03

Efficient graph clustering and distributed query optimization

大量なデータ間のつながりから隠れた知識を発見

～高速グラフクラスタリングと分散クエリ最適化～

どんな研究

本研究は、SNSやWebなどのつながりを表す**大規模なグラフデータを従来技術よりも高速に分析する技術**です。グラフデータ分析することにより、大規模なデータから仲良し友人グループや影響力のある人物などのデータに隠れた知識を発見します。

どこが凄い

グラフデータの統計的な性質を用いることで、**高速なグラフデータ分析手法**を確立しました。このアルゴリズムを用いることで、**データ数が数百万～数千万件規模**のグラフデータを従来手法よりも**70%～90%短い時間で分析**することができます。

目指す未来

TwitterやSNSなどの大規模なソーシャルメディアやログデータ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。

● ページランク高速化技術（関連文献[1]）

人間関係などの大規模な**グラフデータの中から影響力（ページランク値）の高いデータ**を**従来手法より高速**(Wikipedia データで1桁高速)に発見します.

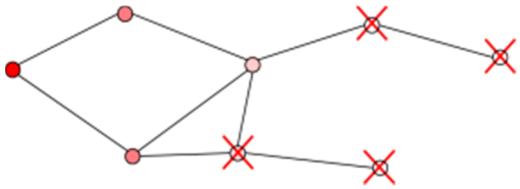
- 以下の式展開に基づきページランクの上限値と下限値を再帰的に算出します

$$s = cWs + (1-c)e$$
$$= (1-c)\left(e + cWe + c^2W^2e + \cdots\right)$$

s :ページランクベクトル　c :ランダムウォークの確率

W :グラフの隣接行列　e :全ての値が**1/n**のベクトル

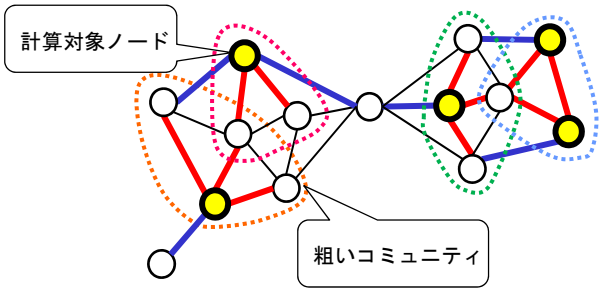
- ページランクの小さなノードを枝刈りしグラフを徐々にコンパクトにします



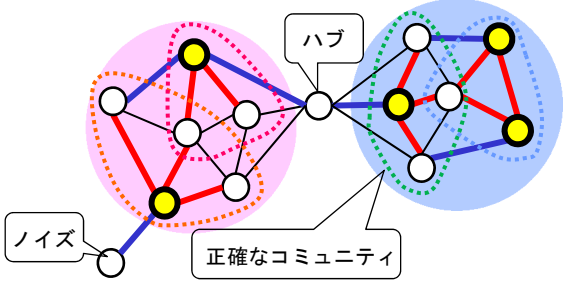
● クラスタリング高速化技術（関連文献[2]）

人間関係や購買履歴などの大規模な**グラフデータの中に隠れたコミュニティやハブ、ノイズ**となるデータを**従来手法より70%以上高速**に発見します。

- 最短距離が**2**ホップ離れたノードのみを計算し、粗いコミュニティを算出します



- 複数のコミュニティに所属するノードを見つけ正確なコミュニティを修正します



関連文献

[1] Y. Fujiwara, M. Nakatsuji, H. Shiokawa, T. Mishima, M. Onizuka, “Fast and exact top-k algorithm for PageRank,” in *Proc. the 27th AAAI Conference on Artificial Intelligence (AAAI2013)*, 2013.

[2] 塩川浩昭, 藤原靖宏, 鬼塚真, “構造的類似度に基づくグラフクラスタリングの高速化,” 第6回データ工学と情報マネジメントに関するフォーラム(DEIM Forum 2014), D6-2, 2014.

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 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@に置き換えてください)