超高速なビッグデータ解析法

Data Skewnessを捉えた超高速・省メモリな大規模データ処理

塩川浩昭（筑波大学 計算科学研究センター・准教授）
shiokawa@cs.tsukuba.ac.jp

筑波大学 University of Tsukuba

研究概要

● 解決したい課題

限られた計算資源の中でいかにして
大規模データを高速に計算するか？

【現在の大規模データ処理の課題】
・大規模データ処理に必要な計算コストと
普及した計算機の性能にはギャップがある

【本研究の目的】
・高速・省メモリなデータ処理アルゴリズムの
開発により，ギャップの解消を目指す

● 成果概要（ACT-I & ACT-I加速フェーズ）

数億件のデータを約数千倍高速に計算可能

【グラフデータ処理】
・Modularityクラスタリングの高速化 【IJCAI'19】
・構造的クラスタリングの高速化【DEXA'18 & '20】
・不確実グラフの信頼性検索高速化 【DEXA'20】

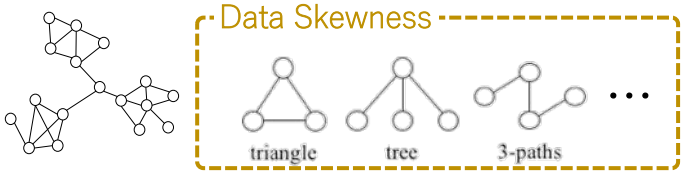
【多次元データ処理，データベース処理】
・クラスタリングの高速化 【AAAI'21】
・グラフデータベースの高速化【DEXA'20, WISE'20】
・知識データベース類似度検索の高速化 （投稿中）

研究成果の詳細

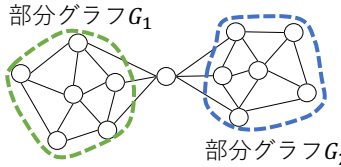
● 基本的なアプローチ：Data Skewnessを捉えた計算コスト削減

・Data Skewness ＝「実データのみが持つ頻出する（特徴的な）部分構造」に着眼
・Data Skewnessを捉えることで，冗長な計算を大幅に削減する

【着眼点①】実データ偏った局所構造を持つ



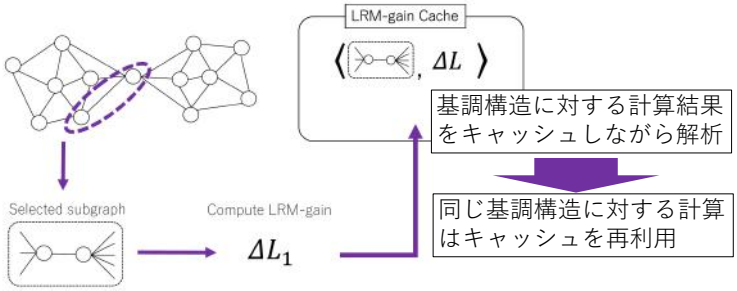
【着眼点②】局所構造に対して決定性がある

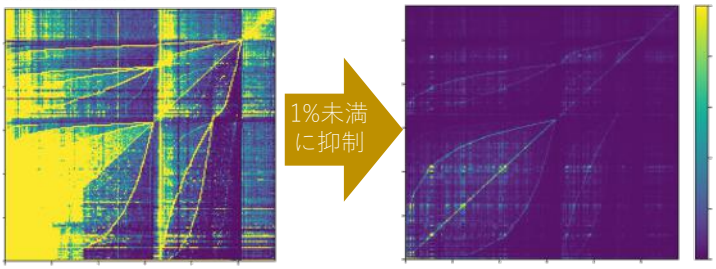


G_1とG_2が同型なので
$f(G_1) = f(G_2)$

● 研究成果①：gScarf法

・Modularityクラスタリングに対する高速化手法
・正確な処理結果を1,000倍以上高速に計算可能





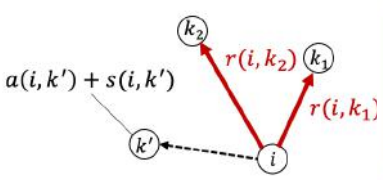
[Duan et al., KDD 2015]　　提案手法 gScarf

H. Shiokawa, T. Amagasa, H. Kitagawa, "Scaling Fine-grained Modularity Clustering for Massive Graphs," In Proc. IJCAI 2019.

● 研究成果②：ScaleAP法

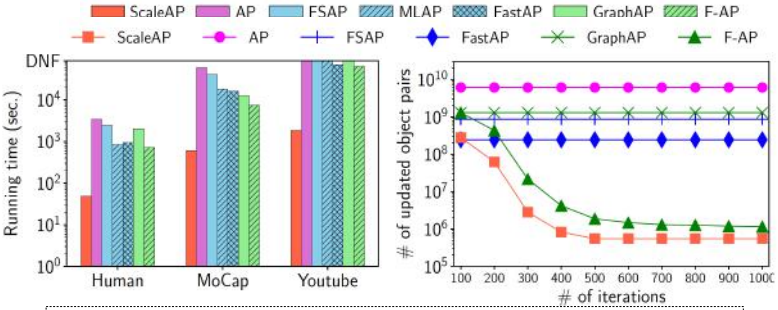
・Affinity Propagation (AP)に対する高速化手法
・APと同じ結果を1,000倍程度高速に計算可能

AP内部で生じるメッセージ更新を決定性性質
を利用することでN分の1にまで抑制する



$k_1, k_2 \neq \operatorname{argmax}_{k'} \{a(i,k') + s(i,k')\}$
ならば以下の性質が成り立つ

決定性性質
$s(i,k_1) = s(i,k_2) + \delta$
$\Rightarrow r(i,k_1) = r(i,k_2) + \delta$



H. Shiokawa, "Scalable Affinity Propagation for Massive Datasets," In Proc. AAAI 2021.