

大規模なグラフデータをあっという間に分析

～グラフデータを対象とした高速なマイニングアルゴリズム～

どんな研究

実世界にはソーシャルメディアやSNSなどの情報のつながりを表す膨大なグラフデータが存在します。従来の技術では、大量のグラフデータの分析のために、膨大な時間を要していました。今回は、このグラフデータを対象にした高速な分析アルゴリズムを紹介します。

どこが凄い

グラフデータの統計的な性質を用いて、グラフデータをよりコンパクトな形に変形していくことで、**世界最高速のグラフデータ分析アルゴリズム**を確立しました。このアルゴリズムを用いることで1億人分の友人関係を表すグラフデータも分析精度を落とさず3分程度で解析できます。

めざす未来

TwitterやSNSなどの極めて大規模なソーシャルメディアや人間の**行動履歴などを瞬時に分析し、情報推薦や情報予測、情報理解に活用**できるような未来を目指しています。本技術は、従来考えられなかった規模のデータの分析に貢献できる可能性があります。

■ パーソナライズドページランク^注 高速化技術 (関連文献[1])

POINT

- 人間関係や購買履歴などのグラフから重要なデータを高速に発見
- 従来技術よりも**10～100倍**程度高速に処理可能

- ⊙ グラフを行列にし、行列分解と並べ替えてグラフをコンパクトにして重要度を計算

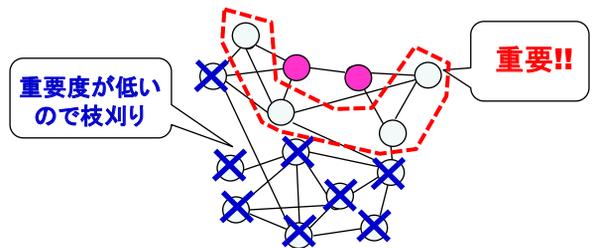
$$S = (1-c)G \begin{bmatrix} S \\ cN \end{bmatrix}$$

S: ノードの重要度ベクトル **M**: 趣向ベクトル
G: 隣接行列 **c**: ランダムウォーク再スタート確率

$$S = C \begin{bmatrix} P^T \\ \left[\begin{array}{c} Q \\ R \end{array} \right]^{-1} P \end{bmatrix} N$$

P: 置換行列 **Q**: QR分解における直交行列
R: QR分解における上三角行列

- ⊙ 重要度の低いデータを枝刈りし、グラフをさらにコンパクトに変換



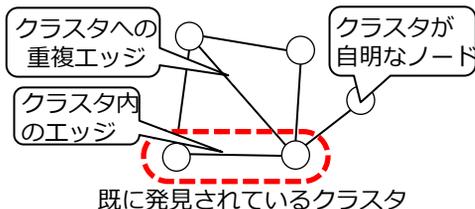
注) パーソナライズドページランク: 個人に有用な情報検索技術の1つ

■ クラスタリング高速化技術 (関連文献[2])

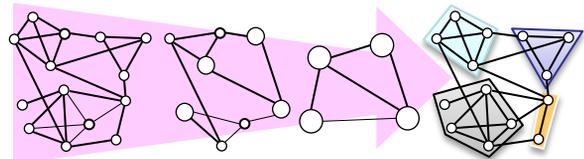
POINT

- 人間関係や購買履歴などのグラフの中からコミュニティを高速に発見
- 1億人規模のデータも3分**で処理可能

- ⊙ グラフから重複計算の生じるノードとエッジを削減



- ⊙ 計算不要なデータを逐次的に集約しデータをコンパクトにすることで計算を高速化



関連文献

- [1] Y. Fujiwara, M. Nakatsuji, T. Yamamuro, H. Shiokawa, M. Onizuka, "Efficient Personalized PageRank with Accuracy Assurance," in *Proc. the 18th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2012.
 [2] H. Shiokawa, Y. Fujiwara, M. Onizuka, "Fast Algorithm for Modularity-based Graph Clustering," in *Proc. the 27th AAAI Conference on Artificial Intelligence (AAAI)*, 2013.

連絡先

鬼塚 真 (Makoto Onizuka) NTT ソフトウェアイノベーションセンター 分散処理基盤技術プロジェクト
 E-mail: onizuka.makoto[at]lab.ntt.co.jp ({at}の部分をも@に置き換えてください)