

COPINE: 共通パターンを持つネットワーク抽出

関 美緒[†] 瀬々 潤[†]

[†] お茶の水女子大学大学院人間文化創成科学研究科 〒112-8610 東京都文京区大塚 2-1-1

E-mail: [†]{seki,sesejun}@sel.is.ocha.ac.jp

あらまし ノードにアイテムの集合を持つネットワークを考える．このネットワークは，購買した商品を持つ友人関係ネットワークや，活性化条件を有する遺伝子ネットワークが表せ，多くの応用が考えられる．このようなネットワークにおいて共通のアイテム集合を持つノードから成る部分ネットワークは，商品の浸透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ていると考えられる．本研究では，このような部分ネットワークを効率よく抽出するアルゴリズム COPINE の開発を行った．本論文では，効率よいグラフ探索手法を導入し，そのスケーラビリティの高さを示す．更に，酵母の遺伝子ネットワークに適用することで，既知の知見と高い一致を示す部分ネットワークが発見でき，本手法が有用であることを示す．

キーワード グラフマイニング，共通パターン

COPINE: Mining Networks Sharing Common Patterns

Mio SEKI[†] and Jun SESE[†]

[†] Graduate School of Humanities and Sciences, Ochanomizu University Otsuka 2-1-1, Bunkyo-ku, Tokyo, 112-8610 Japan

E-mail: [†]{seki,sesejun}@sel.is.ocha.ac.jp

Abstract We specifically examine an undirected network where nodes have a set of items, such as a social network in which the nodes are associated with each person's bought products or a gene network in which the nodes are related to each gene's activated conditions. For a given network, the subnetworks whose vertices contain common patterns have various applications. In order to extract such subnetworks, we introduce an algorithm called COPINE. COPINE can extract subnetworks from network having ten thousand nodes and two hundred thousand edges. COPINE can also discover useful biological networks from protein-protein interaction network in yeast.

Key words Graph mining, common patterns

1. はじめに

近年，データ収集技術と記憶装置の大幅な進歩により，情報収集が進み，大規模なデータが蓄積されてきている．中でも，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など幅広い分野において，グラフとして表現されたデータが扱われる場面が増え，グラフマイニングの研究が盛んになっている [1]．グラフ構造はグラフだけでなく，他のデータと結びついていることも多い．ソーシャルネットワークを考えた場合，友達のつながりだけでなく，各会員が属するコミュニティや購入した商品などの付帯状況があることが考えられる．遺伝子ネットワークであれば，遺伝子のネットワークだけでなく，各遺伝子には活動時期があり，時期により利用されるネットワークが異なる．しかし，これらの異なった複数のデータを同時に扱うデータマイニング手法は限られている．そこで本

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，それぞれ買った商品がアイテムである友人関係を表すネットワークや，各ノードが遺伝子，それぞれの活性化する実験環境がアイテムであるたんぱく質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，大規模ネットワークには多くの部分ネットワークが存在し，抽出が困難である．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う．また，新たな2つの枝刈り手法を導入する．これにより探索空間を大幅に減らす，COmmon Pattern Itemset NEtwork mining (COPINE) を考案した．

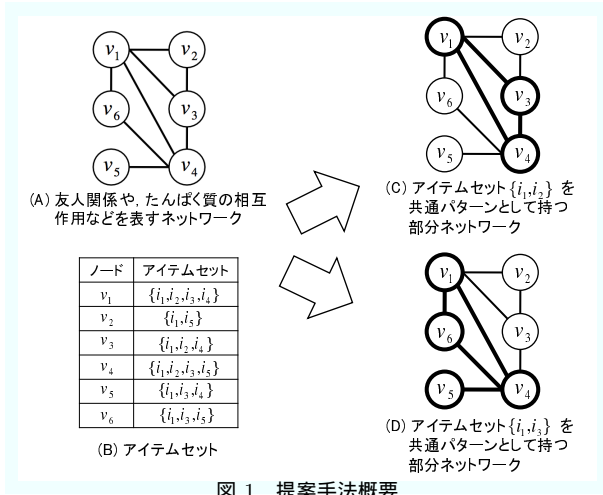


図 1 提案手法概要

2. 提案手法

2.1 概要

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つ。この商品や、活性化した条件を各ノードのアイテム、アイテムの集合を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を表す。これら、グラフデータとアイテムセットデータは、既に調べられていたり、まとめられている情報を利用する。本研究の目的は、これらのデータが与えられたとき、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で、図 1(D) の太線で表されている部分ネットワークはアイテムセット $\{i_1, i_3\}$ を持つ部分ネットワークである。また、ある部分ネットワーク G に G には含まれない連結しているどのノードを加えても、共通アイテムセットが変わらない場合、その部分ネットワーク G は閉じていると定義する。図 1(C)(D) の部分ネットワークは、他のどのノードを加えても、共通アイテムセットが変わってしまうので、閉じた部分ネットワークである。

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。よって、共通アイテムセットを持つ大きい部分ネットワークは、そのネットワークと共通アイテムセットの強い関連を示す。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。ここで、ネット

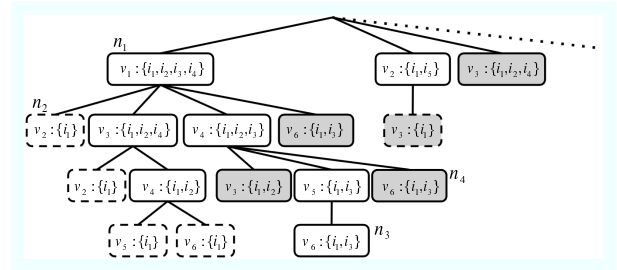


図 2 DFS アイテムセット木

ワークの大きさはネットワークのエッジ数とし、大きさ上位 N 個とは、閉じた部分ネットワークの大きさ上位 N 個を示す。

しかし、与えられたネットワークの大きさが大きくなるに従い、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と、探索済みパターンテーブルというデータ構造を保持している。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。

2.2 深さ優先探索アイテムセット木

COPINE は、ネットワークを深さ優先探索し、深さ優先探索アイテムセット木を生成する。深さ優先探索アイテムセット木は、木の各ノードがネットワーク中のノード 1 つと、そのノードを探索した時の共通アイテムセットから成る、深さ優先探索木である。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、単調減少である。深さ優先探索アイテムセット木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。

図 2 に、図 1(A)(B) のデータを用いて、COPINE で探索を進めていく様子を表した。ここでは共通アイテムセットの大きさの閾値を 2 とする。探索は、辞書順で 1 番小さいノード v_1 から始まり、深さ優先探索アイテムセット木に v_1 と、そのアイテムセット $\{i_1, i_2, i_3, i_4\}$ を持つノード n_1 が根の子どもとして追加される。深さ優先で v_2 に探索が進み、 n_1 の子どもに v_2 と、 v_1 と v_2 の共通アイテムセット $\{i_1\}$ を持つ n_2 が追加される。 n_2 における部分ネットワークの持つ共通アイテムセットは $\{i_1\}$ のみであり、共通アイテムセットの大きさの閾値 2 を満たさない。単調減少の性質により、 n_2 より先には閾値を満たす共通アイテムセットを持つ部分ネットワークは存在しないので、探索終了となる。同様に、図 2 の点線で囲まれたノードが探索された場合、探索終了となる。

2.3 探索済みパターンテーブル

先に述べた共通アイテムセット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。例えば、図 1(A) の v_6 は、 $\langle v_1, v_4, v_5, v_6 \rangle$ 、 $\langle v_1, v_4, v_6 \rangle$ 、 $\langle v_1, v_6 \rangle$ といった

ノード v	$P(v)$
v_1	$\{i_1, i_2, i_3, i_4\}$
v_3	$\{i_1, i_2, i_4\}$
v_4	$\{i_1, i_2\}, \{i_1, i_2, i_3\}$
v_5	$\{i_1, i_3\}$
v_6	$\{i_1, i_3\}$

表 1 探索済みパターンテーブル
(図 2 において, n_3 を探索した時)

異なるパスで繰り返し探索される．このような重複探索を除くため，探索済みパターンテーブルを導入し，次に述べる性質を利用して探索空間をさらに減らした．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 を深さ優先探索アイテムセット木のノードとし， n_1 は n_2 より前に生成されているとする． n_1 と n_2 両方がネットワーク中のノードを持ち， 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 の子孫に，閉じた共通アイテムセットを持つ部分ネットワークは存在しない．

表 1 は，図 2 の n_3 を探索した直後に保持している探索済みパターンテーブルである．例えば，表 1 の探索済みパターン $P(v_3)$ は， v_3 を探索した時の共通アイテムセットが $\{i_1, i_2, i_4\}$ であったことを示している．先に述べた性質により，あるノード v を探索したときの共通アイテムセットが，探索済みパターンテーブルに保持されている $P(v)$ 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．例えば，図 2 の n_3 において表 1 が生成された後，図 2 の n_4 へ探索が進む．このとき v_6 が探索され，共通アイテムセットは $\{i_1, i_3\}$ であるが， $P(v_6) = \{i_1, i_3\}$ が既に保持されているため，探索終了となる．同様にして，図 2 の灰色のノードが探索された際，探索終了となる．これにより列挙する部分ネットワークの増加を大幅に抑えることができた．

3. 実験

3.1 疑似データを用いた性能評価

COPINE の性能評価を行った結果を示す．COPINE は Java で実装し，Java1.5.0_16，CPU は AMD Opteron 2.2GHz，OS は Linux2.6 を利用した．COPINE の実行に際し，メモリは最大 1GB を利用した．

様々な大きさのネットワークにおける COPINE の性能を評価するために，疑似データを生成し，利用した．表 2 に疑似データと COPINE 実行の際に用いるパラメーターと，それぞれの初期値を示す．以下の手順で疑似データを生成した．

- (1) $|V|$ 個のノードを生成する．
- (2) ランダムに選んだ 2 つのノード間にエッジを張る．これを $|E|$ 回繰り返す．
- (3) 答えとなる最大の部分ネットワークに持たせる $|I|$ 個のアイテムをランダムに生成する．
- (4) (3) で生成したアイテムを，(1) で生成したノードからランダムに選んだ 1 つのノード v に持たせる．
- (5) v と隣接しているノードからランダムに選んだ 1 つのノード v' にも (3) で生成したアイテムを持たせる． v' を v として，(3) で生成したアイテムを持つノードが $|L|$ 個になるまで繰り返す．または，連結しているノード

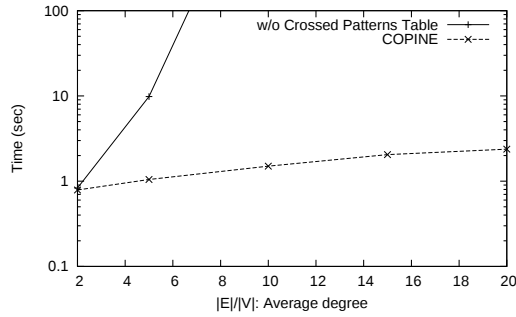
記号	定義	初期値
$ V $	ノード数	1,000
$ E $	エッジ数	5,000
N	アイテム数	100
$ T $	各ノードの平均アイテム数	20
$ I $	答えとなる最大部分ネットワークが持つアイテム数	10
$ L $	答えとなる最大部分ネットワークの大きさ	20
θ	共通アイテムセット閾値	$ I $

ド全てがこのアイテムを持ったら終了．

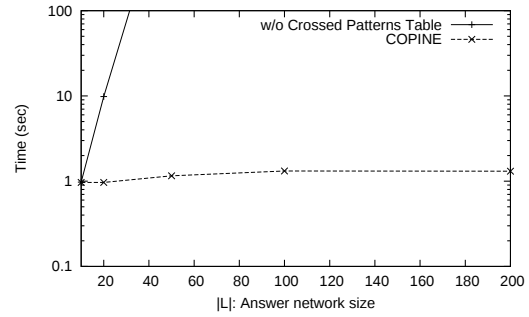
- (6) ネットワーク中の全てのノードに，ランダムにアイテムを持たせる．各ノードに持たせるアイテム数は，平均 $|T|$ のポアソン分布によって求める．

COPINE に導入した探索済みパターンテーブルには，テーブルのスキャンやメモリの使用量などのコストがかかる．そこで，探索済みパターンテーブルを用いた枝刈りの有効性を調べるため，探索済みパターンテーブル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を行った．図 3 に結果を示す．探索済みパターンテーブルによる枝刈りを導入しない場合と COPINE を利用した場合を，それぞれ “w/o Crossed Patterns Table”，“COPINE” と表している．縦軸が実行時間で，対数目盛を用いて表している．図 3(A) は，ネットワーク中のノードの平均度数を変化させ，ネットワークの複雑さに対する探索済みパターンテーブルの効果を調べた結果である．横軸が平均度数を表す．平均度数が増え，共通アイテムセットを持つ部分ネットワークも増加する．そのため，探索済みパターンテーブルを用いない場合，実行時間は平均度数の増加に従って指数関数的に長くなるが，COPINE の場合は抑えられており，ネットワークが複雑になっても一定した性能を保つことがわかる．図 3(B) は，抽出する共通アイテムセットを持つ部分ネットワークの大きさによる実行時間を計測したものである．横軸は，答えとなる第 1 位の部分ネットワークの大きさを表す．大きい部分ネットワークの探索は，多くの重複探索を含む．これらの結果は，COPINE が探索済みパターンテーブルを用いない場合より大幅に実行時間を速くすることを示している．

次に，COPINE のスケーラビリティを調べる実験を行った．図 4(A) は，ネットワークの大きさによる COPINE の実行時間を示している．縦軸が実行時間，横軸がネットワークの平均度数を表す．平均度数の増加に伴い，部分ネットワークの数は指数関数的に増加するが，実行時間の増加は緩やかである．この結果より，COPINE は 1 万ノード，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．図 4(B) は，答えとなる第 1 位の共通アイテムセットを持つ部分ネットワークの大きさによる COPINE の実行時間を示している．縦軸は実行時間，横軸は，答えとなる部分ネットワークの大きさを表す．答えとなる部分ネットワークの大きさの増加に伴い，その部分ネットワークの数も探索にかかるコストも指数関数的に増加すると考えられる．しかし COPINE では，実行時間の増加がほぼ抑えられている

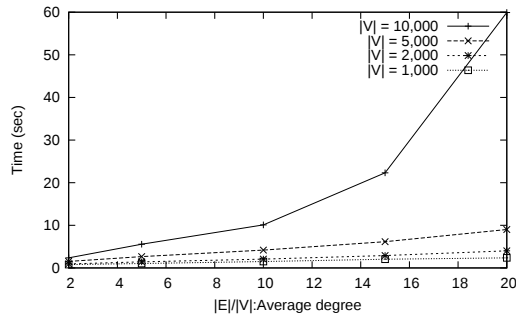


(A) ネットワークの大きさ

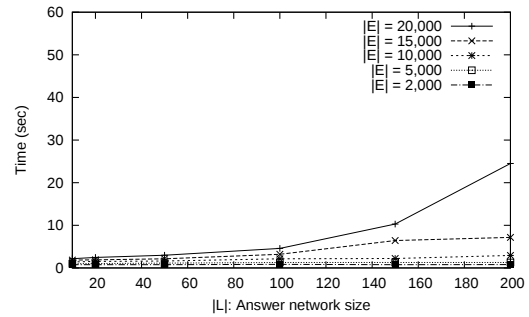


(B) 抽出部分ネットワークの大きさ

図 3 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) 抽出する部分ネットワークの大きさ

図 4 COPINE のスケーラビリティ

ことが図 4(B) からわかる。

3.2 実データを用いた実験

重複のない部分ネットワークの抽出

2.2, 2.3 節で述べた新たな枝刈り手法を用いて列挙した、共通アイテムセットを持つ部分ネットワークには、それら部分ネットワークに含まれるエッジや共通アイテムセットの多くが重複しているものもある。そこでエッジや共通アイテムセットの重複が少ない部分ネットワークを抽出するため、重複を判断する指標 $overlap(G_1, G_2)$ を次のように定義し、導入した。

$$\frac{|E(G_1) \cap E(G_2)|}{\min\{|E(G_1)|, |E(G_2)|\}} \times \frac{|I(G_1) \cap I(G_2)|}{\min\{|I(G_1)|, |I(G_2)|\}},$$

ここで、 G_1 と G_2 は共通アイテムセットを持つ部分ネットワーク、 $|E(G_1)|$ と $|I(G_1)|$ は G_1 のエッジ数と共通アイテムセットの大きさを表す。 $overlap(G_1, G_2)$ は 0 から 1 の値を取り、 G_1 , G_2 間に共通のエッジやアイテムが多い時、1 に近い値となる。実データを用いた実験では、重複を判断する閾値 γ を設定し、 $overlap(G_1, G_2) > \gamma$ となる、上位の部分ネットワークのみを抽出する。

結果

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 [2-4] 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 [5] を、遺伝子発現量 1.0 以上の実験環境を、その遺伝子が働く環境と考え、アイテムとして利用した。 $\theta = 3$ として COPINE を実行し、重複を判断する閾値 $\gamma = 0.1$ として重複の少ない部分ネットワークを抽出した。図 5 は実験結果を可視化したものである。中心上側は実験に用いたネットワーク全体を表し、抽出された

部分ネットワークのうち、上位 5 つを拡大し、それぞれをまとめた。このように、実データにおいても COPINE は共通アイテムセットを持つ部分ネットワークを抽出することができた。また、折れ線グラフは各部分ネットワーク中の全遺伝子の発現量平均値と標準偏差を表し、それぞれの順位はネットワークの順位と対応している。グラフの横軸は、173 種類の実験環境のうち 12 種類のみを示した。上位 5 つの部分ネットワークそれぞれの共通アイテムセットとなっている実験環境 12 種類である。この結果より、特定の環境下で共通して働いているネットワークを抽出できたことがわかる。

また、結果の精度評価のために、図 5 の上位 5 つの部分ネットワークについて各ネットワークに含まれる遺伝子と、遺伝子の機能や情報伝達経路のデータベース [6, 7] において定義されている各カテゴリーに属する遺伝子との一致を二項検定を用いて調べた。その結果、二項検定により得られた p 値が $9.07 \times 10^{-20} \sim 6.53 \times 10^{-3}$ となり、COPINE の結果と既知の知見との高い一致を見ることができた。

4. 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 [8, 9] やグラフのクラスタリング [10]、制約付きクラスタリング [11] 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、頻出しているもの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もの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

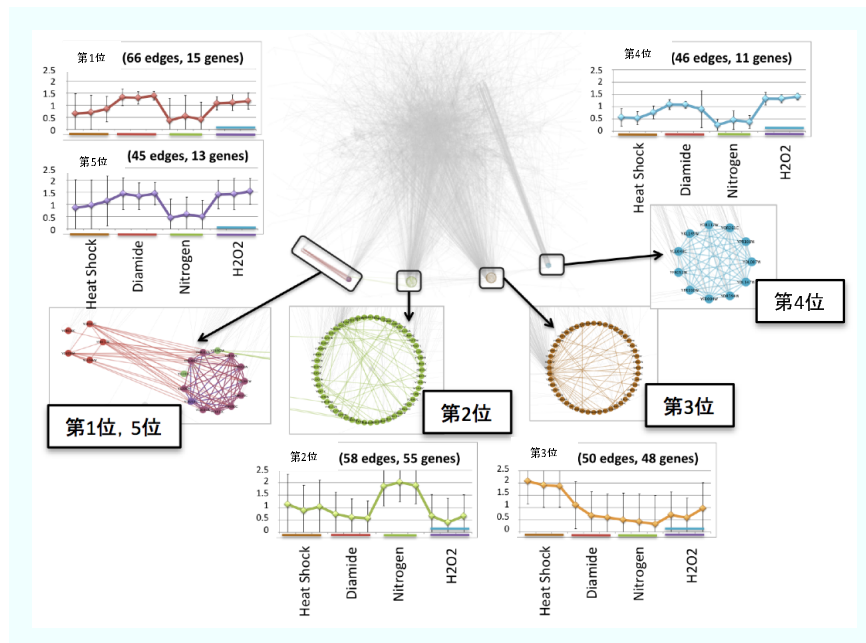


図5 実験結果

ラフのクラスタリングは、口コミや、生物学的なパスウェイにおいて、このような情報は低い次数を持つノードを通し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特徴であ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。

アイテムセットを利用したマイニングでは、頻出するアイテムセットを抽出する Apriori アルゴリズム [12] や、FP-tree を用いた手法 [13] が考えられている。しかし、これらの手法によって抽出された頻出するアイテムセットを持つノードがグラフ上では離れてしまうことが起こりうるため、グラフとアイテムセットを関連付けた解析は難しい。また、実行にはサポート値など、アイテムセットの出現度合に関する閾値を設定する必要がある。しかし、本研究において抽出を目的としている部分ネットワークを持つ共通アイテムセットは、必ずしも高いサポート値を満たすアイテムセットであるとは限らず、そのため低いサポート値を設定すると意味の無いアイテムセットも多く生成されてしまうため、マイニングが困難となる。

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。

文 献

- [1] M. E. J. Newman. *The structure and function of complex networks*. SIAM Review, 45:167, 2003.
- [2] T. Ito, et al. *A comprehensive two-hybrid analysis to explore the yeast protein interactome*. Proc. Natl. Acad. Sci. vol. 98, 4569-4574, 2001.
- [3] P. Uetz, et al. *A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae*. Nature, vol. 403, no. 6770, 623-627, 2000.
- [4] N. J. Krogan, et al. *Global landscape of protein complexes in the yeast Saccharomyces cerevisiae*. Nature, vol. 440, 637-643, 2006.
- [5] 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.
- [6] 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.
- [7] M. Kanehisa, and S. Goto. *KEGG: Kyoto Encyclopedia of Genes and Genomes*. Nucleic Acids Res, 28: 27-30, 2000.
- [8] A. Inokuchi, T. Washio, and H. Motoda. *An apriori-based algorithm for mining frequent substructures from graph data*. PKDD '00, 2000.
- [9] X. Yan and J. Han. *gSpan: Graph-Based Substructure Pattern Mining*. ICDM '02, 721, 2002.
- [10] M. J. Rattigan, M. Maier, and D. Jensen. *Graph clustering with network structure indices*. ICML '07, 783-790, 2007.
- [11] 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.
- [12] R. Agrawal and R. Srikant. *First algorithms for mining association rules*. Proc. VLDB '94, 487-499, 1994.
- [13] J. Han, J. Pei, and Y. Yin. *Mining frequent patterns without candidate generation*. Proc. SIGMOD '00, 1-12, 2000.