COPINE: 共通パターンを持つネットワーク抽出

関 美緒　(指導教員：瀬々 潤)

博士前期課程 理学専攻情報科学コース (学籍番号: 0740644)

1　はじめに

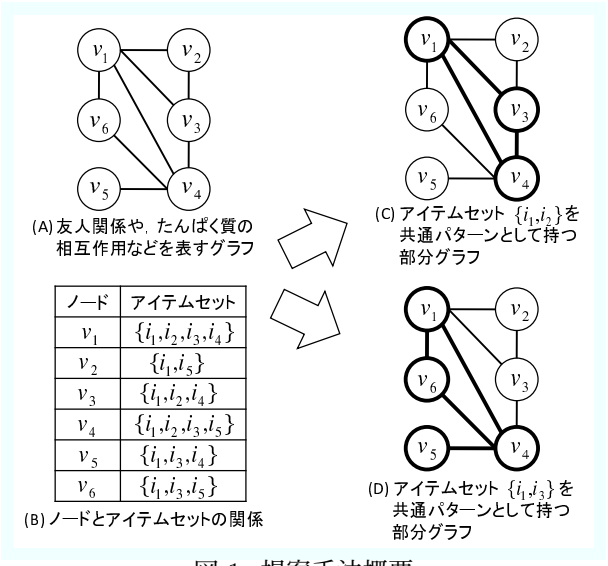
近年，グラフマイニング研究は，ソーシャルネットワーク，生物学的ネットワーク，コンピュータネットワークといった幅広い分野に渡っている [1]．本研究では，ノードにアイテムの集合を持つ新たなグラフモデルを考える．例えば，各ノードが人を表し，それぞれ買った商品をアイテムとして持つ友人関係を表すネットワークや，各ノードが遺伝子を表し，それぞれの活性化する実験環境をアイテムとして持つたんぱく質の相互作用ネットワークなどである．これらのネットワークにおいて，共通のアイテム集合を持つ部分ネットワークは，商品の広まりと関係のあるグループや，同じ環境下で働いている細胞内ネットワークを示し，創薬研究などに役立つ．しかし，大規模なネットワークには多くの部分ネットワークが存在し，同型部分グラフも多く含まれるため抽出は困難である．本研究では，深さ優先探索木を用いて部分ネットワークの列挙を行い，新たな 2 つの枝刈り手法を導入することで探索空間を大幅に減らす COmmon Pattern Itemset NEtwork mining (COPINE) という手法を考案した．

2　関連研究

グラフマイニングの手法として，頻出部分グラフの抽出 [2, 3] やグラフのクラスタリング [4]，制約付きクラスタリング [5] がある．頻出部分グラフの抽出は，全ての部分グラフを列挙し，頻出している部分グラフを見つけることを目的とするが，本研究では，全ての部分グラフを列挙し，その中で共通のアイテムセットを持つ部分グラフを見つけることを目的としている．グラフのクラスタリングでは，エッジの重みを利用するが，本研究で扱う共通パターンを持つグラフは，1 つのエッジを異なる 2 つの部分グラフで共有することがある．また，ノードの次数を用いるグラフのクラスタリングは，口コミや，生物学的なパスウエイにおいて，このような情報は低い次数を持つノードを通って伝達されることがあるため有用でない．制約付きクラスタリングは，ネットワークの同時クラスタリングや，ノードと関係のある数値ベクトルを見つけることを目標としている．これらの手法と，本研究との違いの 1 つは，全てのノードに関連づけられているアイテムが離散の値を持つことである．制約付きクラスタリングでは離散の値が扱えないので，パターンの一部を共有するグラフを見つけることは困難である．

3　研究内容

図 1(A) のグラフは，友人関係やたんぱく質の相互作用を表すネットワークを例示したものである．各ノードが人や遺伝子で，友人関係や相互作用のあるノード間にエッジが張られている．また，ノードは各人が買った商品や各遺伝子が活性化した条件を持ち，ノードのアイテムセットと呼ぶ．図 1(B) は，各ノードのアイテムセット表す．本研究の目的は，このデータから，共通のアイテムセットを持つ連結した部分ネットワークを見つけることである．共通のアイテムセットとは，部分ネットワークに含まれる全てのノードが持つアイテムセットの積集合を表す．図 1(C)(D) に図 1(A)(B) における共通アイテムセットを持つ部分ネットワークを示す．図 1(C) の太線で表されている部分ネットワークは，含まれる全てのノードがアイテムセット $\{i_1, i_2\}$ を持つ部分ネットワーク，図 1(D) は $\{i_1, i_3\}$ を持つ部分ネットワークである．



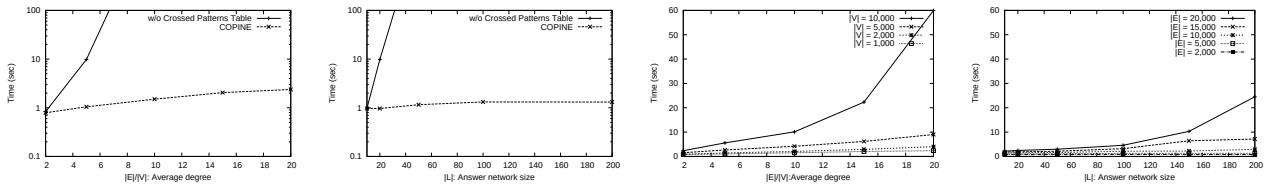
ノード	アイテムセット
v_1	$\{i_1,i_2,i_3,i_4\}$
v_2	$\{i_1,i_5\}$
v_3	$\{i_1,i_2,i_4\}$
v_4	$\{i_1,i_2,i_3,i_5\}$
v_5	$\{i_1,i_3,i_4\}$
v_6	$\{i_1,i_3,i_5\}$

図 1: 提案手法概要

与えられたネットワークには，様々な大きさ，共通アイテムセットを持つ部分ネットワークが存在する．人々が購入する商品がランダムな場合，二人が共通した商品を買うことはほとんどなく，また，隣接するノードが持つアイテム間に関連が無ければ，共通アイテムセットを持つ部分ネットワークは小さいものとなる．そこで本研究では，大きさ上位 N 個の，ユーザの指定した閾値以上の共通アイテム数を持つものを，有意な共通アイテムセットを持つ部分ネットワークとした．

しかし，与えられたネットワークの大きさにより，部分ネットワークの数は指数関数的に増加するため，有意な共通アイテムセットを持つ部分ネットワークの探索は困難である．この問題を解決するため，本研究では新しいアルゴリズム COPINE を導入する．COPINE は，深さ優先探索アイテムセット木 (DFSIT) と，探索済みパターンテーブル (CPT) というデータ構造を保持している．

DFSIT は，木の各ノードがネットワークのノード v と v を探索した時の共通アイテムセットを持つ，深さ優先探索木である．COPINE は，深さ優先探索に基づいて探索を行い，DFSIT を生成する．また，共通アイテムセットの大きさは，探索においてノードが 1 つ加わると，その部分ネットワークの持つ共通アイテムセットの大きさは，元の部分ネットワークの共通アイテムセットの大きさと同じか小さくなり，単調減少で



(a) ネットワークの大きさ (b) 抽出ネットワークの大きさ (A) 探索済みパターンテーブルの効果

(a) ネットワークの大きさ (b) 抽出ネットワークの大きさ (B) COPINE のスケーラビリティ

図 2: COPINE の性能評価

ある．DFSIT とアイテムセットの単調減少の性質を用いることで，共通アイテムセットの大きさによる枝刈りを行うことができる．

CPT は，既に探索したノードとアイテムセットとの関係を保持している．共通アイテムセットの大きさによる枝刈りを用いても，同じエッジの重複探索を省くことはできない．そこで，CPT と次に述べる性質を用いた枝刈りも導入した．ただし，テーブルを引く時間やメモリの使用量が増えるため，本テーブルを利用して実行時間が確かに速くなることは，疑似データを用いて確かめる．n_1，n_2 を DFSIT のノードとし，n_1 は n_2 より前に生成されているとすると，n_1 と n_2 両方がネットワーク中のある共通のノードを持ち，n_2 のアイテムセットが n_1 のアイテムセットの部分集合である場合，n_2 の子孫は既に探索されている．これにより，あるノード v を探索したときの共通アイテムセットが，保持されている探索済みパターンの部分集合となっていたら，探索を終了することができる．

この 2 つを組み合わせることにより，探索空間を大幅に減らすことができ，大規模なデータからも有意な共通アイテムセットを持つ部分ネットワークの抽出が可能となった．

4 実験

性能評価

疑似データを用いて，探索済みパターンテーブルの有用性と，COPINE のスケーラビリティを調べる実験を行った．COPINE は Java を用いて実装し，実行には 2.2GHz AMD Opteron，1GB メモリの Linux 2.6 を用いた．図 2 に結果を示す．縦軸は実行時間を，横軸は図 2(A)(B) ともに (a) はネットワークの平均次数，(b) は抽出する第 1 位のネットワークのエッジ数を表す．平均次数が増加すると，部分ネットワークの数は指数関数的に増加し，抽出するネットワークが大きくなると，重複探索が多くなる．

図 2(A) は，CPT を用いた枝刈りを導入しない場合と COPINE を利用した場合の実行時間の比較結果である．CPT を用いない場合，実行時間は平均次数や，抽出するネットワークが大きくなると指数関数的に長くなるが，COPINE の場合は抑えられており，COPINE が大幅に実行時間を速くすることを示している．

図 2(B) は，COPINE のスケーラビリティを調べた結果である．平均次数や，抽出するネットワークが大きくなっても，実行時間の増加がほぼ抑えられていることがわかる．また，図 2(B) の (a) より，COPINE は 20 万エッジものネットワーク探索においても，現実的な実行時間で部分ネットワークを抽出できることがわかる．

実データを用いた実験

COPINE により抽出された，部分ネットワークの有用性を調べるため，生物学の実データを用いて実験を行った．エッジ数 7,564 の酵母のたんぱく質たんぱく質相互作用データ [6, 7, 8] をネットワークデータ．173 種類のストレス環境下で 6,152 遺伝子について実験を行った遺伝子発現データ [9] を，アイテムセットデータとして利用した．その結果，特定の環境下で共通して働くネットワークを抽出でき，既知の知見 [10, 11] との高い一致を見ることができた．

5 まとめ

本研究では，新たなグラフモデルを提案し，共通アイテムセットを持つ部分ネットワークの抽出を行うアルゴリズム COPINE を考案した．COPINE は，ネットワークの大きさに伴って指数関数的に増加する部分ネットワークの探索を行うため，深さ優先探索アイテムセット木と探索済みパターンテーブルというデータ構造を持ち，これにより，探索空間を大幅に削減した．疑似データを用いて，COPINE による探索の有用性を示した．また，酵母のたんぱく質たんぱく質相互作用ネットワークと様々なストレス環境下により得られた実データを用いて，COPINE による解析を行い，特定の環境下で共通して働く部分ネットワークを抽出することができた．今後，COPINE を用いて，ソーシャルネットワークなどの様々なネットワークの解析を行い，COPINE の有用性のさらなる評価を行いたい．

参考文献

[1] M. E. J. Newman. *The structure and function of complex networks*. SIAM Review, 45:167, 2003.

[2] A. Inokuchi, T. Washio, and H. Motoda. *An apriori-based algorithm for mining frequent substructures from graph data*. PKDD '00, 2000.

[3] X. Yan and J. Han. *gSpan: Graph-Based Substructure Pattern Mining*. ICDM '02, 721, 2002.

[4] M. J. Rattigan, M. Maier, and D. Jensen. *Graph clustering with network structure indices*. ICML '07, 783-790, 2007.

[5] M. Shiga, I. Takigawa, and H. Mamitsuka. *A spectral clustering approach to optimally combining numerical vectors with a modular network*. KDD '07, 647-656, 2007.

[6] T. Ito, et al. *A comprehensive two-hybrid analysis to explore the yeast protein interactome*. Proc. Natl.Acad. Sci. vol. 98, 4569-4574, 2001.

[7] P. Uetz, et al. *A comprehensive analysis of protein-protein interactions in Saccharomyces cervisiae*. Nature, vol. 403, no. 6770, 623-627, 2000.

[8] N. J. Krogan, et al. *Global landscape of protein complexes in the yeast Saccharomyces cerevisiae*. Nature, vol. 440, 637-643, 2006.

[9] A. P. Gasch, et al. *Genomic expression programs in the response of yeast cells to environmental changes*. Mol. Biol. Cell, vol. 11, no. 12, 4241-4257, 2000.

[10] M. Ashburner, et al. *Gene ontology: tool for the unification of biology. the gene ontology consortium*. Nat Genet, vol. 25, no. 1, 25-29, May 2000.

[11] M. Kanehisa, and S. Goto. *KEGG: Kyoto Encyclopedia of Genes and Genomes*. Nucleic Acids Res, 28: 27-30, 2000.