical Study of Sections in Classifying Disease Outbreak Reports”, *Annals of Information Systems*, Special Issue ″Web-based Applications in Health Care & Biomedicine″, Springer.
[14] Hartley D, Nelson N, Walters R, Arthur R, Yangarber R, Madoff L, Linge J, Mawudeku A, Collier N, Brownstein J, Thinus G. and Lightfoot N. (2010), “The landscape of international event-based biosurveillance”, J. *Emerging Health Threats Journal*.
[15] “BioCaster: early detection of public health events on the Web”, Invited talk at the Japan Science and Technology Agency, Austria-Japan ICT Workshop, Japan
[16] “Text mining in action: Global disease surveillance and alerting from online news”, Invited talk given at the Centre for Epidemiology and Risk Analysis, Veterinary Laboratories Agency, UK (2010.7).
[17] “Text mining in action: Global disease surveillance and alerting from online news”, Invited talk given at JEITA, Japan Electronics and Technology Industries Association, Japan (2010.6).
[18] Chanlekha, H. and Collier, N. (2010), “Analysis of syntactic and semantic features for fine-grained event-spatial understanding in outbreak news reports”, *Journal of Biomedical Semantics*, 1:3, DOI: 10.1186/2041-1480-1-3.
[19] Chanlekha, H. and Collier, N. (2010), “A methodology to enhance spatial understanding of disease outbreak events reported in news articles”,*International Journal of Medical Informatics*, 79(4): 284-296.
[20] “High throughput analysis and alerting of disease outbreaks from the grey literature”, Invited talk given at the European Bioinformatics Institute, Cambridge, UK (2010.1).

研 究 報 告 書

「時空間解析に基づくインターネットトラフィック異常検出とそのデータベース化」

研究期間: 平成 20 年 10 月～平成 24 年 3 月
研 究 者: 福田 健介

1，研究のねらい

インターネットバックボーン上では，大多数の通常のトラフィックに隠れた異常なトラフィックが存在することが知られており，これらの異常トラフィックをより効率的に発見する手法の確立が求められている．本研究では，上記の目標のために，下記のねらいを定め研究を行った．

(1) 異常トラフィックが持つ特徴量の連続的な変化に着目し，その軌跡を画像処理的アプローチにより検出する異常検出器の実現

(2) 理論的な背景の異なる複数の異常検出器の出力を組み合わせ，その性能の比較を可能とし，また，それらの性能向上を図ることが可能な，ベンチマークアーキテクチャの実現

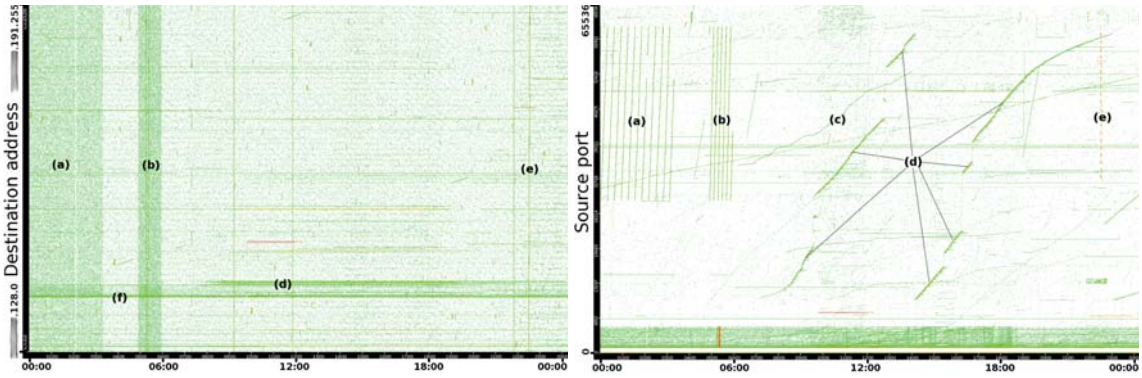
(3) 公開されている 10 年にわたるインターネットトラフィックデータに対して，提案アーキテクチャを適用することで，異常イベントのデータベースを構築し研究コミュニティに公開

2，研究成果

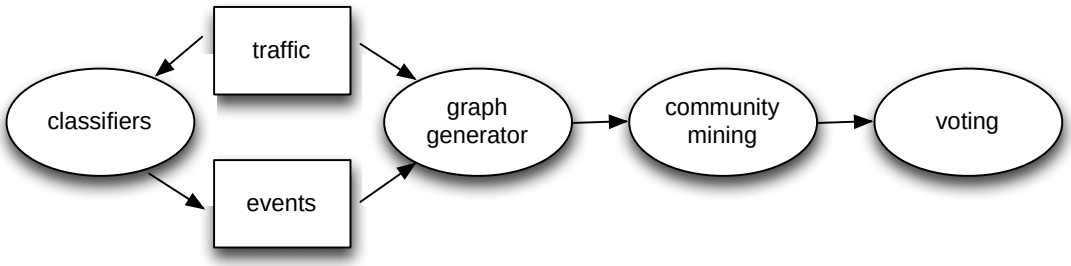
(1) インターネットバックボーントラフィックにおける異常(イベント)は，ウィルス，ワーム，DDos(分散サービス不能攻撃)，機器の故障・設定間違い，フラッシュクラウド(特定コンテンツへのリクエストの集中)等さまざまである．しかしながら，これらの異常イベントは大部分の正常なトラフィックに比べると，トラフィック量としてはそれほどの量とはならない．そのため，トラフィック量を監視しているだけでは，埋もれた異常イベントを検出することが難しい．本研究では，これらの埋もれた異常イベントを構成するパケット群の特徴量の連続する変化を二次元平面上の軌跡と捉え，その軌跡を検出することで異常イベントを検出する，新しいタイプの異常検出器を提案・実装評価した．

下図は，異常トラフィックの時間変化(各点はパケットトラフィックに対応)を示している，左図では特徴量として異常トラフィックの送信元アドレス，右図では送信元ポートを表している．例えば，左図の水平線上の線状の集合は特定ホストへのアクセス，垂直線上の集合は多数のホストへのアクセスに対応している．左図では個々の異常に対応するアクティビティはそれほど明らかではないが，右図のように適切な特徴量を選ぶことで，異常を二次元上の軌跡として捉えることが可能となる．提案アルゴリズムは複数のステップから構成される．(a) トラフィックトレースをランダムハッシュし，複数のサブトレースを生成，(b) サブトレースごとに時間・特徴量(4 種類)空間のデータの切り出し，(c) ハフ変換によるエッジ(軌跡)検出，(d) サブトレースごとに得られた軌跡に属する送信元 IP アドレスをリストアップし，それらのインターセクションを異常イベントに係わる送信元として同定する．提案アルゴリズムを 10 年間にわたるインターネット日米国際リンクトラフィックデータ(MAWIトレース)に適用し，その性能を評価した．提案アルゴリズムは，他のアルゴリズムと比較して，3-5 倍の異常を検出することに成功した．同様に，時間軸と特徴量の変化量に着目し，異常イベントの到達速度を推定する手法を開発した．これ

により，異常イベントの 80–90%は一定の速度をもち，残りの異常イベント(主に特定のオペレーティングシステムによるもの)はランダムな到着であることが明らかとなった．また，提案アルゴリズムを任意の二つの特徴量に拡張し，ユニークなパケットの到着速度に着目することで画像サイズを自動的にチューニングする手法を開発し，オリジナルの手法と比べて，さらに 20%の性能向上が可能となった．α フロー(トラフィック量の多い単一のフロー)に対する検出精度はあまり高くならないものの，他の異常イベントの検出精度が向上したことで，他の検出器では検出困難な異常イベントを検出することが可能となった．



(2) 既存のインターネットトラフィックにおける異常検出器研究に関する問題点として，(a) 共通のトラフィックデータが用いられていない，(b) 異常イベントがラベリングされたデータ(正解データ)が存在しない，(c) 複数検出器の出力の粒度が異なるため単純な比較が困難，(d) これら共通の土俵での検出器の比較検討がなされていないため，どの異常に対してどの異常検出器が優位であるかが不明，などが挙げられる．
以上の問題点を解決するべく，複数異常検出器の出力を比較検討可能とするベンチマークアーキテクチャ(MAWILab)を提案・実装・評価した．提案システムの処理ステップの概略は以下のとおりである(下図参照)



(a) 共通公開トラフィックデータ(MAWI トレース)を入力とした，個々の異常検出器の出力を xml ファイルとして保存．xml ファイルには，異常検出器名，データファイル名，個々の異常イベントの開始，終了時間，トラフィックキー(送信元・送信先アドレスおよびポート，プロトコルの全てもしくはその一部)，イベントのヒューリステックラベル(異常・正常・不明およびその理由)が記録される，個々の異常検出器の出力精度を制御パラメータとするために，個々の検出器ごとに，3 種類の異なるパラメータセット(出力数: 大，最適，小)を用いる．(b) xml ファイル中のイベントからイベント

グラフを生成する．グラフのノードは個々の検出器出力から得られた異常イベント，ノード間のリンクは 2 つのノードに共通するパケットによる重みである．すなわち，複数の異常検出器で共通して得られるイベントはクリークとなり，単一の異常検出器のみから得られるイベントは孤立したノードとなる． イベントに共通するパケットに着目することで，出力粒度の異なる異常検出器を同じように比較検討することが可能となる．（c）得られたグラフから，稠密なサブグラフをコミュニティマイニングアルゴリズムにより検出する．これにより，複数検出器により検出されるイベントを一つのクラスタとして扱うことが可能となる．もし，異常検出器の性能が同じであれば，この結果はクリークとなる．(d) 検出されたクラスタが異常もしくは正常であるかを複数の投票方式(多数決，最大，最小，SVD ベースの教師無し学習（SCANN）)を用いて判定する．SVD を用いる利点は，単一の異常検出器が多くの異常を検出し，その多くが他の検出器で検出されない場合，その結果は最終判定では低い重みがつけられる点である．(e) 得られた異常クラスタに属するパケットデータに，アプリオリアルゴリズムを適用することで，異常クラスタに対応するパケットフィルタルールを生成し，ルータ・スイッチでのフィルタリングを可能とする．
4 種類の理論的なバックグラウンドの異なる異常検出器(画像処理，PCA, Sketch gamman, KL 統計量に基づくもの)を実装し，10 年間わたる公開トラフィックデータ(MAWI トレース)を用いることで，提案ベンチマークシステムの評価を行った．その結果，検出された異常の多くは，少数の検出器によってのみ検出されること，精度の低い検出器の出力は最終出力として採用されず，各々の検出器の寄与が明らかとなった．また，SCANN の結果は必ずしも最適ではないが，平均的に優れた性能を示すことがわかった．すなわち，これらの結果は，単一の精度の高い異常検出器を構築することは困難であり，複数の検出器を組み合わせることによってのみ，精度の向上が可能であることを示すものである．

(3) 上記ベンチマークアーキテクチャを 10 年にわたるトラフィックデータ(MAWI トレース)に適用することで，データ中の異常イベントを抽出・ラベリングし，異常トラフィックデータのデータベースを構築，研究者への公開を 2010 年末より開始した(http://www.fukuda-lab.org/mawilab)．2012 年 2 月現在 15 カ国 2000 を越えるアクセスがあり，本データベースのデータをベンチマークデータとして使用した新たな異常検出器の提案が複数行われている．

上記成果の他に，領域会議中に他研究者より，異常パターンをトラフィックデータから抜いた”準正常データ”を学習データとして用いた異常検出器の改良についての助言があり，追加課題として取り組んだ．しかしながら，異常パターンを単純に抜いたデータでは，トラフィックに不必要なギャップが増えること，不明トラフィックの扱い方による差，輻輳時には正常パターンも異常パターンとなること等，いくつかの問題に直面し，期待した成果を得ることはできなかった．しかしながら，今後，既存異常パターンを抜き出すのではなく，アドレス単位のランダムハッシュを用いて，異常トラフィックを少数のハッシュに分離する方向でさらに研究を進めていく予定である．同様に多次元ハフ変換に基づく異常検出器，非線形次元圧縮に基づく異常検出器に関しても研究を進めたが，主として計算量の問題で，期待した成果を得ることができなかった．

3，今後の展開

本研究では，インターネットトラフィックにおける異常検出を予め収集されたパッシブデータを用いて行った．しかしながら，本研究で行った研究結果を生かすには，リアルタイムにデータを収集し，異常を検出することが望ましい．現在，学術情報ネットワーク(SINET)のバックボーン回線においてデータ収集環境を整えつつあり，今後，開発したシステムを本ネットワークに適用し，リアルタイムでの実証実験を行うことで，開発システムの実ネットワークへの適用を目指す．同様に，機械学習的アプローチを用いた複数異常検出器の出力結果の組み合わせによる精度向上，理論的バックグラウンドの異なる異常検出器の追加，ヒューリスティックラベルの精度向上を行うことでシステムとしての精度向上，他研究者が開発した異常検出器の性能を MAWILab と比較可能な Web インターフェイスの構築，等を行う予定である．

4，自己評価

当初の提案では，画像処理アプローチに基づく異常検出器をメインターゲットとして研究を進める予定であり，実際，精度の高い，自動パラメータ設定可能な検出器を提案できたことから，当初の目的を達成できたと考える．さらに，異常検出器の研究を進めるにつれ，現在のインターネットトラフィック異常検出におけるさまざまな問題点(共通したデータセットの欠如，正解データの欠如，異常検出器間の性能比較方法の欠如)に直面し，これらの問題を解決するよう，研究トピックを追加した．その結果，提案時にはなかった，複数異常検出器の比較ベンチマークアーキテクチャおよび共通データベースの精度向上に関しての研究が進み，トップ国際会議への採択に至った．また，複数検出器出力に基づいた，インターネットトラフィック異常データベースを公開したことで，他研究者からベンチマークに使用される標準データベースとなりつつある点は，当初の目的を達成できたと考える．反面，画像処理ベースの異常検出器の実ネットワーク(リアルタイム)への適用が遅れ，今後の課題となった．

5，研究総括の見解

インターネットトラフィックの時系列を時空間パターンにして解析し、大量のデータに埋もれた少量の検出対象を見つけるという課題である。具体性があり、研究を評価するためのデータの準備も整っており、インターネットに限らず、汎用性のある画像認識によるネットワークダイナミックの異常を検知する技術、また、トラフィック解析のみにとどまらず、幅広い場面での適用が可能な技術となることを期待していた。

当初の提案では、画像処理アプローチに基づく異常検出器をメインターゲットとして研究を進める予定であり、実際、精度の高い、自動パラメータ設定可能な検出器を提案できたことから、当初の目的を達成できていると評価する。さらに、現在のインターネットトラフィック異常検出におけるさまざまな問題点(共通したデータセットの欠如、正解データの欠如、異常検出器間の性能比較方法の欠如)を解決するよう、研究トピックを追加した。その結果、提案時にはなかった、複数異常検出器の比較ベンチマークアーキテクチャおよび共通データベースの精度向上に関しての研究が進み、トップ国際会議への採択に至っている。また、複数検出器出力に基づいた、インターネットトラフィック異常データベースを公開したことで、他研究者からベンチマークに使用される標準データベースとなりつつある。この分野に大きく貢献したと考える。

一方、異常検出器の実ネットワーク(リアルタイム)への適用が遅れており、今後の研究に期待

する。

6，主な研究成果リスト

（1）論文（原著論文）発表

1. R.Fontugne, Y.Himura, K.Fukuda, Evaluation of anomaly detection method based on pattern recognition, IEICE Transactions on Communications, vol.E93-B, no.2, pp.328-335, IEICE, 2010

2. Y.Himura, K.Fukuda, K.Cho, H.Esaki, An evaluation of automatic parameter tuning of a statics-based anomaly detection, International Journal of Network Management, vo.20, no.5, pp.295-316, Wiley, 2010

3. Y.Himura, K.Fukuda, K.Cho, H.Esaki, Characterization of host-based traffic with multi-scale gamma model, IEICE Transactions on Communications, vol.E93-B, no.11, pp.3048-3057, IEICE, 2010

4. R.Fontugne, T.Hirotsu, K.Fukuda, A Visualization tool for exploring multi-scale traffic anomalies, Journal of Networks, vol.4, no.4, pp.577-586, Academy publisher, 2011

5. R.Fontugne, K.Fukuda, Hough-transform-based anomaly detector with an adaptive time interval, ACM Applied Computing Review, vol.11, no.3, pp.41-51, ACM, 2011.

（2）特許出願

研究期間累積件数：0 件

（3）その他の成果（主要な学会発表、受賞、著作物等）

1. R.Fontugne, P.Borgnat, P.Abry, K.Fukuda, Uncovering relations between traffic classifiers and anomaly detectors via graph theory, Proceedings of TMA2010, pp.101-114, Zurich, Apr, 2010

2. K.Fukuda, R.Fontugne, Estimating speed of scanning activities with a Hough transform, Proceedings IEEE ICC2010, p.5, Capetown, Jun., 2010

3. Y.Kanda, K.Fukuda, T.Sugawara, An evaluation of anomaly detection based on sketch and PCA, Proceedings of IEEE GLOBECOM2010, p.5, Miami, Dec., 2010

4. R.Fontugune, P.Borgnat, P.Abry, K.Fukuda, MAWILab: Combining diverse anomaly detectors for automated anomaly labeling and performance benchmarking, Proceedings of ACM CoNEXT2010, p.12, Philadelphia, Dec., 2010

5. K.Fukuda, An analysis of longitudinal TCP passive measurements, Proceedings of TMA2011, pp.29-36, Vienna, Apr., 2011

研　究　報　告　書

「マルチソースデータ高度利用のための統計的データ融合」

研究期間:　平成 20 年10月～平成 24 年3月
研　究　者:　星野　崇宏

1．研究のねらい

製品開発やマーケティング戦略のために得られる様々なデータは複数の情報源から得られるマルチソースデータであることがほとんどである（図1）。マルチソースデータからは各データ間をまたがる変数間の関連を見ることができず、これは実務上では非常に大きな問題となる。

本研究では、各データセットで測定対象となっている個々人の背景情報を積極的に測定し、その情報を利用することでマルチソースデータからシングルソースデータをシミュレートし補完する統計的なデータ融合手法を開発することを目的とする。具体的には、近年統計学において発展の目覚ましいセミパラメトリックな手法を用いたデータ融合手法を開発する。また、これまで注目されなかった問題として、各データセット間での対象者の異質性が重要である。例えば購買データ上には購入頻度の高い顧客が多いのが通常である。従って各データセットで得られる対象者の違いによるバイアスを除去する必要がある。加えて、先行研究では重要視されていなかった「背景情報＝共変量情報」の積極的な探索と利用を行うことで、これまで提案されてきた欠測を補完する方法の様々な問題点を克服する。さらに、単に数理的な手法の開発にとどまらず、実際の製品開発とマーケティング戦略に関する実験調査への応用可能性について実データを積極的に用いて検証する。具体的には、本研究に協力頂ける企業が有する大規模な調査パネルに対する市場調査を実施し、企業が取得している Web 閲覧データと融合させる、あるいは POS 等の実績データなどと融合させることでマーケティング分野での実際の予測を行い、一定の汎用性が得られるかを検討する。

データ融合が対象とするマルチソースデータと同様のデータ構造、解析ニーズは社会科学一般において存在する。そこで、経済学・社会学・教育学分野においてデータ融合の応用研究を行う。例えば repeated cross-sectional dataset から panel data によってのみ得られる情報を抽出する、疑似パネルデータ解析はデータ融合が対象とするデータ構造と基本的に同じものである。そこで、既存の疑似パネルデータ解析では可能ではなかった、各データセットに対する対象者の割り当てが無作為でない場合にも利用可能な方法を、データ融合の方法論を応用し開発する。

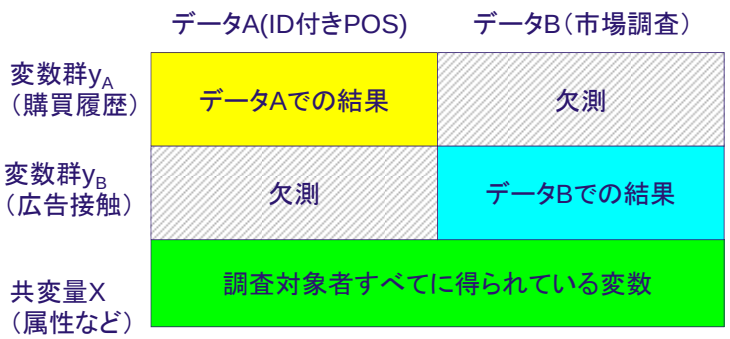


図1：もっともシンプルなマルチソースデータセットの形式

2．研究成果

得られた研究成果は大別すると以下の5つである。

（1）データ融合についての先行研究での前提条件の解明とその緩和

これまでデータ融合の具体的方法として利用されてきたのは「マッチング」「潜在変数モデル」「回帰的モデル」である。まず先行研究で提案されている方法を調べ、これらがすべて以下の仮定を置いていることを確認した。

【1】ランダムな欠測(Missing at random)

どのデータで観測されるかが、共変量に依存しており、アウトカムには依存しない。

【2】条件付き独立(conditional independence)

共変量を所与としたアウトカムは各データセットで独立である。

しかし、実は上記の条件は厳しすぎるため、現実のデータセットに適用できるかどうかは疑問である。そこで、データ構造を欠測データセットの考え方のもとに整理した結果、それぞれ特定のパラメトリックモデル、たとえば各データへの所属インディケータの背後に連続量を仮定するプロビットモデルを考え、また各アウトカムが正規分布に従うと仮定した場合には

【1‘】ランダムでない欠測

どのデータで観測されるかは、基準となるデータセットでのアウトカムには依存してもよい。

【2‘】連関の許容

各データセットでのアウトカムが連関することを許容する。

という形でそれぞれの条件を大幅に緩和してよいことが分かった。

さて、アウトカムのモデルには特定の分布仮定を置くことは消費者行動理論や計量経済学などマーケティングサイエンスの基礎となる諸研究から正当化することはできる。一方、どのデータでその対象者が観測されるか（あるいは欠測となるか）については特定の根拠をもとにパラメトリックなモデルを仮定することは難しい。そこで本研究では対象者のデータセット割り当て（欠測）モデルについてはパラメトリックな仮定を置かず、アウトカムの周辺同時分布についてはパラメトリックな仮定を置くセミパラメトリックモデルを利用することとした。このようなモデルで最も効率的であり、かつ欠測データの精度の高い予測も可能にするモデルとして、ディリクレ過程混合分布を用いたセミパラメトリックベイズ推定法を開発した。シミュレーションの結果からは既存の手法よりも大幅に推定の誤差を減少させることがわかった。また製品の購入とその製品についてのインターネットサイトの閲覧の関係を調べる広告効果測定の実データにおいても、既存法では連関指標の推定において明らかな過小推定が生じてしまうが、提案手法は真値を復元することがわかった（図2）。

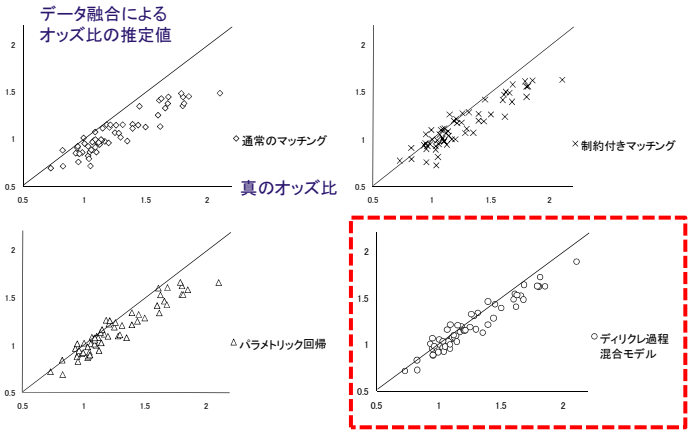


図2：既存法と提案手法（赤枠）の比較

（2）感度分析法の開発

上記の方法論によりデータ融合の仮定は大幅に緩和されたが、実務では各データセットでのアウトカムが高い関連を有する場合も多く、基準となるデータセットがどれであるか不明な場合もある。このような場合には観測されている共変量だけではなく、観測されない未知の共変量によって複数のアウトカムが説明されると考えるのが自然である。但し、潜在変数を利用したモデルでは識別性が無くなってしまう（データのみから母数が決定されない）。一方、潜在変数が観測確率に与える影響に関する母数のみ値を固定すれば、それ以外の母数については推定を行うことができるため、その母数の値を一定の範囲で変化させながら推定を行うことで、アウトカム間の連関がどの範囲に変動するか（そしてその信頼区間はどの程度か）を調べる感度分析法を開発した。この方法を用いることで、欠測を例えば最大 50%説明する未知の共変量が存在したとしても、どの程度の範囲に 95%信頼区間が収まるかを推定することが可能になる（図3）。

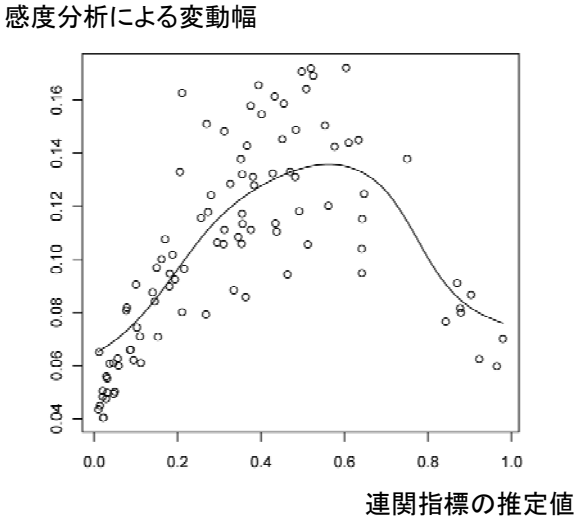


図3：感度分析法による 95%信頼区間の変動幅の推定

（3）マーケティング実務におけるニーズを満たす応用：属性情報の付加

マーケティングの実務において必要とされるデータ融合の応用を複数実施した。一例として、フリークエントショッパーズプログラム(FSP)データにおける属性情報の付加に関する応用を行った。FSP データでは同一の消費者に対して複数時点での購買履歴がひも付けされるという点で非常に豊かなデータであるが、属性情報は一般に性別や年齢、住所など会員登録時に得られる最低限の情報にとどまる。一方、企業がマーケティング戦略を策定する際には、収入や家族構成、職種、居住形態やライフスタイルなどの情報を利用し適切なセグメント設定を行うことや、細かな属性別の市場規模の予測を行うことが適切な広告戦略や価格戦略につながるが、このデータではそれを可能にするに十分な属性情報がない。従って実務上では FSP データに対する属性情報の付加には大きなニーズがある。そこで ID 付きPOS データのうち、市場調査も行うパネル型調査データから、「FSP データ」と、「細かな属性情報と比較的粗い購買履歴情報を測定する市場調査データ」の2つのデータセットにあえ

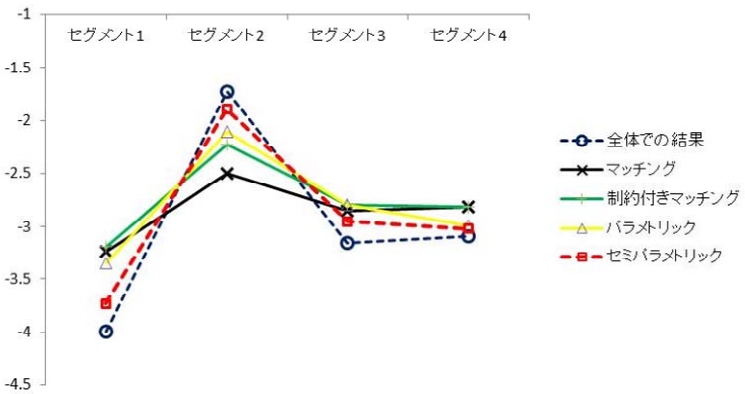


図4：データ融合を用いた価格弾力性の復元

て分離し、その2つのデータを融合させ、FSP データ上の対象者に対する意味のあるセグメンテーションを行い、セグメント別のロイヤルティ係数や価格弾力性の推定を行った。結果としては提案手法はパネル調査データから得られる真のデータ構造を正しく復元することができた（図4）。

（4）疑似パネルデータセットでの因果効果推定

経済学や社会学、心理学などでは同一対象者を追跡調査するパネル調査研究を行うことで、個人差を除去した「特定の独立変数の従属変数への因果効果」を推定することが多い。実際には標本の代表性を担保しながら、ドロップアウトのないパネル調査を実施することは難しいため、パネル調査ではなく、同じ標本抽出法を利用して抽出された、対象者の異なる２時点の調査データ（repeated cross-sectional data）から解析を行わざるを得ないことがある。このような研究関心を一般的に疑似パネルデータ解析と呼ぶが、経済学等で利用されてきた手法はマッチング及び線形の回帰モデルを用いるものであり、（1）で述べたような観点からはデータ融合の考え方を用いることでより頑健かつ効率的な推定法を得ることが可能であると考えられる。

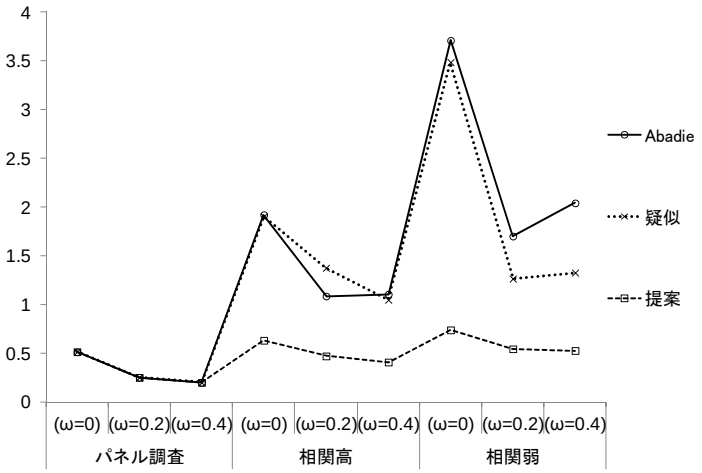


図5：様々な設定での二乗誤差の手法間比較

そこで、本研究で提案したセミパラメトリックなデータ融合法を用いることで、対象者の異なる repeated cross-sectional data を用いて因果効果を推定する方法を開発したところ、既存の手法（一時点のデータによる Abadie 法や既存の疑似パネルの解析法)よりも二乗誤差が小さい推定法を得ることが出来た（図5）。

（5）マルチレベルデータでのデータ融合

社会科学や疫学においては学校-教室-生徒、地域-病院-患者、のような階層データが得られることが普通である。ここで、学校への教育費の投入や教員の質、病院の設備など「上位のユニット(level-2 unit)の効果」と生徒の家庭環境や患者の健康状態など「下位のユニット(level-1 unit)の効果」、そしてその相互作用がどのように結果に影響を与えうるかを調べるためにマルチレベル分析が用いられるが、例えば「地方政府からの教育費投入の程度によって能力別学級の効果はどのように異なるか」といった効果の推定は教育政策や医療政策において非常に重要である。しかしどの「上位ユニットへの下位ユニットの割り当て」はランダムに行えないので、通常は上記の効果と「下位ユニットの性質」が交絡して正しい推定ができない。

そこでデータ融合の考え方を用いたマルチレベルにおける新しい因果効果の推定法を開発した。具体的には、既存の方法では上位ユニットごとへの割り当てをモデリングする必要があるが、上位ユニットは通常は極めて多く、実用的ではない。そこで、上位ユニットの特徴を説明する変数についての割り当てをモデリングするセミパラメトリックな inverse probability weighting

estimator を開発し、既存の方法よりも頑健かつ安定した推定値を与えることを確認した。

3，今後の展開

本研究では当初の予想を超えて数理的な手法論開発に重点的に取り組む必要が生じたため、マーケティングにおける応用研究としてはインターネットのサイト閲覧についてのパネル調査データ、購買履歴に関する調査データ、消費者に対する市場調査データのデータ融合のみとなった。今後はスマートフォンなどのモバイル端末からのセンサーデータや、消費者生成メディアなど様々なデータとのデータ融合についての応用を行う予定である。

また社会科学等他分野への応用という観点からは、教育学や疫学等の調査データ、心理学や行動経済学での実験データ等に対しての応用が可能であるため、これに取り組みたい。

共変量選択法の開発については、クロスバリデーションなど開発された手法から自然に導出される方法を用いることが可能であることは分かった。しかし、本来は「データ取得や調査設計段階でどのような共変量を測定すれば良いかについて一般的な示唆を与える」ことが実務的な観点からは要請されている。現状の共変量選択はあくまで事後的な選択法でしかなく、この点については実務的な要請に答える段階にはない。従って、今後より多くの応用例を積み重ねることで、どのように収集された（どのように偏った）データセットのどのようなアウトカム間の関連を調べる際には、どのような共変量を事前に測定しておくことが重要かについての経験則を得ることが、より広範囲の実務家に利用されるためには必要であると考えられる。

4，自己評価

数理的な方法論の開発と言う観点では、「既存のパラメトリックな手法をセミパラメトリックモデルで代替する」という当初目標にとどまらず、「既存の手法の仮定を大幅に緩和することが可能である」ことを発見できたため、非常に大きな進展があったと評価できると考える。また、疑似パネルデータ解析やマルチレベル分析において、今回開発した方法を応用することで既存の方法論よりも頑健かつ効率の良い因果効果推定の方法を開発することができたという点で社会科学への応用研究と言う点で当初目標を達成したと考えられる。

一方、実務に応用できるパッケージ化という観点では、4の今後の展開で記載したように事前の共変量選択に対する示唆を得ることが出来なかったという点が残念であるが、これも4で記載したように、様々な応用例の積み重ねを行うことで今後一般則を見出して行きたい。

5，研究総括の見解

マルチソースデータからシングルソースデータをシミュレートし補完する手法であり、断片的な大量データから各種の統計的推定を安定に行うために不可欠な技術である。必ず開発していかなければならない技術であり、コア的研究といえる。研究成果を具体的な調査活動につなげることを意識して研究を進めることを期待していた。

数理的な方法論の開発と言う観点では、「既存のパラメトリックな手法をセミパラメトリックモデルで代替する」という当初目標にとどまらず、「既存の手法の仮定を大幅に緩和することが可能である」ことを発見できたため、非常に大きな進展があったと評価する。また、疑似パネルデータ解析やマルチレベル分析において、今回開発した方法を応用することで既存の方法論よりも頑健かつ効率の良い因果効果推定の方法を開発することができたという点で社会科学への応用

研究と言う点で当初目標を達成していると評価する。この分野に大きく貢献していると考える。

実務に応用できるパッケージ化という観点では、事前の共変量選択に対する示唆を得ることが出来なかったという点が残念であるが、今後、様々な応用例の積み重ねを行うことで一般則を見出していくことを期待する。

6, 主な研究成果リスト

(1)論文(原著論文)発表

1. Kei Miyazaki and Takahiro Hoshino, "A Bayesian Semiparametric Item Response Model with Dirichlet Process Priors", Psychometrika, 74　375-393. 2009.

2. Kei Miyazaki, Takahiro Hoshino, Shin-ichi Mayekawa and Kazuo Shigemasu. "A New Concurrent Calibration Method for Nonequivalent Group Design under Nonrandom Assignment", Psychometrika, 74　1-19. 2009.

3. 星野崇宏 "調査不能がある場合の標本調査におけるセミパラメトリック推定と感度分析:日本人の国民性調査データへの適用" 統計数理, 第 58 巻 1 号　3-23.

4. 猪狩良介・星野崇宏 (印刷中) "非集計 Web アクセスデータを用いたサイト普及モデル:多時点・複数サイトの階層ベイズモデリング" マーケティング・サイエンス, 2012.

5. Tetsuro Kobayashi and Takahiro Hoshino"Propensity Score Adjustment for Internet Panel Surveys of Voting Behavior: A Case in Japan",Japanese Journal of Electoral Studies. 27, 104-117, 2011

6.Yusuke Takahashi, Robert W. Brent, and Takahiro Hoshino (in press) "Conscientiousness mediates the relation between perceived parental socialization and self-rated health"

7. Shiro Ojima, Naoko Nakamura, Hiroko Matsuba-Kurita, Takahiro Hoshino and Hiroko Hagiwara (2011)"Age and amount of exposure to a foreign language during childhood: Behavioral and ERP data on the semantic comprehension of spoken English by Japanese children",Neuroscience Research. 2011 Jun;70(2):197-205.

8. Takehiro Nagai, Takahiro Hoshino and Keiji Uchikawa (2011) "Statistical Significance Testing with Mahalanobis Distance for Thresholds Estimated from Constant Stimuli Method" Seeing and Perceiving, 24, 91-124.

9. Shiro Ojima, Naoko Nakamura, Hiroko Matsuba-Kurita, Takahiro Hoshino and Hiroko Hagiwara (2011). "Neural correlates of foreign-language learning in childhood: A 3-year longitudinal ERP study", Journal of Cognitive Neuroscience. 23, 183-199.

10. Atsunobu Suzuki, Takahiro Hoshino, and Kazuo Shigemasu (2010) "Happiness is unique: A latent structure of emotion recognition traits revealed by statistical model comparison" Personality and Individual Differences, 48. 196-201.

(2)特許出願

なし

(3)その他の成果(主要な学会発表、受賞、著作物等)

【著作】

・星野崇宏「調査観察データの統計科学：因果推論／選択バイアス／データ融合」岩波書店

【学会発表】

・Takahiro Hoshino, "Causal Inference Framework for Latent Variable Modeling" Invited Talk, The 76th Annual and the 17th International Meetings of the Psychometric Society, Hong-Kong Institute of Education, Hong-Kong, 2011.

・星野崇宏「反実仮想モデルを用いた統計的因果推論について」、第 13 回情報論的学習理論ワークショップ(IBIS2010),東京大学,2010

・猪狩良介、星野崇宏「非集計 Web アクセスデータを用いたサイト普及モデル－多時点複数サイトの階層ベイズモデリグ」,第 86 回日本マーケティングサイエンス学会研究大会、電通ホール 2009

・太田悠太・星野崇宏「広告効果に影響を及ぼす CM 内容：クリエイティブの科学的管理に向けて」,日本消費者行動研究学会第 41 回消費者行動研究コンファレンス、関西学院大学,2010

・直井映里砂・星野崇宏「ネガティブ情報を積極的に提示すべきか？：企業イメージや購買意図への影響」日本消費者行動研究学会第 41 回消費者行動研究コンファレンス、関西学院大学,2010

・猪狩良介、星野崇宏「サイト属性と訪問経路情報を用いた Web 閲覧行動モデル」第 88 回日本マーケティングサイエンス学会研究大会、電通ホール, 2010

・太田悠太、星野崇宏「プロスペクト理論を考慮した同時購買行動での価格プロモーション戦略」第 90 回日本マーケティングサイエンス学会研究大会、株式会社電通 電通ホール,2011

・Miyazaki K., Hoshino T, & Shigemasu K. A Bayesian Equating Method for Nonequivalent Group Design under Nonrandom Assignment via Data Augmentation Algorithm. The 75th Annual Meeting of the Psychometric Society. (The University of Georgia. Athens, Georgia, USA., 2010

・Miyazaki, K., Hoshino, T, & Shigemasu, K. Determining the direction of the path using a Bayesian semiparametric model. 19th International Conference on Computational Statistics. (Conservatoire National des Arts et Métiers, Paris, France, 2010

・宮崎慧. 星野崇宏. 複数商品購買行動のための階層ベイズグレンジャー因果性分析 日本マーケティングサイエンス学会第 88 回研究大会, 電通ホール, 2010

・宮崎慧. 星野崇宏. 動的階層ベイズモデルを用いた複数商品カテゴリー購買とブランド購買の同時分析 2011 年度統計関連学会連合大会, 九州大学, 2011

・宮崎慧. 階層ベイズ動的ポアソンモデルによる複数商品購買行動の分析 行動計量学会第 39 回大会（岡山理科大学, 2011 年 9 月）・星野崇宏「観察研究・調査データからの因果効果の推定」ICPSR 国内利用協議会統計セミナー、 関西学院大学, 2009

・星野崇宏「Web 調査の偏りの補正: 行動経済学における調査研究への適用」, 関西大学ソシオネットワーク戦略研究機構 第 5 回 RISS 経済政策特別講義、 関西大学, 2010

・星野崇宏「標本調査法への統一的なアプローチと新展開」, 2010 年度統計関連学会連合大会、早稲田大学, 2010

・星野崇宏「課題研究：教育調査の在り方を問い直す－量的研究の課題と展望－：統計学の観点から見た量的研究の課題と今後」, 日本教育社会学会 第 62 回大会、 関西大学,2010

・星野崇宏「欠測データと因果効果の推定」、ICPSR 国内利用協議会統計セミナー、

立教大学, 2010

【学会賞】

・星野崇宏　日本行動計量学会　出版賞

研 究 報 告 書

「ネットワーク理論と機械学習を用いたウェブ情報の構造化・知識化」

研究期間: 平成 20 年 10 月～平成 24 年 3 月
研 究 者: 松尾 豊

1，研究のねらい

本研究では，ネットワーク理論と機械学習に基づく，画期的なウェブ情報の統合・知識化アルゴリズムの構築を目指す。特に，エンティティ(人物や組織，物質，製品名等)のネットワークに着目し，目的に応じた予測のための構造化技術を構築する。大量のウェブ情報に書かれたエンティティ間の構造を抽出し，目的に応じて知識として利用するための基盤であり，ウェブの次世代「知識エンジン」につながる技術である。

2，研究成果

研究成果としては、1)予測に基づく知能の実現の仕組みに関する思考、2) 具体的なアルゴリズムとしての実装、3) ウェブ情報をマイニングすることによる社会予測の研究、の大きく3つに分けられる。以下ではそれらを順に説明する。

研究の動機は、「考える脳、考えるコンピュータ」で語られているような予測に基づく知能のアルゴリズムを構築したいということである。現状の機械学習における各手法では、訓練データとして十分に工夫した素性に基づくデータを与えるとうまく学習でき、そうでなければうまく学習できない。機械学習のパフォーマンスを決める最も重要な部分は、人間がよって行われる適切な素性の構築であり、アルゴリズムでうまく行うことができない。一方、人間は何らかの方法で素性の構築を行っており、このことと人間が世界を構造化して捉えていること（エンティティとその関係性という認識をしていること）、ならびにその構造化が次に何が起こるかをより的確に予測するために用いられているということが相互に密接に関連しているのではないか。こうした研究の動機から出発し、領域会議等でのさまざまな議論や文献等の調査、それに基づく思考を経て、この問題に対する重要な要素として以下のものを抽出した。

1. ニューラル・ダーウィニズム：予測の精度が高いニューロン群だけが生き残る。
2. 補助問題の生成：訓練データそれ自身の予測を行うことで、便宜的に問題数を増やし、訓練データの数を増やす。その後、素性の転移等によって、より効率的な学習を行う。
3. ネットワークと中心性：素性間のネットワークを構築した際に、他の素性と関連の強い（中心性の高い）素性は、さまざまな問題に有用なロバストな素性である。また、中心性が高いと、多くの他の素性にアンカーされることになり、動きにくい
4. 言語化：こうしたネットワークにおける部分集合（サブグラフ）が、それを指すポインタとしての何らかの言語表現と紐付けられ、任意のタイミングで活性化することが可能となる。

次に、これをアルゴリズムとして構築する方法についての成果を述べる。ここで構築するアルゴリズムは、自然言語の文書を入力することで、予測をベースにして自動的にシラブルや単語がチャンク化されるアルゴリズムである。入力したデータの一部（テキストの場合、文字や単語など）に由来する「センサーノード」と、その発火を予測する「予測ノードの」の2種類を使うことで、

入力したデータのチャンク化、抽象化等を行うものである。言語によるラベルは、この2部グラフのクラスタとの関連で作られることになる。

以下のような一連の手続きでネットワークが構築され、それが予測に使われる。

1. ひとつの文字に対応するセンサーノードを設定する。各センサーノードに対する予測ノードを作る。
2. 共起が高い2つのセンサーノードに対して、センサーノードと予測ノード間にエッジを張る。
3. センサーノードの発火を学習させる。予測ノードへのほかのセンサーノードからのエッジの重みを修正する。
4. エッジの重みが重いものだけ残す。
5. 予測ノードと、予測ノードに対応したノードが同時に on になったら(予測に成功したら)発火するセンサーノードを作る。2 に戻る。

1 は準備フェーズであり、2 のフェーズは、あり得る素性の解空間を効率的に探索するためのヒューリスティックである。3 は、通常のニューラルネットワークや SVM 等での重み(判別関数)の学習である。4 は、枝刈りのフェーズである。5 が一番特徴的な点であり、予測に成功したことを表すノードを新たに作ることで、重要な文脈は保存されていくことになる。そしてこの文脈自体がセンサーノードとなって、新たにそれを予測するような一連のノード群が構成されていく。

これを具体的に可視化したものが図1である。大規模な文書データを入力することで、次の文字を予測するために、シラブル、単語に該当するノードが現れ、相互に連結されている様子が分かる。

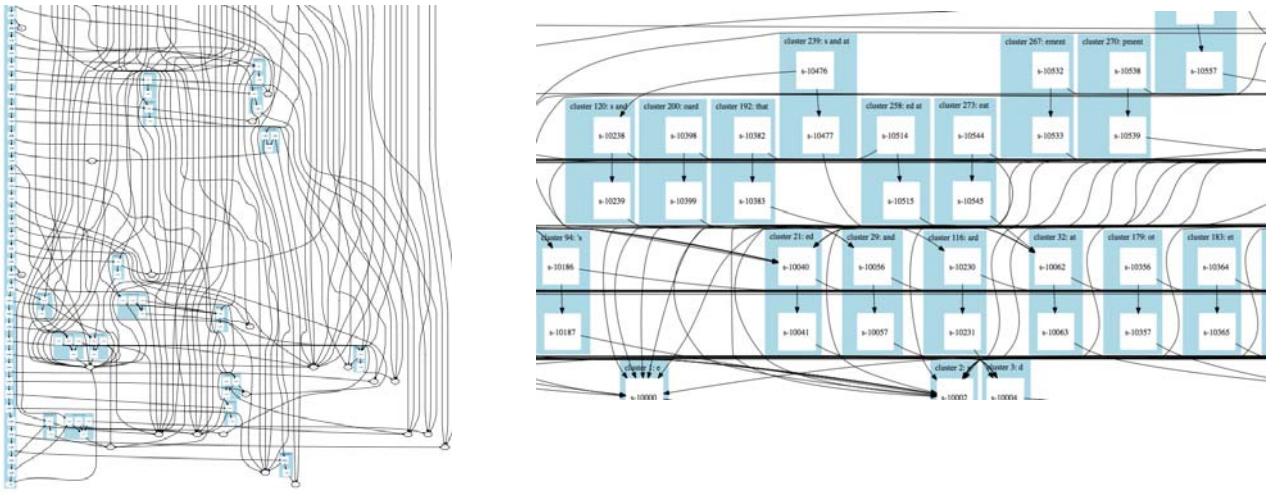


図 1 文書から得られたシラブル、単語のネットワーク構造

これが下図のように階層をもったネットワークとして構造化されていくことが期待できる。また、特定の2部グラフの構造を他のノードが指すことにより、概念を指すポインタとして機能し、それによってより高次の抽象化や関係概念の発見が可能になる。

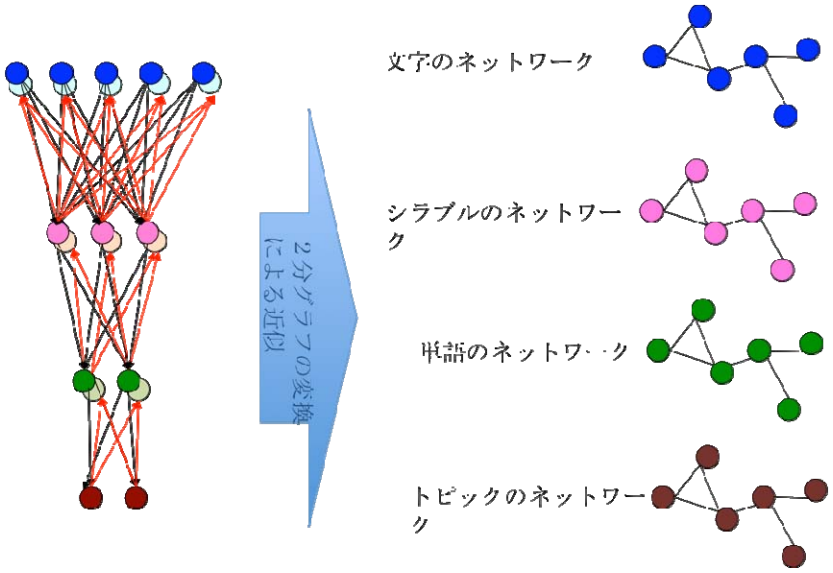


図 2 階層間の2部グラフと各階層におけるネットワーク

本研究では、提案するアルゴリズムをウェブ上のデータに適用するための試みとして、ウェブ上の情報に対して機械学習を用いて予測を行う応用例をいくつか示した。ひとつは、ブログ上の情報からの選挙予測、もうひとつは、twitter からの地震の早期検出の研究である。

下記は国内のブログデータから、国政選挙に関するブログを抜き出し、過去の国政選挙の結果とそのときのブログの特徴（候補者名および党名に対する言及の程度と内容）から選挙の得票数を予測しようという試みである。この結果、ブログ上での言及と得票数は高い相関にあり、精度の高い予測が可能であることが明らかになった。

	# predictions	# correct	# wrong	# unsure	Hit ratio
Blog analysis (ours)	254	241	39	36	92.33%
Asahi Shimbun	260	245	15	40	94.62%
Nikkei Shimbun	288	264	24	12	91.67%

図 3 ブログからの選挙結果の予測の精度

また、下記は、twitter におけるツイートから、地震に関するツイートだけを抽出したものである。ツイート中に含まれるキーワードやその出現位置等を素性として用いることで、地震に関するツイートを一定の精度で抽出することができ、早期に検知することができる。この成果は WWW2009 において論文が採択され、その後、同様の研究における先進的な取り組みとして国内外にインパクトを与えることができた。

図 4 地震に関するツイートの抽出

以上、本研究では、予測に基づく知能の実現の仕組みに関する思考と、それに基づくアルゴリズムの構築、さらにウェブ情報からの社会予測の応用例を示した。相互の課題が直接的な基礎−応用の関係になっているわけではないが、本研究で目的としていたネットワーク構造を用いた画期的な学習、およびその先に広がる世界の像を少なからず示すことが出来たのではないかと考えている。

3, 今後の展開

本研究の今後の展開としては2つの方向がある。ひとつは、提案したアルゴリズムの評価である。アルゴリズムは、データから構造を抽出しながら予測精度の向上を目指すもので、これまでの機械学習のアルゴリズムと比べ、パラメータが多い。この部分の最適化とともにアルゴリズムの評価をしていく必要がある。

もうひとつは、さまざまな社会予測（選挙やイベント）という観点での応用であり、その際には、予測したい対象とは（一見）関係がないものの予測に寄与する素性を抽出することで、本来予測したい対象の予測に寄与するという枠組みが有用であることが想定される。本アルゴリズムを活用する形でこういった社会予測への応用を進めて行きたいと考えている。

4, 自己評価

提案する手法に関して議論を深め手法の大枠を示すことができた点、またウェブマイニングに関してインパクトのある応用例を実現できた点では当初の目標を達成したと考える。一方で、アルゴリズムとしての精錬や評価の点ではまだ課題が多く、当初目標通りではない。設定したテーマが難しい課題であり、3年間でどこまでの成果が上がるかは想定しにくいところではあったが、アドバイザーの先生方等の意見は非常に参考になり、研究助成を受けなかったら進められなかったであろう部分まで大きく進めることができた。大きなチャレンジの土台にあたる部

分であり、今後この期間の研究が花開くように努力していきたい。

5，研究総括の見解

ネットワーク理論と機械学習を用いたウェブ情報の構造化・知識化についての研究である。ウェブマイニングにおける「エンティティ」発見をネットワーク構造上の機械学習の観点から捉えていて独創的であり、ウェブマイニングの核となるアルゴリズムを発見・構築したいという意気込みに期待していた。

提案する手法に関して議論を深め、手法の大枠を示すことができ、また、ウェブマイニングに関して社会的にもインパクトのある応用例を実現し、公表している。これらは彼ならではの優れた成果と考える。また、当初の目標を達成しており、この点も高く評価したい。一方で、構想自体は大きなものであるため新しい今後の研究課題も多く出ている。アルゴリズムとしての精錬や評価の点では、今後、更なる成果につながっていくことを期待する。

6，主な研究成果リスト

（1）論文（原著論文）発表

1. Yutaka Matsuo and Hikaru Yamamoto: Community Gravity: Measuring Bidirectional Effects by Trust and Rating on Online Social Networks, Proc. 18th International World Wide Web Conferenc (WWW2009), 2009
2. Danushka Bollegala, Yutaka Matsuo and Mitsuru Ishizuka: Relational Duality: Unsupervised Extraction of Semantic Relations between Entities on the Web, Proc. 18th International World Wide Web Conference (WWW2010), 2010
3. Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo: Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors, Proc. 18th International World Wide Web Conference (WWW2010), 2010
4. 岡 瑞起，松尾 豊: 検索エンジンを用いた関係の重みづけ，人工知能学会論文誌，Vol. 25, No. 1, 2010
5. Yingzi Jin, Ching-Yung Lin, Yutaka Matsuo, Mitsuru Ishizuka: Mining Longitudinal Network for Predicting Company Value. IJCAI 2011: 2268-2273