様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　　元 年　　６ 月 １０ 日現在

機関番号：６２６１５
研究種目：若手研究(B)
研究期間：2016～2018
課題番号：１６Ｋ１６１１５

研究課題名（和文）統計的に有意な部分構造を発見する巨大グラフマイニング手法の研究

研究課題名（英文）Finding Significant Subgraphs from Big Graph data

研究代表者
杉山　麿人（Sugiyama, Mahito）

国立情報学研究所・情報学プリンシプル研究系・准教授

研究者番号：１０７３３８７６

交付決定額（研究期間全体）：（直接経費）　　3,100,000円

研究成果の概要（和文）：本研究プロジェクトでは，グラフ構造をもつ巨大なデータに対して，部分構造（部分グラフ）を発見し，それらの統計的有意性を保証することを目的とし，研究をおこなった．まず，情報幾何の理論をこの問題に導入することで，部分グラフの探索において不必要な領域を積極的に取り除くことを可能にするための基礎理論を構築した．また，確率的論理プログラミングによる学習を利用することで，得られた部分グラフ集合を簡潔に表すことに成功した．さらに，大量に得られる部分グラフに対する解析をより進めるために，部分グラフ間の類似度を測るためのグラフカーネル手法について，広く利用可能なR及びPythonのパッケージを開発し，公開した．

研究成果の学術的意義や社会的意義
ソーシャルネットワークから分子構造まで，様々な対象がグラフと呼ばれる構造で表現される．多くの場合，グラフ中の特定の部分構造が，重要な役割を担っている．しかし，これまでは，統計的有意性を保証しつつそれらの重要な部分グラフを発見する手法は確立されていなかった．本研究では，この目的を達成するための手法を研究するとともに，いくつかの実用的なアルゴリズムを提案することに成功した．本研究は，情報学にとどまらず，生命科学や化学など，様々な分野へ応用可能な基盤技術となることが期待される．

研究成果の概要（英文）：This research project has developed methodologies for finding statistically significant substructures from massive scale graph structured datasets. We have established an information geometric formulation that enables us to remove unnecessary candidates during the search for significant subgraphs. Moreover, we have successfully constructed a learning method that can compress detected subgraphs using probabilistic logic programming. Furthermore, to further address data analysis for subgraphs, we have implemented various graph kernels that can measure the similarity between subgraphs and published as R and Python packages.

研究分野：知能情報学

キーワード：グラフ　グラフマイニング　統計的有意性　多重検定　検定可能性　パターン　情報幾何　半順序集合

１．研究開始当初の背景

巨大グラフデータが実世界の様々な分野で獲得され，大量に蓄積されてきている．例えば，Web 上のソーシャルネットワークや，タンパク質間相互作用ネットワーク，遺伝子ネットワークなどが挙げられる．これらグラフ構造を持つデータからの知識発見を目的とするグラフマイニングは，データマイニングにおける主要なトピックの 1 つとして盛んに研究されてきた．しかし，単一の巨大グラフに特化したグラフマイニング手法は発展途上であった[文献①]．一方，データ解析において，A/B テストに代表される統計的仮説検定は，基本的な解析手法の 1 つとして重要性を増し，データマイニングと仮説検定を融合した手法が出現してきた．グラフマイニングによるグラフ構造データの解析においても，単に部分グラフの頻度だけから重要性を評価するのではなく，仮説検定と組み合わせることで，偽陽性の割合を適切に制御しつつ統計的に有意に現れる部分グラフを発見することが可能になってきた[文献②]．しかし，依然として，巨大グラフから統計的に有意に現れている部分グラフを発見する問題は未解決であった．

２．研究の目的

本研究の目的は，巨大グラフに現れる統計的に有意な部分グラフを効率的に列挙する手法を構築することである．さらに，構築した手法を実世界の巨大グラフデータへ適用して有用性を確認する．この目的を達成するために，(1)計算量の爆発，(2)多重検定に起因する偽陽性の増加，(3)頻度の非単調性，という 3 つの問題を解決する．

３．研究の方法

まず，(1)既存手法の融合によるプロトタイプの設計と検証に取り組む．巨大グラフの特殊系である木構造データに着目し，統計的に有意に出現する木パターンを発見する手法を構築する．木パターン列挙アルゴリズムに，研究代表者らがこれまでに開発した多数の小規模グラフから偽陽性の割合を制御しつつ統計的に有意な部分グラフを発見する手法を融合することで，巨大木構造データを扱うことができる手法のプロトタイプを構築して実装する．このプロトタイプを用いて，問題の詳細な分析をおこなう．

その後，プロトタイプで得た結果をもとに，(2)巨大グラフを扱うことができる理論的基盤を構築する．特に，各部分グラフの頻度が，情報幾何で扱う二重平坦構造の座標系になっていることに着目することで，部分グラフ群がなす空間の性質を情報幾何的に捉える．

また，データによっては，非常に多く（数百万個以上）の有意な部分グラフが発見される可能性があるため，(3)出力された部分グラフの圧縮・要約による解釈可能性の向上に取り組む．グラフが論理プログラムとして記述できることに着目し，確率的論理プログラミングを導入することで，得られる部分グラフの要約を実現する．

４．研究成果

最初に，グラフの一種である木構造データに着目し，木パターンのマイニングアルゴリズムと Tarone の検定可能性と呼ばれる多重検定手法を融合することで，統計的に有意な木パターンを発見する手法を確立した．しかし，大規模なデータでは，(1)当初の想定以上にアルゴリズムの実行時間が増大してしまい，かつ(2)統計的に有意な部分グラフが大量に見つかり結果の解釈が困難である，という 2 つの課題があるという結果を得た．木構造データはグラフの特殊系なので，巨大グラフからのマイニングは，木構造データからのマイニングよりもさらに困難な問題となる．したがって，マイニングアルゴリズムと Tarone の検定可能性を直接融合するだけでは，目的の達成が困難であるという見通しを得た．

課題(1)を解決するためには，木パターンのマイニングアルゴリズムそのものを改善する必要があるが，これは本研究の主要な目的からは外れてしまう．そこで，より本質的な問題解決のために，情報幾何の理論を導入することで，解の探索において不必要な領域を積極的に削除するための基礎理論を構築した．より具体的には，巨大グラフ中の部分グラフ集合を半順序集合として扱うことで，情報幾何で基本的な構造として用いられている二重平坦な多様体の構造が自然に導入できることを示した．二重平坦構造を構

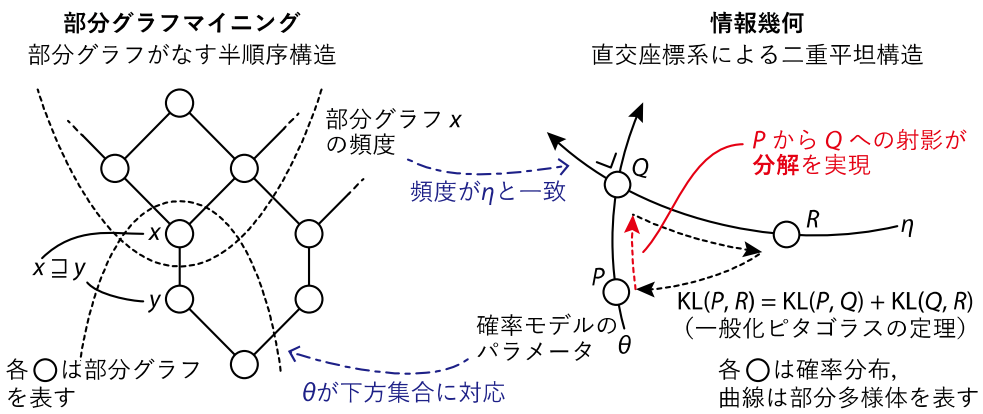

図 １．部分グラフマイニングと情報幾何の関係

成する 2 つの座標系が，それぞれ指数型分布族の自然パラメータと，各部分グラフの出現回数に対応することを明らかにした（図 1）．これによって，探索領域の組合せ爆発を回避しつつ，統計的有意性の判定を実行することができるようになった[発表雑誌論文②④⑤].

さらに，課題(2)を解決するために，部分グラフの集合を要約するための手法を構築した．特に，確率的論理プログラミングによる学習を利用することで，部分グラフ集合を表す簡潔な表現を獲得することに成功した．具体的には，中間結果として得られる検定可能な部分グラフを利用することで，統計的に有意だった部分グラフを正例，検定可能だが統計的に有意でない部分グラフを負例として，正例を高精度で表現することができる確率的論理プログラミング表現を学習する手法を構築した（図 2）[発表雑誌論文①]．この手法を用いることによって，数百～数千個の部分グラフを，わずか数個の節で表現することが可能となった．

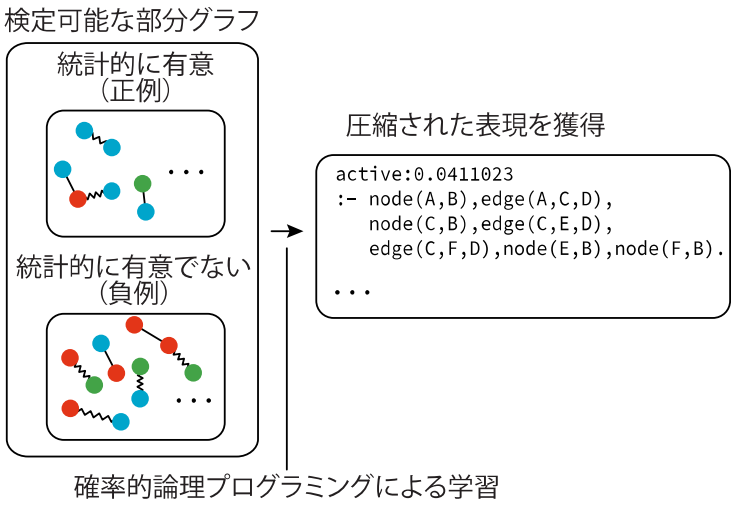

図 2．確率的論理プログラミングによる部分グラフの要約

上記の研究と平行して，得られた部分グラフ集合を適切に扱うための機械学習手法の研究を進めた．得に，部分グラフ間の類似度を測るためのグラフカーネル手法について，広く利用可能な R 及び Python のパッケージを開発し，公開した[発表雑誌論文③].

＜引用文献＞

① GRAMI: Frequent Subgraph and Pattern Mining in a Single Large Graph
Elseidy, M. and Abdelhamid, E. and Skiadopoulos, S. and Kalnis, P.
Proceedings of the VLDB Endowment 7 517-528 2014 年

② Significant Subgraph Mining with Multiple Testing Correction.
Sugiyama, M., Llinares-López, F., Kasenburg, N., Borgwardt, K. M.
Proceedings of the 2015 SIAM International Conference on Data Mining, 37-45 2015 年

５．主な発表論文等

〔雑誌論文〕（計　6　件）

① Summarizing Significant Subgraphs by Probabilistic Logic Programming
Bellodi, E., Satoh, K., Sugiyama, M.
Intelligent Data Analysis　2019 年（査読有，印刷中）

② Legendre Decomposition for Tensors
Sugiyama, M., Nakahara, H., Tsuda, K.
Advances in Neural Information Processing Systems (NeurIPS2018) 31 8825-8835 2018 年 12 月（査読有）

③ graphkernels: R and Python Packages for Graph Comparison
Sugiyama, M., Ghisu, E., Llinares-López, F., Borgwardt, K. M.
Bioinformatics 34(3) 530-532 2018 年 2 月（査読有）

④ Tensor Balancing on Statistical Manifold
Sugiyama, M., Nakahara, H., Tsuda, K.:
Proceedings of the 34th International Conference on Machine Learning (ICML 2017)　3270-3279 2017 年 8 月（査読有）

⑤ 統計的有意性を担保するパターンマイニング技術
杉山麿人
オペレーションズ・リサーチ誌 62(4) 226-232 2017 年 3 月

〔学会発表〕（計　10　件）

① Machine Learning and Information Geometry

杉山麿人
日独先端科学（JGFoS）シンポジウム　2018 年 9 月（招待有）

② Information Geometric Analysis on Partial Order Structures
Mahito Sugiyama
Kyoto University Informatics Seminar　2017 年 10 月 19 日（招待有）

③ ネットワーク構造上の統計モデルと情報幾何的な解析
杉山麿人
第 10 回情報ネットワーク科学研究会　2017 年 10 月 17 日（招待有）

④ Significant Pattern Mining on Graphs
Mahito Sugiyama
10th International Conference on Multiple Comparison Procedures　2017 年 6 月

〔図書〕（計　2　件）

① Searching for Bacterial Pathogens in the Digital Ocean—Executive Summary
Giuliano, L., Dorman, C., Bowler, C., Sugiyama, M., Vezzulli, L., Czerucka, D., Le Roux, F., D'Auria, G., Troussellier, M., Briand, F.
CIESM Workshop Monograph 49 5-25 2017 **年（依頼有）**

② Finding Statistically Significant Patterns from Data
Sugiyama, M.
CIESM Workshop Monograph 49 53-58 2017 **年（依頼有）**

〔産業財産権〕
○出願状況（計　0　件）
○取得状況（計　0　件）

〔その他〕
ホームページ等
https://mahito.info

６．研究組織

(1)**研究分担者**
なし

(2)**研究協力者**
なし

※科研費による研究は、研究者の自覚と責任において実施するものです。そのため、研究の実施や研究成果の公表等については、国の要請等に基づくものではなく、その研究成果に関する見解や責任は、研究者個人に帰属されます。