



たくさんのデータの中から素早く知識を発見 ～計算の枝刈りを活用した高速かつ厳密な大規模データ解析～

Fast knowledge discovery from big data
– Large-scale data analysis with accuracy guarantee via efficient pruning methods –



NTT コミュニケーション科学基礎研究所 メディア情報研究部

藤原 靖宏

Yasuhiro Fujiwara

●プロフィール

NTT コミュニケーション科学基礎研究所 メディア情報研究部 特別研究員。
2001年早稲田大学理工学部電気電子情報工学科卒業。2003年早稲田大学
大学院理工学研究科電気工学専攻修士課程修了。同年日本電信電話株式
会社入社。2011年東京大学大学院情報理工学系研究科電子情報学専攻博士
後期課程修了。博士(情報理工学)。データベース、人工知能、機械学習の研究に
従事。電子情報通信学会、情報処理学会、日本データベース学会各会員。

今、データが知の源泉となっています。「データは新しい石油である。」というフレーズに注目が集まり、人工知能に基づいたデータ解析を効果的に活用することに関心が高まっています。しかし、膨大な収集データに対する解析には多大な計算コストが必要です。莫大な計算コストを低減するために近似的な処理を行うことが一般的ですが、近似処理では正確な解析結果を得ることができないという問題があります。本講演では、近似と異なり結果を変えない計算の枝刈りを行うことにより高速性と厳密性を両立したデータ解析を実現する代表的な手法を紹介します。

「データは新しい石油である。」この言葉は21世紀のデジタル時代における重要な概念の一端を象徴しています。近年デジタルデータの量は急増しています。市場調査会社によるとデジタルデータの量は2020年には約59ゼタバイト(約590億テラバイト)に達し、2025年には163ゼタバイト以上に達すると予測されています。データに潜むパターンやトレンドを発見し人間の意思決定をサポートすることができるデータ解析は、データという新しい石油を利活用していくうえで重要な技

術になります。

しかし膨大なデータに対してデータ解析を行うには莫大な計算コストが必要になるという課題があります。この課題はデータ解析の計算コストが高いことに起因します。計算コストを低減するためには近似計算を行うことが一般的です。しかし近似計算によって高速に得られた解析結果は厳密な解析結果に対して誤差が生じてしまうという問題があります。データ解析は人間の意思決定をサポートするために用いるため、解析結果の厳密性を犠牲にするアプローチは好ましくありません。

そこで、私たちは高速性と厳密性を両立したデータ解析のための機械学習基盤を構築に向けた研究開発を進めています(図1)。この機械学習基盤において高速性と厳密性を担保するために鍵となるのが計算の枝刈り手法です。これは計算結果の厳密性を侵さない範囲で不要な計算を省くことで高速化を実現します。その代表的なものとして、1. 上限値・下限値による計算の省略、2. 解になりえない計算の打ち切り、3. 楽

観的処理による高速計算を紹介します(図2)。

上限値・下限値による計算の省略の例としてCUR分解の高速化の研究があります[1]。CUR分解は入力された一つの行列をその行列に含まれる少数の行で表現する行列分解アルゴリズムです。オリジナルのCUR分解では、最適化の過程で行毎にスコアを計算し、そのスコアの絶対値が閾値より小さければその行の全ての要素をゼロにする処理を繰り返します。しかしこの方法は計算コストが高いという問題がありました。そこで私たちはスコアの上限值と下限値を軽量に計算し、ゼロになる行を高速に特定し、非ゼロになる行を集中的に学習することで高速化を実現しました。

解になりえない計算の打ち切りの例として部分グラフの高速マッチングの研究が挙げられます[2]。部分グラフマッチングは大規模なデータグラフから問い合わせのグラフと同じ構造を持つ部分を特定する処理になります。この方法は問い合わせグラフのノード一つ一つをデータグラフにマッピングして照合するため、最悪の場合の時間計算量はデータ数に対して指数的に増加します。そこで私たちはノード一つ一つのマッピングにおいて照合が失敗したパターンを保持し、マッピング処理において失敗したパターンが再度現れた場合に処理を早期に終了する手法を提案しました。

また、楽観的処理による高速計算の例として近傍グラフにおけるエッジの重みの高速計算が挙げられます[3]。近傍グ



図1：機械学習基盤

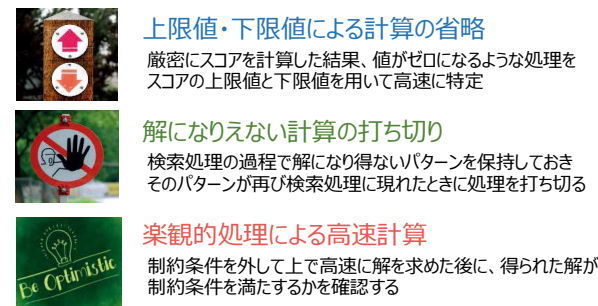


図2：代表的な計算の枝刈り手法

●参考文献

- [1] Y. Ida, S. Kanai, Y. Fujiwara, T. Iwata, K. Takeuchi, H. Kashima, “Fast deterministic CUR matrix decomposition with accuracy assurance,” in Proc. the 37th International Conference on Machine Learning (ICML), 2020
- [2] J. Arai, Y. Fujiwara, M. Onizuka, “GuP: fast subgraph matching by guard-based pruning,” in Proc. the ACM on Management of Data (SIGMOD), 2023
- [3] Y. Fujiwara, A. Kumagai, S. Kanai, Y. Ida, N. Ueda, “Efficient algorithm for the b-matching graph,” in Proc. the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2020