戦略的創造研究推進事業　CREST
研究領域「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」
研究課題「データ粒子化による高速高精度な次世代マイニング技術の創出」

研究終了報告書

研究期間　2014年　10月～2020年　3月
（新型コロナウイルス感染症の影響を受け2021年3月まで延長）

研究代表者：宇野　毅明
（国立情報学研究所
情報学プリンシプル研究系、教授）

§1 研究実施の概要

(1)実施概要

データの粒子は、巨大なデータが内包する最も低いレベルの抽象度を持つ構造であり、データから粒子を網羅的に抽出する作業をデータ粒子化とよぶ。データの多様性を獲得し、粒度の高い意味を抽出し、データ全体を広く深く理解するために当研究課題によって命名された概念である。例えば、ニュース記事群のクラスタ、という構造では、通常のアルゴリズムで求める構造は経済、社会などの大きなカテゴリに対応し、粒子は記事のトピック、特定のスポーツの記事群や、ある自然災害や事故の記事群などに対応する。ネットワークのコミュニティも粒子に対応すると考えられる。これまで、大きなカテゴリを見つけるためのアルゴリズムやクラスタリングなどの研究が盛んに行われてきたが、粒子に対応するような構造を精度良く見つける技術の研究は存在しなかった。実用上の観点から、粒子化には、少ない解でデータ全体を網羅すること、計算に再現性や一貫性があること、解の粒度がそろっていること、個別の解の意味解釈が容易であること、などが求められるが、これらをすべて満たす計算手法は存在しなかったのである。当研究課題では、新たにデータ研磨というアルゴリズムを開発することで、これらの要因をすべて満たすことに成功した。このデータ研磨技術を様々なデータに適用できるよう技術開発を行い、理論面から性能や挙動を解析し、意味構造の獲得技術を研究し、実利用のための可視化・インタラクションを設計し、さらに実利用における様々な問題を学術的な観点から研究していき、データ粒子化を基礎から応用まで網羅する技術体系を構築することが、この研究課題の目標である。そのために、モデル・アルゴリズムと理論、実利用、社会還元、インタラクションの各領域で研究を行った。

モデル・アルゴリズムと理論においては、宇野グループが中心となり、データ研磨アルゴリズムの開発と改良を行った。実データで高速で動作し、意味解釈性の高い解を生成することに成功し、収束性や速度についても理論解析を行った。計算が収束しないインスタンスが存在すること、現実的なべき乗則が成り立つデータにおいては計算が線形時間で収束することなどを示した。また、山本グループではKl閉パターンという新しい抽象構造の定義と計算アルゴリズムの開発にも成功した。木データへの拡張や、特徴学習への適用に関しては山本グループが中心となって研究を行った。これら開発したアルゴリズムは、羽室グループ、中小路グループにおいて、実応用とインタラクション設計に用いられている。

粒子よりも抽象度の低いレベルでの、より理論面の支持の強い解の生成アルゴリズムについては、宇野グループ中心に積極的な研究を行い、閉路や非閉路的な部分構造、k 縮退、内周が大きな部分グラフ、2 部グラフなどに対する多項式時間の効率的な列挙アルゴリズムを開発した。特に極大解を列挙するメタアルゴリズム proximity search は、当分野のトップカンファレンスである STOC に採択され、その技術の高さが評価された。

また、クラスタリングの解を複数用いて平均化することにより、クラスタリングアルゴリズムを安定化する技術を開発した。これにより、任意のクラスタリングアルゴリズムに対して再現性を持った計算を行うように改良することができるようになり、現実問題での利便性を高めることとなった。

データ粒子化技術の学術応用では、宇野グループが触媒化学でのデータ解析や、自然言語解析分野への適用がある。特に自然言語解析においては、データ研磨の持つ計算の一貫性、同一のデータからは同一の解を出し、類似するデータからは類似する解を出す、という性質を利用し、データ研磨の作るマイクロクラスタによる多様性の指標化を提案した。さらに投稿数と多様性の増加率の違いからデマ的なトピックを同定する技術を開発した。

産業応用としては、宇野グループが愛媛結婚支援センターでの婚活サイトでのリコメンデーション、株式会社 DAC での広告ターゲッティングに利用された。また、羽室グループと宇野グループの協業で、㈱光洋、㈱ブランシェスとの共同研究も行っており、データマイニング技術の小売り現場での実利用技術を開発した。

社会還元的な研究活動として、データ研磨アルゴリズム技術を中心としたデータマイニングツール群の開発を行った。これはデータ解析プラットフォームNysolに統合され、利便性の高い状態で提供されている。また、データ解析の入門書の執筆も行った。これらツールを用いたデ

ータ解析のハッカソン、データハッカソンを複数回開催し、一般の人々に対してデータマイニングの技術教育と啓蒙活動を行った。

データと解析結果を理解するためのインタラクション技術について開発を行った。ユーザが何を理解したいかわからない状況で、多量の情報を表示する必要のあるマイニング型データ解析のインタラクションは、今までの研究に全く存在しなかった困難性をはらんでおり、大きな挑戦であった。巨大な情報に対応するため、ユーザの着目点の推移に着目し、動きや色といったものによる誘発や誘導、意味解釈にとどまらず、因果推論やストーリーメイキングを行うための時系列提示などを用いたインタラクション設計を行い、複数のプロトタイピングによりビジュアルインタラクティビティ機構の開発を行った。

(2)顕著な成果

＜優れた基礎研究としての成果＞

1．データ研磨クラスタリングアルゴリズムの開発
概要：

粒度の高い小さなクラスタを網羅的に発見するあらたな手法データ研磨クラスタリングを開発した。既存手法と異なり、再現性・一貫性を持ち、解の数が爆発することもなく、高速であり、実データでの適用で納得のいく、意味解釈しやすいクラスタが網羅的に得られるようになった。トピックやコミュニティのマイニングの精度が飛躍的に向上した。収束性や計算量などの数理的側面も明らかになっている。

2．Proximity Search
概要：

解から解へと移動することで網羅的に解を全列挙するタイプのアルゴリズムの構築手法 proximity search を開発した。解の間に新たな近接性 proximity を導入し、それによって簡単な移動でも網羅性が担保できることを簡単に証明できるようになった。これにより、以前より未解決であった、極大誘導木、極大コーダルグラフ、極大連結非閉路的グラフ、極大な k 縮退グラフなどの列挙問題に対し、効率良いアルゴリズムの開発に成功した。

3．少数の例外を含む閉集合粒子の数理的定義と列挙
概要：

クラスタリング手法により求められたクラスタでは、クラスタ内の異なるデータがいくつかの属性を共有することが多く、この共通の属性を用いれば別のクラスタが構成できるはずである。閉集合はこの考え方に基づいて数理的に定義された理想の粒子であり、粒子化の数理的意味の基盤となる。閉集合はノイズを全く認めない密な粒子であるため、実用性を考慮し、少数の例外をノイズとして含む(k、l)−閉集合を数理的に定義し、その列挙アルゴリズムを与えた。

＜科学技術イノベーションに大きく寄与する成果＞

1．データハッカソンの開催
概要：

「データ解析ハッカソン」を開催することにより、データ研磨手法の産業界での利用促進と応用上の課題発見するための方法論を確立した。企業から出された実課題と実データに基づいたデータ解析を、多様なバックグラウンドを持つ参加者が実施することで、解析手法の利点・問題点が様々な観点から浮かび上がり、手法改良の指針を与えてくれる。このように、単なるデータ解析イベントとしてのハッカソンではなく、アカデミアからの情報発信と産業界からのフィードバックを効率よく実現する方法を確立した。

2. データ空間の複眼的ブラウジング機構による量子化学分野での知見創出
概要:
　粒子化されたデータ空間の複眼的ブラウジング機構を展開した、化学反応予測マップデータのためのインタラクティブな可視化環境 RMapViewer が、量子化学分野における知見発見に寄与した。RMapViewer の利用により、D-グルコースの分子構造の配座解析において、C14 と C41 の間の最小エネルギー反応経路の同定と熱力学的および動力学的な 1C4 の優位性を示すことにつながり、その成果は Journal of Chemical Theory and Computation に採択された。

3. 文書データの多様性解析技術への応用
概要:
　データ研磨が生成する、マイクロブログなどの多量の短い文章のマイクロクラスタを利用することで、トピックの多様性を形質として捉えることに成功した。データ研磨によるクラスタリングが、一貫性・再現性を持ちながら、網羅性高くマイクロクラスタを適切な数に抑えて見つけることができるために実現したものである。これにより、あるデマの流行や、ユーザの総体としての興味などが検知できるようになった。

＜代表的な論文＞
Takeaki Uno, Hiroki Maegawa, Takanobu Nakahara, Yukinobu Hamuro, Ryo Yoshinaka, Makoto Tatsuta, Micro-clustering by data polishing. BigData 2017: 1012-1018, 2017

Alessio Conte, Takeaki Uno, New polynomial delay bounds for maximal subgraph enumeration by proximity search. 51st ACM Symposium on Theory of Computing, 1179-1190, 2019.

Madori Ikeda and Akihiro Yamamoto: Extending Various Thesauri by Finding Synonym Sets from a Formal Concept Lattice, Journal of Natural Language Processing, Volume 24 Number 3, 323-350, 2018

§2 研究実施体制

(1)研究チームの体制について

(1)「計算技術とモデル化」グループ
① 研究代表者:宇野 毅明 (国立情報学研究所情報学プリンシプル研究系 教授)
② 研究項目
・様々なマイニングタスクに対するデータ粒子化モデルと研磨アルゴリズムの開発
・粒子化したデータの利用における補助モデルとアルゴリズムの開発
・匿名化など他の情報処理技術への応用技術開発

(2)「意味構造解析」グループ
① 主たる共同研究者:山本 章博 (京都大学情報学研究科 教授)
② 研究項目
・データの粒子化の意味論
・属性選択の高速化と意味論の構成
・構造データに対する意味論

(3)「実応用」グループ
① 主たる共同研究者:羽室 行信 (関西学院大学経営戦略研究科 准教授)
② 研究項目
・データ整備
・実データでの効果・効率の検証
・実データへの適用に関わる手法開発
・現実のビッグデータと手法の性質・特性の解明
・ハーデングメカニズムに関する理論構築

(4)「データインタラクション」グループ
① 主たる共同研究者:中小路 久美代 (公立はこだて未来大学システム情報科学部 教授)
② 研究項目
・洞察を誘導し着目点や思考の変化に柔軟に対応する効果的なビジュアルインタラクティビティの解明
・ユーザの着目点の抽出と連携および複合化のためのフォーカシング表現技術の確立
・粒子化されたデータ空間の複眼的ブラウジングを実現するデータインタラクション環境の構築

(2)国内外の研究者や産業界等との連携によるネットワーク形成の状況について
・ 宇野グループはイタリアピサ大学 Roberto Grossi 教授のグループと交流が深く、基礎的なマイニングアルゴリズムの理論について多くの成果を上げてきた。また、宇野グループは、オーベルニュ・クレモンフェラン大学の数理科学グループと交流が深く、基礎的なマイニング問題の数理構造や、計算量・アルゴリズムについて数多くの成果を上げている。
・宇野グループ、山本グループは、千葉商科大学橋本教授や情報学研究所小林助教などと、SNS 自然言語解析の共同研究を行っており、ここでデータ研磨の技術が中心的に使われている。
・宇野グループは、複数の自治体・企業と共同研究を行っており、データ研磨をはじめとするマイニング技術の開発と利用について議論を重ねてきた。
・CREST が支援してきたマイニングアルゴリズムの国際ワークショップにより、ヨーロッパと日本を中心としたコミュニティが育成されつつある。今まであまり日が当たってこなかったマイニングアルゴリズムの理論的研究に対して、今後大きな進展が見込まれる。

・中小路グループは、データを利用した科学計算における知識創出活動として、チューリヒ大学の Jurg Hutter 教授（Computational Chemistry）らとの連携を始めている。反応マップ経路の可視化環境開発を継続している。

・中小路グループは、開発したツールをアクティビティデータ履歴共有環境として位置付け、東京大学の岡田猛教授（教育学）らと協力し、学習プロセスの振り返りや知識創出の支援への展開に着手している。

・中小路グループは、時系列データをブラウジングする環境を、バスの運行経路データに適用し、GTFS 対応とした。これにより、函館バスをはじめとする地方都市のバス経路情報との連携に着手している。

・当初 2019 年度に開催予定であった 3 件のシンポジウム等については、新型コロナウイルス感染拡大により 2020 年度に延期したが、状況が改善せずいずれも実施を見合わせた。