「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

H26 年度
実績報告書

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

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)研究代表者グループ
① 研究代表者:宇野　毅明 （国立情報学研究所・情報プリンシプル研究系・教授）
② 研究項目
・ クラスタリングを中心とするデータ粒子化の基礎モデル、アルゴリズム開発とその並列化、機械学習・予測・匿名化・圧縮などの基礎情報処理への応用

(2)共同研究グループ(1)
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
・木構造データはその特徴により、様々なクラスがあり、座標として用いることができる部分構造も多様であることが予想される。そこで、できるだけ一般的な部分構造とクラスに特有の部分構造について部分構造座標を用いて類似性尺度が定義可能な形にする。
・データ粒子の意味解析の最も単純な場合として、データ粒子に対する評価値導入手法を研究する。データ粒子化の具体的な領域として、プロセス解析に着目し、データ、各データ粒子に対して妥当な評価手法構成を目指す。

(3)共同研究グループ(2)
① 主たる共同研究者:羽室　信行 （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)共同研究グループ(3)

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）

① 研究項目

・着目点や思考の変化に柔軟に対応する可視化システムの構築
・認知対象間の距離と構造が導く認知のメカニズムの研究
・解析プロセスの保存と再生、共有手法の開発
・ユーザ操作に関するビッグデータの創出

§2. 研究実施の概要

データ活用の良さは、人間が見つけにくいような事実や仮説を見つけやすいところにある。計算機を使った場合、解析によって多面的な方面から、そのような発見の「手助け」ができるところに大きな利点がある。ビッグデータにおいてもこれは同じであるが、データの巨大さ、複雑さ、いいかげんさなどによって、このような解析の難しさが飛躍的に高くなってしまう。今までのデータマインニング手法では、この状況では解の質が非常に悪くなり、知識の候補となるパターンを1億個以上生成して検証を事実上不可能にしてしまう、細かく分類したはずが非常に多様な様々なものが混ざった大きなグループを作ってしまう、などの本質的に克服が難しい弱点を持っている。この課題では、データの中の構造(データの「粒子」とよぶ)を明確にすることによってこの難しさを解消する方法を研究する。具体的には、データの揺らぎをなくすように、周辺情報からデータを正しそうな形に整形し、そのことによって様々な構造を浮かび上がらせる。例えば図 01 のようにネットワークの中のコミュニティ(密度の濃い部分)を視覚化することができる。

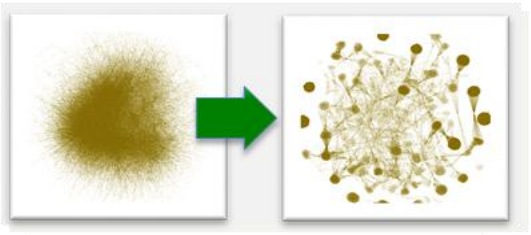

図 01: 研磨によるグラフの抽象化

今期は研究のスタートにあたり、様々な種類の研磨技術に対して基礎的な設計を行った。類似性を用いて頻出パターンを浮かび上がらせる方法、ネットワークのような離散的でない構造でも塊を浮かび上がらせる方法などの、現実的かつ効果的な設計はどのようになるべきかを議論し、基礎的な手法論の原型構築を行った。また、アルゴリズムの効率化についても設計を行い、現実的なデータの性質を上手に利用した高速化手法と、並列アルゴリズム構築に関わる基礎的な設計を行った。

また、データ研磨のプロトタイプを用いて、文献データからの研究者コミュニティ抽出を行った。従来コミュニティ抽出は論文共著ネットワークを用いて行われていたが、これでは face to face のコミュニケーションに基づいたコミュニティは発見できない。そこで研究会で顔を合わせた可能性がある、という情報を軸にマイニングすることで、よりおおきなコミュニティの発見を試みた。結果、研究会の 5 倍程度の数の、研究室のスタッフよりも 5 倍以上大きなコミュニティを見つけることができ、それらの人がある研究会をベースに他の研究会にも活動範囲を広げている様を確認することができた。データ研磨により、簡潔な解析のみで明確に事象を捉えることができたと考えており、解析結果は電子情報通信学会 I-Scover チャレンジにおいて最優秀賞を獲得することができた[受賞 2]。同様の解析はビジネスデータ、プログラムのモジュールのデータなどについても行った。結果は良好で、既存手法よりはるかに意味理解の容易な解を生成することができ、ビジネスデータのものについてもコンペティションにおいて優秀賞を獲得した[受賞 3]。

また、データ研磨によって作られる構造意味構造の理解を勧めるため、関係性が導く構造の中の基本的なものである、木構造について研究を行った。木構造の類似性を調べる新しい手法を開発し、木構造データ研磨に道を開き、同時に研磨構造の類似性の評価を可能にした[口頭講演 2]。また、研磨を利用するユーザの解析を進めるために、データ解析を行うさいの行動ログ取得に着手した。手近に手に入る、比較的簡単かつ明確なインターフェースと目的を持つ解析タスクとして京都の観光・歴史アーカイブデータに着目し、そのデータ取得を開始した。