アイテム集合を付与したグラフからの頻出グラフ発見

福崎　睦美† 　関　美緒† 　鹿島　久嗣†† 　瀬々　潤†

† お茶の水女子大学理学部情報科学科　〒 112-8610 東京都文京区大塚 2–1–1
†† IBM 東京基礎研究所　〒 242-8502 神奈川県大和市下鶴間 1623-14
E-mail: †{fukuzaki,seki}@sel.is.ocha.ac.jp, sesejun@is.ocha.ac.jp ††hkashima@jp.ibm.com

あらまし　アイテム集合マイニングとグラフマイニングは双方が独立して研究されてきた．そこで，本研究では各頂点にアイテム集合を持つ重み無しグラフという新たなデータ構造を導入する．このデータから，連結したサブネットワーク集合とそれらの頂点に共通するアイテム集合の組み合せを列挙することにより，薬の副作用を伴う組み合せを求める．この問題を解くため，本研究では新たなアルゴリズム ROBIN を導入する．ROBIN は，まずアイテム集合を共有する連結した部分グラフの列挙を行い，その後それらの組み合せを列挙する．本論文では実験結果によって，ROBIN は 100 万エッジを超えるグラフからも問題を解くことが可能であり，酵母の実データから生物学的に有意な部分グラフの列挙ができることを示す．

キーワード　グラフ，アイテム集合，遺伝子発現量

Finding Itemset-Sharing Patterns in a Large Itemset-Associated Graph

Mutsumi FUKUZAKI†, Mio SEKI†, Hisashi KASHIMA††, and Jun SESE†

† Dept. of Computer Science, Ochanomizu Univ.　2–1–1 Otsuka, Bunkyo-ku, Tokyo, 112–8610 Japan
†† Tokyo Research Laboratory, IBM Research　1623–14 Shimotsuruma, Yamato-shi, Kanagawa, 242–8502 Japan
E-mail: †{fukuzaki,seki}@sel.is.ocha.ac.jp, sesejun@is.ocha.ac.jp ††hkashima@jp.ibm.com

Abstract　Both itemset mining and graph mining have been studied independently. Here, we introduce a novel data structure, which is an unweighted graph whose vertices contain itemsets. From the data, we enumerate combinations of connected subnetworks whose vertices share itemsets because the combinations imply side effects of drugs. To solve the problem, we introduce a novel algorithm ROBIN, which first enumerates connected graphs sharing itemsets, and then enumerates the combinations of the graphs. Our experimental result shows that ROBIN can solve the problem from the graph having more than one million edges, and enumerate biologically meaning subgraphs from real yeast dataset.

Key words　Graph, itemset, gene expression

1. はじめに

データマイニングで頻繁に研究される問題として，頻出アイテム集合の列挙がある．また，近年この頻出アイテム集合マイニングの手法を応用した，頻出グラフマイニングが発展してきている．しかし，アイテム集合とグラフは独立して解析されている．本研究では，このアイテム集合の解析とグラフの解析を融合させた解析を行う．

本研究では，重み無しグラフの頂点にアイテム集合が付与してある構造を考える．このようなグラフ構造は，実データに頻繁に現れる．たとえば，創薬情報であれば遺伝子間の関係を示す遺伝子ネットワーク（ノードが遺伝子，辺が遺伝子間関係）に対し，各遺伝子がどの薬に反応するかを付与したグラフである．他にも，SNS であれば頂点に参加者，辺を友人関係とし，参加者が購入した商品を付与したものである．

本論文では，この様なグラフからアイテムを共有していながら全体としては非連結である複数の部分グラフを列挙する手法，RelatiOn Between Items and Network (ROBIN) を提案する．この手法により，遺伝子ネットワークからは，既存の知識では連結していないが，実は協調して働く可能性のある遺伝子ネットワーク，つまり，副作用の可能性を示す遺伝子ネットワークを調査することが可能である．

ROBIN は以下のような手順で構成される．(1) 重みなし無向グラフに変換した遺伝子ネットワーク上のノードに活性化条

件をアイテム集合として関連づける．(2) 共通したアイテムを持つノードによって構成される連結した部分グラフを列挙する．(3)(2) で列挙した部分グラフを組み合せ，部分グラフ間で共通するアイテムを列挙する．(3) で得られた部分グラフ集合は，非連結であっても共通した条件で活性化する．つまり，副作用を起こすと考えられるのである．本研究では，(2) で COPINE [5] を応用することで部分グラフとその共通アイテムを高速に列挙し，(3) における効率よい共通アイテム集合を持つ部分グラフ集合の列挙方法を提案する．

1.1 解 析 例

本手法の解析結果例を図 1 に示す．図 1(a) は無向グラフであり，(b) はその各頂点のアイテム集合を表す．ROBIN では結果のネットワークの信頼性を高めるため，共通アイテム数が閾値 θ_I，大きさが閾値 θ_S を満たす部分グラフのみ列挙する．図 1 を，$\theta_I = 2$，$\theta_S = 2$ で解析した一例が図 2 である．図 2 において太線で示した二つの部分グラフは，左の部分グラフで全ノードに $\{i_1, i_2, i_3\}$ が共通しており，右は $\{i_2, i_3\}$ が共通している．これらは，$\{i_2, i_3\}$ という共通の条件で副作用を起こす可能性のあるネットワークを表すと言える．このように二つの独立した部分グラフの持つアイテム集合が完全一致していない場合であっても，それらの持つアイテム集合の共通部分の大きさが閾値を満たすならば副作用を起こす可能性があるといえ，列挙は容易ではない．

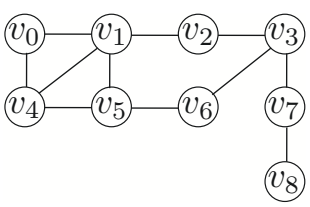
1.2 関 連 研 究

グラフデータベース解析手法として主要なものには，頻出サブグラフ列挙手法がある [2], [3]．しかし，この手法はグラフの構造のみについて解析するものであり，遺伝子発現量と統合した解析を行うことはできない．また，頻出アイテム集合マイニング [1], [4] を応用することも考えられるが，そのためには頻出するアイテム集合の全てのパターンについてネットワークでの連結を調べる必要があり，効率的でない．COPINE [5] は，二つのデータベースを統合した解析を可能にするが，連結グラフのみを考慮しているため，本研究の求める複数の部分グラフ間の共通アイテムを発見することはできない．本研究では COPINE を応用し，共通のアイテム集合を持つ独立した複数のネットワークを列挙する手法を提案する．

2. 問 題 設 定

本研究で使用する定義を以下に示す．G は非連結でラベルや重みのない無向グラフであり，その各頂点はアイテム集合を持つ．このようなグラフ G を itemset-associated graph(IA graph) と定義する．$V(G), E(G), \mathcal{I}(G)$ および $I(v)$ は，それぞれ G の頂点集合，G の辺集合, G に含まれる全アイテムの集合，$v \in V(G)$ のアイテム集合を表す．$|E(G)|$ は, G の大きさを表す．また，G の部分グラフの中でも重要なのはより大きなグラフであると言えるので，次の Itemset-Sharing Subgraph (ISS) を，G の興味深い部分グラフであるとし，その共通アイテムとともに定義を示す．

定義 1：(ISS)　G' を IA graph G の部分グラフとする．$I(G')$ を $I(G') = \bigcap_{v \in V(G')} I(v)$ と定義したとき，$I(G')$ をグラフ G' の共通アイテムとし，$I(G') \neq \phi$ かつ，G' に隣接する全てのノード v' について $I(v') \not\supseteq I(G')$ のとき，G' を $I(G')$ についての ISS と定義する．■



(a) 遺伝子ネットワークを表すグラフ

v_0	$\{i_1, i_2, i_3\}$
v_1	$\{i_1, i_2\}$
v_2	$\{i_1, i_2\}$
v_3	$\{i_2, i_3, i_4\}$
v_4	$\{i_1, i_2, i_3\}$
v_5	$\{i_1, i_2, i_3, i_4\}$
v_6	$\{i_2, i_4\}$
v_7	$\{i_1, i_2, i_3\}$
v_8	$\{i_2, i_3\}$

(b) アイテム集合

図 1　IA graph の例

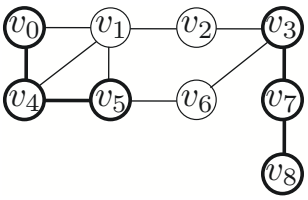


図 2　ISS 集合の例:太枠の部分グラフは共通アイテム $\{i_2, i_3\}$ を持つ

この ISS を用い，本研究で求める部分グラフ集合を定義する．

定義 2：(ISS 集合列挙問題) $\mathcal{G} = \{G_1, \ldots, G_n\}$ を ISS の集合とする．ここで G_i は ISS である．θ_S をユーザ定義の値とすると，(1) $V(G_i) \cap V(G_j) = \phi$ (ただし i と j は異なる), (2) $\mathcal{G}$ に隣接する頂点 v に関し $I(v) \not\supseteq I(\mathcal{G})$，(3) $|G_i| \geqq \theta_S$，(4) $\mathcal{G}$ 内の ISS 以外に θ_S 以上の大きさの $I(\mathcal{G})$ に関する ISS は存在しない，を全て満たすグラフを ISS 集合と呼ぶ．本研究ではこの中から，θ_F 個以上の部分グラフを持ち，θ_I 以上の大きさのアイテム集合に関連づけられた ISS 集合を列挙する．この問題を，ISS 集合列挙問題と呼ぶ．■

3. 提 案 手 法

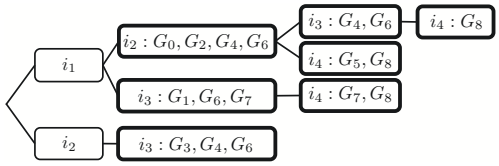
ROBIN では ISS 集合列挙問題を解くため，閾値を満たす ISS を COPINE によって列挙し，その中から共通するアイテムを θ_I 個以上持つ ISS 集合を発見する．このとき ISS の全通りの組み合せを調べる必要がある．その方法の一つとして ISS の深さ優先による組み合せがあるが，ISS 数が多い場合，組み合せを探索する木が深くなる．そこで本手法では，ISS を関連するアイテム集合でグループ化することで，集合の組み合わせを可能にし，高速化する．

定義 3：(陽な ISS 集合と陰な ISS 集合)　ISS に関連したアイテム集合を $\mathcal{I}$ とすると ISS 集合 $\mathcal{G}$ に関し，$I(\mathcal{G}) \in \mathcal{I}$ の時，$\mathcal{G}$ を陽な ISS 集合，それ以外を陰な ISS 集合と呼ぶ．■

ROBIN では列挙した ISS をアイテム集合に着目してグループ化することで陽な ISS 集合を求める．次に，陽な ISS 集合を組み合わせることで陰な ISS 集合を求める．ISS のグループ化に

G_0	$\{i_1, i_2\}$
G_1	$\{i_1, i_3\}$
G_2	$\{i_1, i_2\}$
G_3	$\{i_2, i_3\}$
G_4	$\{i_1, i_2, i_3\}$
G_5	$\{i_1, i_2, i_4\}$
G_6	$\{i_1, i_2, i_3\}$
G_7	$\{i_1, i_3, i_4\}$
G_8	$\{i_1, i_2, i_3, i_4\}$

(a) ISS 集合



(b) (a) から作成される ISS プレフィックス木

図 3　ISS プレフィックス木:太枠のノードが陽な ISS 集合を表す

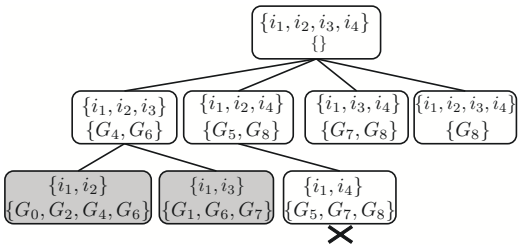


図 4　図 3 から作成される ISS 木

はアイテム集合のプレフィックス木を利用する．また，アイテムは添字通りの順序を持つ．

定義 4：(ISS プレフィックス木)　T_P を木，n をそのノードとする．この木は次の性質を持つ．n はアイテム i_n と ISS 集合 $\mathcal{G}(n)$ を持つ．$n_i, n_j \in T_P$ のノードとし，n_j が n_i の子ノードの時，$i_{n_j} < i_{n_i}$ を満たす．ISS 上の根からあるノードまでのパスは，パス上のノードに関連したアイテムの集合からなるアイテム集合を表す．ISS の共有するアイテム集合は全てこの木上の根からのパスで表される．この木を ISS プレフィックス木と呼ぶ．■

これにより，完全一致またはサブセットになるようなアイテム集合に関連した陽な ISS 集合へ高速にアクセスできるようにする．

ISS プレフィックス木の作成のアルゴリズムを以下に示す．

アルゴリズム 1(ISS プレフィックス木作成)
入力:$\mathcal{G} = \{G_1, ...G_n\}$($G_i$は ISS, $i < j$ のとき $|I(G_i))| \leqq |I(G_j))|$)
1.$T_P \leftarrow null$
2.for each $G_i \in \mathcal{G}$ do
3.　$I(G_i)$ を表すパスがなかったら，パスを作成
4.　$I(G_i)$ の部分集合をあらわす，存在する全てのノードに G_iを追加
5.end for
6.for each $n \in T_P$ do
7.　n 内の他の ISS に包含される全ての ISS を n から消去
8.end for
9.return T_P

ISS プレフィックス木の例を図 3 に示す．ISS とそれに関連するアイテム集合は，図 3(a) によって表されている．ただし $V(G_4) \subset V(G_1)$, $V(G_5) \subset V(G_0)$, $V(G_8) \subset V(G_6)$ という包含関係があるとする．全 ISS 集合に関連するアイテム集合は，ISS 木によって列挙される．

定義 5：(ISS 木)　T_I を木とし，T_I の各ノード n は，アイテム集合 $I(n)$ と ISS 集合 $\mathcal{G}(n)$ を持つ．$n_i, n_j \in T_I$ について n_j が n_i の子ノードの時，$I(n_j) \subset I(n_i)$ を満たす．この性質を満たす木を ISS 木と呼ぶ．■
ISS 木の作成アルゴリズムを以下に示す

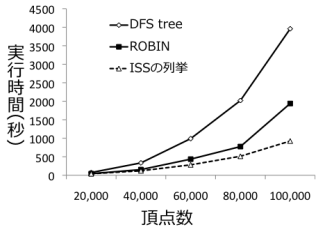
アルゴリズム 2(ISS 木生成)
入力:$\mathcal{G} = \{\mathcal{G}_1, ...\mathcal{G}_n\}$($\mathcal{G}_i$は 陽な ISS 集合)
1. $T_I \leftarrow null$
2.// root ノード n の生成
3. $I(n) \leftarrow \mathcal{I}(G)$; $\mathcal{G}(n) \leftarrow \{\}$; $\mathcal{H}(n) \leftarrow \mathcal{G}$
4.// n の子の再帰的な生成
5. ISSTreeGrowth(n)
6. return T_I

$ISSTreeGrowth$
入力:ISS 木ノード n
7. for each $\mathcal{G}_i \in \mathcal{H}(n)$ do
8.　$n' : ISS$ 木ノードとする
9.　$I(n') \leftarrow I(n) \cap I(\mathcal{G}_i)$
10.　$\mathcal{G}' \leftarrow \{\mathcal{G}_j | \mathcal{G}_j \in \mathcal{H}(n), I(\mathcal{G}_j) \supseteq I(n')\}$
11.　$\mathcal{G}(n') \leftarrow \mathcal{G}(n) + \mathcal{G}'$
12.　$\mathcal{H}(n') \leftarrow \mathcal{H}(n) - \mathcal{G}'$
13.　n'を n の子ノードにする
14.　ISSTreeGrowth(n')
15.end for

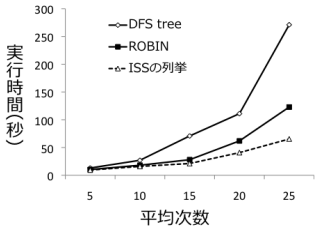
以上のように，ROBIN では新たにノード n' が作成されるとき，$I(n')$ に関連する全ての ISS を $\mathcal{G}(n')$ に加える．よって，親ノードを n，子ノードを n' としたとき，$|I(n)| > |I(n')|$ が成り立つ．よって，ISS 木は以下のような性質を満たす．

性質:(アルゴリズム 2)　ISS 木のノードのうち，以下の条件を満たすノードは枝刈りが可能である．
(1)$I(n')$ が既出のアイテム集合である．
(2)$|I(n')| < \theta_I$.
(3)$|I(n)| = \theta_I$となるノード n を親に持つ．

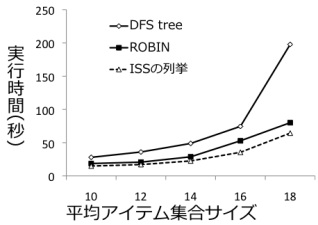
証明:(性質:アルゴリズム 2)　(1)添字をノードの生成順序とし，ノード $n_i, n_j (i < j)$ について $I(n_i) = I(n_j)$ であるとき，$\mathcal{G}_j = \{n_j$とその子孫で生成されたISS の集合 $\}$, $\mathcal{G}_i = \{$ 全て n_i とその子孫で生成されたISS の集合 $\}$ として，$\forall \mathcal{G}'_k \in \mathcal{G}_j$について $I(\mathcal{G}'_k) = I(\mathcal{G}'_l)$ かつ $\mathcal{G}'_k \subset \mathcal{G}'_l$を満たす $\mathcal{G}'_l \in \mathcal{G}_i$が存在する．このとき，ISS 集合の定義により $\mathcal{G}'_k$ は解とならない．よって，既出のアイテム集合が関連づけられた n_j とその子ノードは枝刈りが可能．
(2)(3)親ノードを n，その子ノードを n'としたとき，$|I(n')| < |I(n)|$ が成り立つことから，(2)(3) に該当するノード及びその子ノードは閾値を満たさないため，枝刈りが可能．■



(a) 頂点数を変化し，グラフを DFS で組み合せた場合との実行時間を比較



(b) グラフの平均次数を変化させた場合の提案手法での実行時間の変化



(c) 平均アイテム集合サイズを変化させた場合の提案手法での実行時間の変化

図 5　疑似データによる実行結果

図 3 をもとに ISS 木を作成した例を図 4 に示す．各ノードには，アイテム集合と ISS 集合を示した．$\theta_I = 2$ の時，灰色のノードは既出のアイテム集合であり，$\times$ の子ノードはアイテム集合のサイズが閾値 θ_I を満たさないため，枝刈りされる．

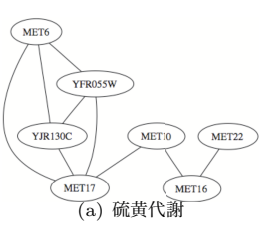
4. 実 行 結 果

4.1 疑似データによる実験

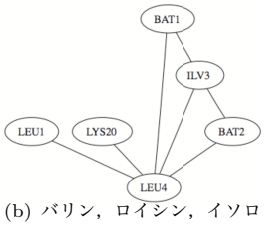
ROBIN の性能の評価のため，疑似データを作成し実験を行った．特に指定しない場合，頂点数 $|V(G)| = 15000$，平均次数 $|D| = 10$，アイテムの種類 $|\mathcal{I}(G)| = 100$，平均アイテム集合サイズ $A = 10$，解のアイテム集合サイズ $|I| = 10$，解の ISS サイズ $S = 7$，解の ISS 集合サイズ $F = 5$，解のパターン数 $P = 10$，閾値 $\theta_I = |I| - 1, \theta_S = S - 1$ とする．疑似データは以下の手順で作成される．

1. 大きさが $|V(G)|$ となる頂点集合 $V(G)$ を作成し，$|V(G)| \times |D|$ 回ランダムに二つの頂点間に辺をはる．
2. 以下を P 回繰り返す
   (1) 大きさ $|I|$ のアイテム集合 $I' \in \mathcal{I}(G)$ を作成
   (2) 大きさ S の I' に関連する ISS を F 個作成
3. 頻出するアイテム集合 $I_f \subset \mathcal{I}(G)$ を $|I_f| = |I| \times 1.7$ となるように作成
4. 以下を $|V(G)|/30$ 回繰り返す
   (1) 大きさ $|I|$ のアイテム集合 $I'' \subset I_f$ を作成
   (2) 大きさ S の I'' に関する ISS を作成
5. 各頂点の平均アイテム数が A になるまで，ランダムに頂点にアイテムを配置

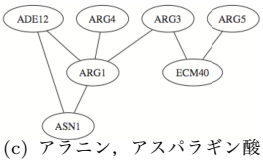
実行結果は図 5 に示した．比較のため，ISS の単純な深さ優先探索による組み合せの実行時間を計測し，さらに組み合せの実行時間を明確にするため ISS の列挙にかかった実行時間も計測，グラフ中にそれぞれ DFS tree，ISS として表示した．図 5(a) は，疑似データ上で頂点数を変化させた場合の実行時間である．



(a) 硫黄代謝



(b) バリン，ロイシン，イソロイシン合成



(c) アラニン，アスパラギン酸代謝

図 6　酵母の実データ解析結果

頂点数が増加するほど列挙される ISS 数も増加するため，提案手法が有利になることがわかる．さらに，ROBIN で 10 万ノード，100 万エッジの大規模なデータを解析することも可能であったことが示されている．また，図 5(b)(c) にはそれぞれ，(b) グラフの平均次数を変化させたとき (c) 平均アイテム集合サイズを変化させたときの実行時間を示した．

4.2 実データによる実験

図 6 は，実データによる実験結果を表す．使用したデータはネットワークのエッジ数が 3,324 の酵母の遺伝子ネットワークと，各遺伝子の持つ活性化条件 (アイテム) 数の平均が 5.7 となっている実験データであり，閾値を $\theta_I = 3, \theta_S = 7$ とした．列挙された 3 つのネットワークは共通に {erg2，yhl029c，ERG11(tet promoter)} の条件において活性化している．また，これらのネットワークの機能を KEGG を用いて調べると，(a)，(b)，(c) の何れもアミノ酸代謝に関連する既知のパスウェイと高い相関があり，生物学的な知識と一致する結果が得られた．

5. まとめと今後の課題

本論文では，遺伝子ネットワークと遺伝子発現量をアイテム集合を付与したグラフにモデル化することで統合して解析し，アイテムを共有する連結部分グラフを求めた後，その部分グラフを効率よく組み合わせて探索する手法 ROBIN を提案した．今後は，SNS など遺伝子以外のネットワークへの応用や，実データの実行結果の生物学的な解釈を進めていきたい．

文　献

[1] J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. In SIGMOD ' 00, pages 1 - 12, 2000.
[2] M. Kuramochi and G. Karypis. Frequent subgraph discovery. In ICDM ' 01, pages 313 - 320, 2001.
[3] X. Yan and J. Han. gspan: Graph-based substructure pattern mining. In ICDM ' 02, page 721, 2002.
[4] J. Han, H. Cheng, D. Xin, and X. Yan. Frequent pattern mining: Current status and future directions. Data Mining and Knowledge Discovery, 14(1), 2007.
[5] M. Seki and J. Sese. Identification of active biological networks and common expression conditions. In BIBE ' 08, 2008.