アイテム集合付き部分グラフに対する事後処理法の提案

山田　恵† 瀬々　潤†

† お茶の水女子大学　〒 112–8610 東京都文京区大塚 2–1–1
E-mail: †yamada.megumi@sel.is.ocha.ac.jp, sesejun@is.ocha.ac.jp

あらまし　Web，遺伝子ネットワークやソーシャルネットワークなど，現代には大規模で多様なネットワークが存在する．これらは，頂点と辺からなるだけでなく，辺の重みや頂点固有の別データなど様々な付加要素を有し，構造の性質を考慮に入れたデータの解析から新たな知識発見の可能性が期待される．本論では，口コミや薬剤開発に利用価値のある，頂点にアイテム集合を付与したネットワークから抽出された，共通のアイテム集合を持つ部分グラフ (ISS) に着目する．大規模な IA グラフを対象に ISS 列挙を行うと，多数の重複の多い解を持つことがあり，結果の解釈が困難であった．本研究では，重度の重複は，本質的に同じ現象を示す結果が不意に別々の結果として現れた例と捉え，多少のミスが生じたとしても柔軟で簡潔な結果となるよう，列挙された大量の ISS から重複が多いものを併合することで解釈容易な ISS を抽出する．また，実際の生物学データに適用し，本手法の有効性を示す．

キーワード　データマイニング，グラフ，アイテム集合，遺伝子発現量

Post-processing for Itemset Sharing Subgraphs

Megumi YAMADA† and Jun SESE†

† Dept. of Computer Science, Ochanomizu Univ. 2–1–1, Otsuka, Bunkyo-ku, Tokyo, 112–8610 Japan
E-mail: †yamada.megumi@sel.is.ocha.ac.jp, sesejun@is.ocha.ac.jp

Abstract　Recently, various large and complex graphs exist such as the Web, gene networks and social networks. One of the common features of these graphs is that the nodes on the graphs have various attributes which are effective feature to analyze the graphs. This study use a graph whose vertices has a set of items and focus on a subgraph whose members share a set of items, called itemset-sharing subgraphs (ISSes) because they are related to viral marketings and drug design. ISS enumeration algorithms often produce enumerous and highly overlapped results, which make it difficut to our understanding of the network properties. Avoiding the difficutly, we here introduce post-processing method for the enumerated ISSes by aggregating the highly overlapped ISSes. The application of our method to real biological dataset shows effectiveness of our aggregation of ISSes.

Key words　Data mining, Graph, Itemset, Gene expression

1. はじめに

観測されるデータが膨大かつ複雑化するにつれて，複数のヘテロなデータが統合されたデータベースからの知識発見の必要性が増している．その最たる物として，今までデータマイニングで研究されてきたアイテム集合マイニング [1], [2] と Web やソーシャルネットワーク，更に遺伝子ネットワークなどに代表されるグラフ構造 [3] の両方の性質を持つ，アイテム集合付きグラフ (IA グラフ) がある．

IA グラフは，グラフ構造の頂点にラベルとしてアイテム集合が付与されているグラフである．図 1(C) に IA グラフの例を示す．

このグラフでは，図 1(A) で示す頂点が 8 個，辺が 8 本のグラフの各頂点に，図 1(B) のアイテム集合が付与されている．たとえば，頂点 1 は頂点 2,5 との間に辺を持ち，更に，A, B, X のアイテムを有している．

このような IA グラフは Web や遺伝子情報をマイニングする際に頻繁に現れる構造である．たとえば，頂点をユーザ，辺をユーザ間の友人関係とするグラフに対し，各ユーザが購入した商品をアイテムとして付与したものは IA グラフであるし，頂点を遺伝子，辺を遺伝子間の繋がりとして，各遺伝子が反応する薬剤をアイテムとして付与したものも IA グラフである．更に，論文の参照関係をグラフとし，論文のキーワードを付与したものも IA グラフである．このように IA グラフは現代の様々なデータに現れる構造でありながら，この構造の研究は限定的である [4]〜[6].

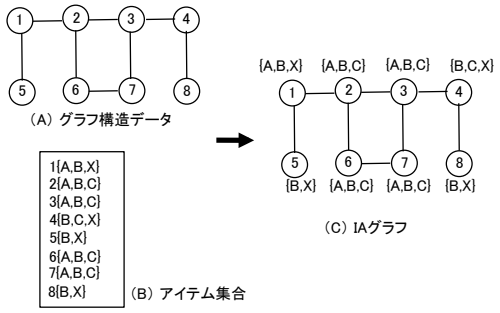

図 1　IA グラフの例

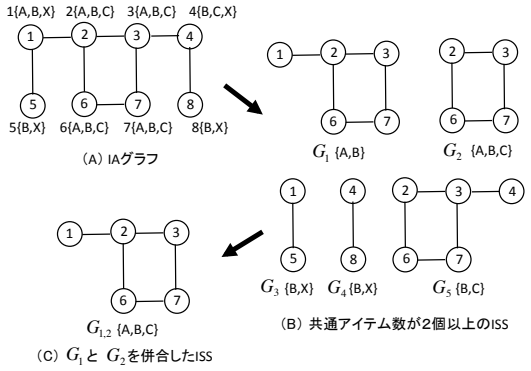

図 2　ISS 集合列挙，併合

既存の研究において IA グラフで重要視されているのは，共通アイテム集合を持つ部分グラフ (ISS) を抽出する問題である．たとえば，図 1(C) では，頂点 2,3,6,7 はグラフ上連結であり，かつ，これらの頂点は全て共通のアイテム集合 {A,B,C} を有するので ISS である．ISS に着目する理由は，IA グラフのデータ解析に有益ながら，ISS がアイテム集合マイニングでもグラフ構造マイニングでも抽出できない構造を抽出できるためである．たとえば，SNS の友達関係に商品購買履歴を付与した IA グラフを考えると，ISS が表すものは共通の商品群を購入していて，かつ，友人関係がある部分グラフである．このグラフは，言い換えると，友人間で話し合うことで商品購入が進んだ，つまり，口コミで広まった商品群とその友人関係と見ることができ，広告の効果測定に有益である．同様に遺伝子ネットワークに関して考えると，ISS は共通の薬剤に反応する遺伝子ネットワークを示しており，抽出した ISS から薬剤の組み合わせによる反応や，異種と思われていた薬剤が実は同じ遺伝子群に対して作用していることが発見できる可能性がある．

この有益な ISS を列挙するアルゴリズムとして COPINE [4] が開発されている．ISS はユーザが与えた部分グラフの大きさと，共通アイテム集合の大きさを満たす ISS を高速に列挙するアルゴリズムである．しかし，COPINE が列挙する部分グラフは，与えるグラフが複雑になると結果の量が膨大になる上，グラフ間の重複が多くなりがちで，どのグラフが本当に重要な部分グラフであるか判定が難しい側面が存在する．たとえば，図 1(C) の IA グラフから共通アイテム集合が 2 個以上の ISS を列挙した場合，図 2(B) に示した G_1 から G_5 のようになるが，この中には，頂点 1,2,3,6,7 からなる，共通アイテムが {A,B} の ISS, 頂点 2,3,4,6,7 からなる，共通アイテムが {B,C} の ISS があり，これらは，共通の頂点として 2,3,6,7 を有している．実際には，これらの間には共通するアイテムがあり，共通する頂点も多いため，単に異なる ISS と見るより，頂点 1,2,3,4,6,7 からなる，共通アイテム {A,B,C} の ISS のアイテムに含まれるノイズにより異なる ISS が生成されたと考えられる．

本研究では，頂点とアイテム集合に重複の多い ISS を併合することで，観測された ISS の背後にある，本当の ISS を求める．

2. 提 案 手 法

本章では列挙された大量の ISS から，グラフ間あるいはアイテム集合間に重複の多い ISS 集合を併合することで，本質的に重要な ISS を求める手法を提案する．

2.1 準　　備

グラフ G を非連結でラベルや重みのない無向グラフとする．$V(G)$，$E(G)$，$I(G)$ および $I(v)$ は，それぞれ G の頂点集合，辺集合，G 頂点の持つ全アイテムの集合，頂点 $v \in V(G)$ が持つアイテム集合とする．この時，グラフ G を Itemset-associated graph (IA グラフ) と呼ぶ．

G' を G の部分グラフとする．このとき，グラフ G' の持つ共通アイテム集合は，$I(G') = \bigcap_{v \in V(G')} I(v)$ と表せ，$I(G') \neq \phi$ かつ，G' に隣接する全てのノード v' について $I(G' \cup \{v'\}) \neq I(G')$ のとき，G' を Itemset-Sharing Subgraph (ISS) と定義する．この ISS を利用し，以下の ISS 列挙問題が定義できる．

［定義 1］ **ISS 列挙問題**: θ_I，θ_S をユーザ定義の値とする．ISS 列挙問題は，与えられた IA グラフ G から，共通アイテム集合の大きさが θ_I 以上，グラフの大きさが θ_S 以上の ISS を全て列挙する問題である．

COPINE [4] は，この ISS 列挙問題を全ての部分グラフを効率的に列挙しつつ解くアルゴリズムであり，100 万を超える辺を持つ大規模なグラフから ISS が列挙できる．一方で，ISS 列挙問題で求められる解は，解の数が多かったり，解の間にグラフの重複が多かったりするため，結果の解釈が困難である難点を有している．

2.2 ISS の重複

一般のグラフにおいて重複とは共通の頂点及び辺を有することであるが，IA グラフにおける重複は，共通の頂点や辺を有し，かつ，共通アイテム集合に重複がある場合に IA グラフが重複していると呼ぶことにする．つまり，共通の頂点及び辺を有していても，共通のアイテムが存在しなければ共通の頂点ではないし，共通のアイテムを有していても頂点や辺に重複が無ければそれらの ISS は重複が無いと考える．

図 2(B) に示した G_1 から G_5 は，図 2(A) から抽出された，共通アイテム集合が 2 個以上という条件を満たす，5 つの ISS

であり，それぞれ {A,B},{A,B,C},{B,X},{B,X},{B,C} が共通アイテム集合である．これらの中で G_1 と G_2 は，頂点 2,3,6,7, かつ，共通アイテム集合のうち A,B が重複しているため，G_1 と G_2 は重複している．同様に G_2 と G_5 も重複している．一方，G_3 と G_4 は，共通アイテム集合に重複が見られるが頂点が異なるために重複はしていないと捉える．

このような重複した ISS が，どのような状況から生じるかを考える．自然現象や観測には観測漏れや揺れが存在する．G_1 と G_2 を考えた場合，頂点 1,2,3,6,7 を持ち共通アイテム集合として A,B,C を有するグラフが存在し，そのグラフに観測ノイズが載った結果 G_1 と G_2 が観測されたと考えられる．たとえば G_1 と G_2 で異なる点は頂点 G_1 が G_2 に比べ頂点 1 を多く有していること，アイテム C を有していないことが異なる点である．頂点 1 のアイテム集合にアイテム C が含まれていれば，G_1 と G_2 は，頂点 1,2,3,6,7 を持ち共通アイテム集合が A,B,C である ISS の部分グラフであり．分かれずに済んだ．

本節では，このような背後に潜むアイテム集合付きグラフ構造を推定するため，ISS と ISS の併合を考える．

［定義 2］ (IA グラフの併合) 2 つの IA グラフを G_i, G_j とする．頂点 $V(G_i) \cup V(G_j)$ 及び辺 $E(G_i) \cup E(G_j)$ を有し，アイテム集合 $I(G_i) \cup I(G_j)$ が付与されている IA グラフ G' を考える．この時，G' を G_i と G_j を併合したグラフと呼ぶ．

併合により作成した G' は，作成方法により必ず連結グラフであることに注意されたい．

本研究では，一定以上の重複を持ったグラフに関して，併合したグラフを本質的な IA グラフであると考える．もちろん，定義ではない併合方法も考えられ，図 2 で G_1 と G_2 は頂点 2,3,6,7 からなるグラフで，アイテム集合として A,B が付与されているグラフにノイズが入ったデータとして生成されたと考えることも可能である．しかし，遺伝子ネットワークや SNS の解析においては利用者が事前知識を持っていることが多いため，過小評価した結果より，多少のミスを含んだとしても大きな範囲で取られたデータの方が利用しやすい点を考慮し，IA グラフの併合として和集合を取っている．

2.3 IA グラフの併合

本節では，IA グラフを併合する具体的手順を提案する．IA グラフ間の重複の多さを測る指標を定義し，その定義を利用して併合する IA グラフを決める．本提案ではボトムアップ型の階層的クラスタリング同様に最も重複した IA グラフを発見し，再帰的に併合する．

［定義 3］ (重複度) 現在存在している IA グラフの集合を $\mathcal{G} = \{G_1, ..., G_n\}$ とする．$G_i, G_j \in \mathcal{G} (i \neq j)$ に対し重複度 $intersec(G_i, G_j)$ を

$$intersec(G_i, G_j) = \frac{|\, V(G_i) \cap V(G_j) \,|}{|\, V(G_i) \cup V(G_j) \,|} \times \frac{|\, I(G_i) \cap I(G_j) \,|}{|\, I(G_i) \cup I(G_j) \,|}$$

と定義する．

この指標は，G_i, G_j の両方に関し頂点及びアイテム集合の両者の重複が大きい場合に値が大きくなる．また，辺の重複を見ていないが，これは IA グラフに共通するアイテム集合が頂点だけから決まり辺は関わらないことから，IA グラフの連結性のみを考慮し，辺の重複割合は考えていない．

表 1　閾値と併合数の関係

閾値	0.9	0.85	0.8
併合数	0	64	1690

図 2 の G_1 と G_2 に関して，具体的に重複度を計算する．$V(G_1) \cap V(G_2)$={2,3,6,7} より $|V(G_1) \cap V(G_2)|$=4, $V(G_1) \cup V(G_2)$ ={1,2,3,6,7} より $|V(G_1) \cup V(G_2))| = 5$. 同様に，$I(G_1) \cap I(G_2)$ ={A,B} より $|I(G_1) \cap I(G_2)| = 2$, $I(G_1) \cup I(G_2)$ ={A,B,C} より $|I(G_1) \cup I(G_2)| = 3$. 以上より，

$$intersec(G_1, G_2) = 4/5 \times 2/3 = 0.533$$

この重複度を $\mathcal{G}$ 内の IA グラフペアに対して求め，最も大きな重複度を持つペアに関し併合した上 $\mathcal{G}$ に挿入．また，併合前の 2 つの IA グラフは $\mathcal{G}$ から削除する．以上の操作を，最大の重複度が予め定義した閾値を切るまで繰り返す．この最大値は，単調減少とは限らないが，実データにおいては大幅に増加することはないため，一度閾値を切った時点で終了している．

図 2(B) から，閾値を 0.5 として具体的に計算すると，最も重複度の大きな IA グラフペアは，G_1 と G_2 もしくは G_2 と G_5 (いずれも 0.533) の 2 種類である．いずれから併合しても構わないが，ここでは添え字の小さい G_1 と G_2 を併合する．我々の実装では，先に見つかった物を優先的に併合している．G_1 と G_2 を併合すると,G_1 と G_2 が除かれ，$G_{1,2}$（図 2(C) 頂点 1,2,3,6,7 からなる，共通アイテム集合 {A,B,C} をもつ）が加わる．次は，G_3，G_4，G_5，$G_{1,2}$ の互いの重複度を求めるが，このうち最も重複の多い IA グラフペア，G_5，$G_{1,2}$ の重複度は 0.444 となり，この値は閾値 0.5 を下回っているので，この時点で併合を終了する．

3. 解　　析

本章では実際の遺伝子データに対し，本手法を適用することで手法の有用性を検証する．

3.1 実験データ

本実験では，IA グラフのグラフとして，iRefIndex [7] の遺伝子ネットワーク，アイテム集合として BioGPS [8] で公開されているヒトの 79 組織に対し 2 回ずつ実験を行ったものを用いた．遺伝子発現量は，各遺伝子に正規分布を当てはめた時に p 値が 0.05 より低い場合に高発現であると見なした．ISS として導出される部分グラフは，組織特異的に働く遺伝子群とその組織を示している．本データから生成される IA グラフは，頂点数 15,519 個, アイテム数 $79 \times 2 = 158$ 種, 辺の数 235,407 本である．

本実験では，このデータに対し COPINE [4] を用い ISS の最小サイズ $\theta_S = 20$, ISS の共通アイテム最小サイズ $\theta_I = 5$ として ISS を列挙した．列挙の結果，7,063 個の ISS が抽出された．この ISS を併合していく．

3.2 実 験 結 果

本節では，併合を止める閾値と併合の数の関係を見ていく．閾値と併合数の関係を表 1 に示す．終了の閾値を下げると指数

級数的に併合する IA グラフの数が増えることが分かる．これは，COPINE [4] で求められた ISS の多くで *intersection* の指標が 0.8 前後の値に集中しており，少し閾値を小さくしただけで大量のグラフ併合が起こるためである．

3.3 併合の実例

本節では，実際に併合された IA グラフを例に挙げ，本手法の有効性を示す．

図 3(A),(B) に示した 2 つの IA グラフが図 3(C) の様に併合された．各頂点は遺伝子を表し，各辺は遺伝子間に関係があることを示している．また，各グラフの下に書かれているアイテム集合は各 IA グラフの持つ共通のアイテム集合である．(A) と (B) の相違は，黄色で示した (B) の頂点 UNC119 のアイテム集合に，Myeloid（(A) の赤字で示したアイテム）が含まれていないことである．ところで，今回使用したデータにおいては，同一のサンプルから 2 度同じ実験を行っている．(A),(B) の IA グラフとも，2 回の内もう 1 回の Myeloid における実験がアイテムとして含まれており，共通アイテム集合に Myeloid を入れることで，実験誤差等による値のぶれで発見できなかったネットワークを補完できた可能性が高いことが分かる．

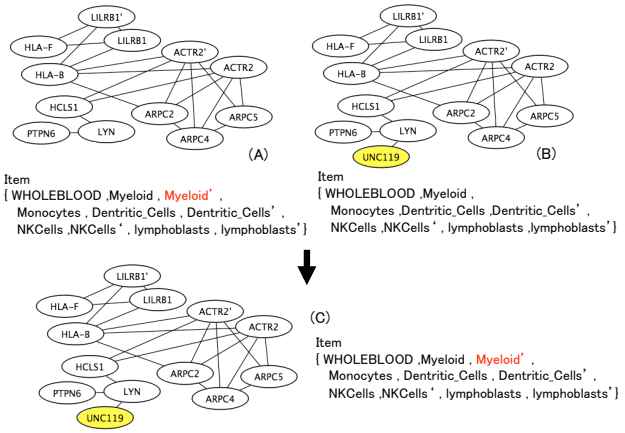

図 3　併合による遺伝子が追加される例

また異なる例として，図 4 に示す 2 つの IA ネットワークも併合されていた．先の例と同様に併合によって増える遺伝子を黄色で，増えるアイテムを赤色で書いてある．この併合で増えている実験は，先ほど同様 2 回実験されており，他方の実験もアイテムとして含まれるため，実験誤差による観測結果の誤差を本併合により補完できた可能性が高い．更に，この 2 つのネットワークの差である LYN 遺伝子の値は 2 回観測されており (頂点 LYN')，つまり本併合により，本質的な組織集合と遺伝子集合に変わりはないうえで，アイテムだけでなく，遺伝子に関しても実験誤差で落ちてしまった可能性のある頂点を拾うことが可能になっている．

また，本事例は疎なネットワークとなっている．多くのネットワーク解析アルゴリズムが密なネットワークを見つけるのに対し，COPINE の利点は疎なネットワークであっても共通点の高いネットワークを見つけることが可能な点である．

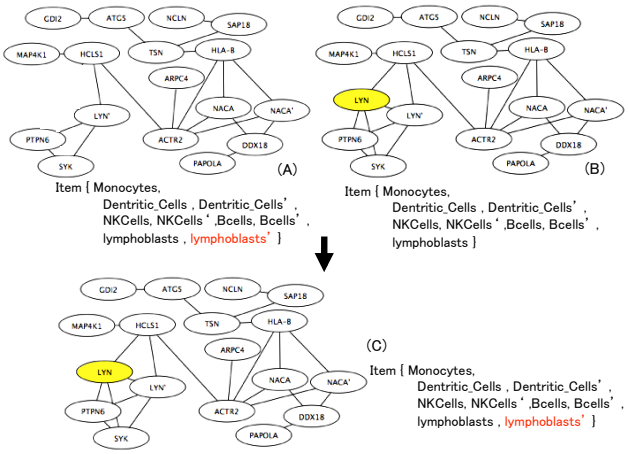

図 4　2 つの同質なネットワークが併合された例

4. ま と め

ノードにアイテムを付与した大規模グラフデータから，共通アイテムをもつ連結部分グラフを COPINE アルゴリズム [4] を用い列挙する．そしてその ISS 集合に対し，それぞれ重複を評価し，重複度を評価して高いものから順に併合，それを繰り返し，解釈容易な ISS 集合を生成した．ネットワーク全体が示す特徴を見出すのに有用である．

文　献

[1] R. Agrawal and R. Srikant. Fast Algorithms for Mining Association Rules. VLDB, 487–499, 1994.

[2] J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. SIGMOD 2000, 1–12, 2000.

[3] , M. Girvan and M. E. J. Newman. Community structure in social and biological networks. Proc Natl Acad Sci, vol. 99, 12, 7821–6, 2002.

[4] M. Seki and J. Sese, Identification of Active Biological Networks and Common Expression Conditions. IEEE BIBE 2008, pp. 1–6, 2008.

[5] T. Itoh, C. Muelder, K. Ma, and J. Sese. A Hybrid Space-Filling and Force-Directed Layout Method for Visualizing Multiple-Category Graphs. IEEE Pacific Visualization Symposium 2009, pp.121-128, 2009.

[6] M. Fukuzaki, M. Seki, H. Kashima and J. Sese, Side Effect Prediction Using Cooperative Pathways. IEEE BIBM 2009, pp. 142–147, 2009.

[7] S. Razick, G. Magklaras, and I. M Donaldson, iRefIndex: a consolidated protein interaction database with provenance, BMC Bioinformatics, vol. 9, pp. 405, 2008.

[8] A. Su, T. Wiltshire, *et al.*, A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci, Vol. 101, No 16, pp 6062-7, 2004.