「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」
平成 26 年度採択研究代表者

H28 年度
実績報告書

宇野　毅明

国立情報学研究所
教授

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

§1. 研究実施体制

(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、教授）
② 研究項目
・クラスタリングアルゴリズムの改良・精度向上
・バイクラスタリングアルゴリズムの開発、精度向上、検証
・アルゴリズムの並列化

(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
・2 部グラフ研磨とバイクラスタリングアルゴリズムの併用の効果検証
・バイクラスタリングアルゴリズムの分類と体系化
・人工データの作成
・教師無し学習における特徴(座標)選択法の開発

(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)「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 （京都大学 学際融合教育研究推進センター デザイン学ユニット、特定教授）

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. 研究実施の概要

本年の大きな進展の一つは、バイクラスタリングに対する技術開発である。昨年度開発したバイクラスタリングに対してアルゴリズムの高速化を行った。クラスタが少ない現実的なデータで高速になるよう、データの質を活かした設計が達成された。また、並列化も成功し、4コア使用した場合に3倍程度の高速化が達成できた。データ解析技術においては、クラスタを用いた属性分析の手法を開発した。また、過去に提案された4つのバイクラスタリングアルゴリズムに対して、各アルゴリズムがどのクラスタ構造をどの程度上手く再現できるかを、人工データを用いて実験的に明らかにし、二部グラフ研磨アルゴリズムを前処理と適用することにより、再現可能性が上昇することをも示すことができた。

また、データをグラフで表現した上で、モジュラ性を数値化する手法を利用して、できるだけ少ない数の特徴で、モジュラ性が高くなるようなものを選択するアルゴリズムを考案した。クラスタ解析手法に関しては、粒子化の基本的な考え方である、データそのものを抽象化し分析するという観点から手法開発を行った。属性を持つデータの中で、抽象化を行い、多様性を獲得し、得られた各グループの特徴持つメンバーを抜き出してグループ化することで、属性を持つデータがなぜその属性を持つことになったのか、その理由を意味づけしやすくなった。

また、研磨アルゴリズムを用いた名寄せアルゴリズムの開発を行った。論文タイトルや広告文句などのセンテンスを単語に分解したものを類似性に基づいて抽象クラスタ化することで、ときに 1000 倍を超える高速化を実現した。また、大規模データを効率的に処理するためのデータ解析ツールである NYSOL に新たなツールを追加し公開した。

応用では、金融データにおいては、類似度グラフに対するグラフ研磨によりハーディング行動の予測精度が高くなることも確認でき、グラフ研磨の有効性を実証した。また、腸内細菌の細菌叢をクラスタリングすると、体質とライフスタイルの関係性を腸内細菌の様相から明らかにし、効果的に健康指導が行える可能性を示した。

データおよび解析結果の可視化では、ユーザの主体的理解醸成のための機構を擁した環境を構成する要素として、markdown、css、および html の記述形式を組み合わせた簡易言語をデザインし、これを年表形式でウェブブラウザ上に表示するツールを開発した。開発した年表形式可視化ツールは、3 日間の情報創出ワークショップをミクロなレベルで記録したアクティビティの可視化表示や、羽室グループらが実施したデータハッカソンの粒子化過程のプロセスデータの可視化に適用した。洞察を誘導する情報表現を探るための動的表現としては、2016 年度に開発に着手した株価の値動き動向の粒子化データを可視化する環境 ViNL 上で、フォーカスしたデータエントリ（銘柄）と同じ粒子に属する他の銘柄を、回転させたり微細に動かすようなアニメーション試行環境の実装に着手した。時間情報を有する事象に関するテキストデータをインタラクティブにブラウンジグする環境 Timezoom を、継続的に展開した。ユーザの事象理解のフレームワークとしてノード連結によるネットワーク構造およびライン交叉によるメッシュ構造という2種類の誘導的な情報表現を実現した。タイムスタンプ付きのテキストデータとその粒子(ViTL: Visual Text Landscape)および株価データ（ViNL: Visual Number Landscape） という二種のデータインタラクション環境として開発した。

代表的な原著論文

1） データ研磨によるバイクラスタマイニング，宇野 毅明，小池 敦，中原 孝信，羽室 行信，第157回情報処理学会アルゴリズム研究会　pp.1-8　2017年3月

2) Takeaki Uno, Yushi Uno, Mining preserving structures in a graph sequence, Theoretical Computer Science, Volume 654, 22 November 2016, Pages 155-163.

3) Kumiyo Nakakoji, Yasuhiro Yamamoto, Yusuke Kita, Visual Interaction Design for Experiencing and Engaging with a Large Chronological Table, Proceedings of the 3rd HistoInformatics Workshop on Computational History (HistoInformatics 2016), pp.47-51, Krakow, Poland, July 11, 2016.