定量的アイテム集合付き単一グラフからの 頻出パターンマイニング

Frequent Pattern Mining from a Single Graph with Quantitative Itemsets

三好 裕樹
Yuuki Miyoshi

神戸大学 大学院工学研究科
Graduate School of Engineering, Kobe University
miyoshi@cs25.scitec.kobe-u.ac.jp

尾崎 知伸
Tomonobu Ozaki

大阪大学 サイバーメディアセンター
Cybermedia Center, Osaka University
tozaki@dcm.cmc.osaka-u.ac.jp

江口 浩二
Eguchi Koji

神戸大学 大学院システム情報学研究科
Graduate School of System Informatics, Kobe University
eguchi@port.kobe-u.ac.jp

大川 剛直
Takenao Ohkawa

(同 上)
ohkawa@kobe-u.ac.jp

keywords: graph mining, quantitative itemsets, data mining

Summary

In this paper, we discuss pattern mining problems from a single graph whose vertices or edges contain a set of numerical attributes. Several networks can be naturally represented in this kind of complex graph. A typical example is a social network whose vertex corresponds to a person with some numerical attributes such as age, salary and so on. Another example is a communication network whose edge represents a communication between devices. We can associate the numbers of communications per certain time period to each edge. For these kinds of complex graphs, it is meaningful to consider not only graph structure but also the internal informations. Although it can be expected that these kinds of data will increase rapidly, most of current graph mining algorithms do not handle these complex graphs directly. Motivated by the above background, we developed algorithms named FAG^V -gSpan and FAG^E -gSpan for finding frequent patterns from a single graph with numerical attributes, by effectively combining techniques of graph mining and quantitative itemset mining. The effectiveness of the proposed algorithms was confirmed by experiments using real world datasets.

1. は じ め に

グラフ構造は、社会ネットワークやタンパク質構造など、構造データを表現する際に利用される汎用的なデータ構造である。これらグラフデータベース中に頻出する部分構造（部分グラフ）を発見することは、グラフマイニング分野における重要な問題のひとつとして認識され、近年、数多くの研究が行われている [Washio 04, Washio 03]。これらの研究の多くは、単純なラベル付きグラフを対象としているが、対象のより自然かつ精密な表現を考えた場合、必ずしもラベル付きグラフで十分であるとは限らない。そこで本論文では、対象のより精密な表現手段として、(1) 頂点に数値属性の集合が付与された頂点数値属性付きグラフ (AG^V) 及び (2) 辺に数値属性の集合が付与された辺数値属性付きグラフ (AG^E) の 2 種類の数値属性付きグラフを採用し、そこからのパターン発見について議論する。

グラフで表現できる代表的なデータの 1 つとして、人

を頂点、人間関係を辺で表したコミュニケーションネットワークが考えられる (図 1(i))。特に近年では、ネットワークの発達なども伴い、人と人とのコミュニケーションが頻繁に行われるようになり、その構造を表すコミュニケーションネットワークや社会ネットワークデータの分析は非常に重要になってきている [Lewis 09, Newman 03, 安田 01]。コミュニケーションは、ただ単に身近な知り合いだけでなく共通の趣味や思考などの繋がりにより生まれる場合がある。そのため、役職のような 1 つの情報だけでなく、所属しているコミュニティ、興味のある分野などといった個人の特徴を表す個人情報も重要な情報となりうる。したがって、これらの情報も含んだ形でより精密なデータ分析を考えた場合、ネットワークを単なるラベル付きのグラフとして表現するだけでは不十分であり、頂点に複数の情報を持ったグラフが必要とされることが容易に想像される。

一方、辺数値属性付きグラフによって自然に表現でき

図1 ネットワークデータの例

るデータの例として、図 1(ii) に示すような電子メールネットワークが考えられる。電子メールネットワークは、各通信機器を頂点とし、ある期間内でメールの送受信が行われた機器同士を辺で結ぶことで表現される。ここでもより精密な分析を行うために、付加的な情報として、各通信機器間における送受信方法 (To, Cc, Bcc) に対する通信回数を辺に割り当てることを考えることができる。また、ネットワークの時間的な変化に関する知見は、図 1(iii) に示すような動的ネットワーク (グラフ系列) を分析することで得られると考えられる。その際、例えば、電子メールネットワークに対する動的グラフでは、各時間における通信回数やその差分などを辺に割り当てることで、グラフ構造の変化だけでなく、通信回数の変化も合わせた形式での知見を得ることが期待できる。

このような、頂点や辺に複数の情報が付与されたグラフデータからの知識発見やデータマイニングに対する需要や期待を背景に、近年では特に、頂点によりリッチな情報を持たせたグラフデータを対象としたマイニング手法の研究が報告されている [Chen 04, Moser 09, 山本 08]。詳しくは後述するが、これらの手法では、グラフ構造と頂点に含まれるアイテム集合や系列データを同時に考慮することで、特徴的な部分グラフを抽出することを目的としている。一方、本論文では、グラフの頂点や辺に付与する情報として数値属性集合を採用し、特に、数値属性付きグラフを対象とした特徴的な部分グラフを発見するアルゴリズムを提案する。グラフデータへ付与する情報として数値属性集合を採用することで、例えばコミュニケーションネットワークでは、年齢や年収といった個人情報だけでなく、所属するコミュニティの情報発信量などの大域的な情報も表現可能となり、個別頂点の情報では検出が困難であったネットワークの大域的な変化も表現可能となると考えられる。これらのことから、数値属性グラフからのパターン発見においては、ラベル付きグラフにはなかった情報を積極的に考慮することで、対

象が持つ豊かな情報を考慮したパターンの発見を期待することができる。

本論文では、頂点数値属性付きグラフ (AG^V) 及び辺数値属性付きグラフ (AG^E) に対し、それぞれ特徴的なパターン発見アルゴリズムを提案する。 AG^V に対するアルゴリズムを FAG^V -gSpan, AG^E に対するアルゴリズムを FAG^E -gSpan と呼び、総称して FAG -gSpan と呼ぶ。 FAG -gSpan は、頻出部分グラフを発見するアルゴリズム gSpan [Yan 02] による (1) ラベル付きグラフパターンの列挙と、高密度クラスタに基づいて定量的アイテム集合を発見するアルゴリズム QFIMiner [Washio 05] を利用した (2) 定量的アイテム集合の割り当てを繰り返すことで、特徴的な部分グラフを発見する。なお詳しくは後述するが、グラフパターンから高密度クラスタを抽出する際、グラフ形式を考慮した定量的アイテム集合のトランザクション化が必要となる。そこで、データグラフに対するグラフパターンの出現を導入することでこの問題の解決を試みる。また抽出された数値属性付きグラフパターンに対し、出現をもとに (i) 支持度の計算と (ii) 標準形判定を導入することにより、重複なく頻出パターンを列挙する。

以下に本論文の構成を示す。2 章で、頂点数値属性付きグラフの定式的な定義を与えると同時に、本論文で扱う問題を定式化する。3 章で、提案する頂点数値属性付きグラフパターン発見アルゴリズム FAG^V -gSpan を提案し、その詳細を説明する。ついで 4 章で、辺数値属性付きグラフの定式的な定義と、それに対するパターン発見アルゴリズム FAG^E -gSpan を与える。5 章で実験結果を示す。また 6 章で関連研究について述べ、最後に 7 章でまとめを行う。

2. 頂点数値属性付きグラフパターン

本論文ではまず、ラベル付きグラフの各頂点に、定量的アイテム集合 (数値属性の集合) が付与されたグラフデータを対象とする。以下では、[Bringmann 08, Washio 05, Yan 02] などに基づき、対象とするグラフデータの形式的な定義などを与え、本論文で対象とするデータマイニング問題を導入する。

2.1 ラベル・頂点数値属性付きグラフ

p を属性名, $[f, t]$ をその値の範囲とし, $\langle p : [f, t] \rangle$ の形式で定量的アイテム (数値属性) を表現する。このとき, $\{\langle p_1 : [f_1, t_1] \rangle, \dots, \langle p_m : [f_m, t_m] \rangle\}$ で構成される m 個の数値属性からなる集合を定量的アイテム集合 [Washio 05] と定義する。例えば, $\{\langle \text{年齢} : [30, 35] \rangle, \langle \text{家族人数} : [2, 4] \rangle\}$ は、2 つの数値属性からなる定量的アイテム集合であり、年齢が 30 ~ 35 歳で、家族人数が 2 ~ 4 人であることを表す。

一方、ラベル集合 $\mathcal{L}$ 上のラベル付きグラフ G とは、頂

図2 頂点数値属性付きグラフにおける同型グラフの例

点集合 V_G 及び辺集合 $E_G \subseteq V_G \times V_G$, ラベル関数 $l_G : V_G \cup E_G \rightarrow \mathcal{L}$ の三項組 $G = (V_G, E_G, l_G)$ で表される.

ラベル集合 $\mathcal{L}$ 及び数値属性集合 $\mathcal{A}$ 上の頂点数値属性付きグラフ (以降, AG^V と略記する) $G = (V_G, E_G, l_G, a_G^V)$ とは, ラベル付きグラフ $s^V(G) = (V_G, E_G, l_G)$ の各頂点 $v \in V_G$ に, 関数 $a_G^V : V_G \rightarrow \text{pow}(\mathcal{A})$ により (空でない) 数値属性集合を割り当てたグラフである. ここで $s^V(G)$ は G の各頂点から定量的アイテム集合を取り除いたグラフを, $\text{pow}(\mathcal{A})$ は $\mathcal{A}$ のべき集合を表す. すなわち頂点数値属性付きグラフとは, 外部構造であるラベル付きグラフに対し, その各頂点に内部構造として数値属性集合を割り当てたグラフである. 例えば図2では, G_1, G_2, G_3 は G' の各頂点に数値属性を割り当てた AG^V である ($s^V(G_1) = s^V(G_2) = s^V(G_3) = G'$).

2.2 頻出頂点数値属性付きグラフパターン

2 つの定量的アイテム $x = \langle p^x : [f^x, t^x] \rangle$ と $y = \langle p^y : [f^y, t^y] \rangle$ に対し, $p^x = p^y \wedge f^x \leq f^y \wedge t^x \leq t^y$ が成り立つ, すなわち両者が同じ属性でありかつ x の範囲内に y の範囲があるとき, x は y に支持されるといい, $x \preceq y$ と表記する. また, 2 つの定量的アイテム集合 X と Y に対し, $\forall x \in X \exists y \in Y [x \preceq y]$ が成り立つとき, X は Y に支持されるといい $X \preceq Y$ と表記する. 2 つの AG^V $P_1 = (V_{P_1}, E_{P_1}, l_{P_1}, a_{P_1}^V)$ と $P_2 = (V_{P_2}, E_{P_2}, l_{P_2}, a_{P_2}^V)$ に対し, 以下の条件を満たす単射関数 $f^V : V_{P_1} \rightarrow V_{P_2}$ が存在するとき, P_1 を P_2 の部分グラフと定義し, $P_1 \subset P_2$ と表記する.

- (1) $\forall v \in V_{P_1} [l_{P_1}(v) = l_{P_2}(f^V(v))]$
- (2) $\forall v \in V_{P_1} [a_{P_1}^V(v) \preceq a_{P_2}^V(f^V(v))]$
- (3) $\forall (u, v) \in E_{P_1} \exists (f^V(u), f^V(v)) \in E_{P_2}$
 $[l_{P_1}(u, v) = l_{P_2}(f^V(u), f^V(v))]$

また, $P_1 \subset P_2$ かつ $P_2 \subset P_1$ が成り立つとき, P_1 と P_2 は同型であると定義する. なお本論文では, 部分グラフのことをパターンとも呼ぶ.

図2において, G_1, G_2, G_3 は互いに同型グラフである. また図4は入力データとして頂点数値属性付きグラフ G を与えたときの属性付きグラフパターンの生成過程の例

P_1	P_2	P_3	P_4
(0,1,X,-,Y) (1,2,Y,-,Z)	(0,1,Y,-,Z) (0,2,Y,-,X)	(0,1,Y,-,X) (0,2,Y,-,Y)	(0,1,Z,-,Y) (1,2,Y,-,X)

図3 ラベル付きグラフパターンのコードワードの例

である. $P_2 \subset G$ より, P_2 は G におけるパターンのひとつであることが分かる.

詳細は [Yan 02] に譲るが, グラフパターン中の各辺を, (i) 辺を構成する2頂点の各訪問順 (追加順), (ii) 各頂点ラベル, (iii) 辺ラベルの5項組みで表現するとともに, それらを辺の追加順に連結することで, 各パターンをコードワードと呼ばれる文字列で表現することが可能である. また, 同型なパターン間で辞書式順で最小のコードワードを持つものを標準形と定義し, 同型グラフ集合の代表元とすることで, 同型グラフの重複列挙の問題に対処することが可能である. 図3において, 4つの同型なパターン間で最小のコードワードを持つ P_1 が標準形となる. 詳細は, [Yan 02] を参照されたい.

標準形グラフ P 及び $P \subset G$ なるグラフ G と, P と G の間で上記の条件 (1)-(3) を満たす単射関数 f^V に対して, P 中の各頂点に対応する G 中の頂点を, P のコードワード生成における訪問順 (追加順) に並べたリスト

$$o^V = [v_1, v_2, \dots, v_{|V_P|}]$$

$$(v_i \in V_G, v_i = f^V(u_i), u_i \in V_P)$$

を G における P の頂点出現と呼ぶ. ここで, o^V 中の i 番目の頂点 $o^V(i)$ は, P の i 番目の頂点に対応する G 中の頂点である. また, この様な頂点出現は複数考えられるので, その集合を $\Phi_G^{P^V}$ と表記する.

以上の準備の下, G に対する P の支持度を

$$\text{sup}_G^V(P) = \min_{1 \leq i \leq |V_P|} |\{o^V(i) | o^V \in \Phi_G^{P^V}\}| / |V_G|$$

と定義する. つまり, P の各頂点 v に対し, v に対応するすべての出現 $o^V \in \Phi_G^{P^V}$ 上の頂点数を求め, その最小値を P の支持度とする. なおこの定義は, 単一のラベル付きグラフにおける支持度 [Bringmann 08] に対する素直な拡張であり, グラフの特殊化に関して逆単調性 ($P \subset Q \rightarrow \text{sup}_G^V(P) \geq \text{sup}_G^V(Q)$) が成り立つ. パターンの支持度の計算方法を図4を使って説明する. $o_{21} \sim o_{24}$ は, G における P_2 のすべての出現である. P_2 の頂点 X に対応する出現上の頂点数は 4 (v_1, v_3, v_4, v_7), Y の場合は 2 (v_5, v_8) であるので, 支持度は Y の2となる.

頂点数値属性付きグラフ G と利用者によって与えられた最小支持度 $\sigma (0 < \sigma \leq 1)$ に対し, $\text{sup}_G^V(P) \geq \sigma$ を満たすと共に, すべての頂点に数値属性集合を含み, かつ

連結である頂点数値属性付きグラフ P を，頻出頂点数値属性付きグラフパターン（以後， FAG^V ）と定義する．本論文の目的の一つは， G と σ が与えられたとき， FAG^V を抽出することである．

3. 頂点数値属性付きグラフパターンの発見

本章では，ラベル付きグラフパターンの列挙，及び定量的アイテム集合における高密度クラスタ，属性集合の割り当てと AG^V における代表元を導入する．その後，これらに基づき，頻出数値属性付きグラフパターン発見アルゴリズム FAG^V -gSpan を提案する．詳しくは後述するが， FAG^V -gSpan は，グラフマイニング手法によるラベル付きグラフパターンの列挙と，高密度クラスタから得られた定量的アイテム集合の割り当てを繰り返すことでパターンを生成する．

3.1 ラベル付きグラフパターンの列挙

本論文では，ラベル付きグラフパターンの列挙のために，最右拡張を用いた逆探索 [Avis 93] に基づく頻出部分グラフ発見アルゴリズム gSpan [Yan 02] を援用する．なお gSpan はグラフデータ集合を対象とした手法であるが，支持度の定義を修正することで，比較的容易に単一グラフからの頻出ラベル付き部分グラフの発見へと応用が可能である．

グラフパターン g において，2 章で導入した頂点の訪問順で，開始点となる頂点を根，最遅の頂点を最右頂点としたとき，根から最右頂点に至る経路を最右パスと呼ぶ．gSpan では，グラフパターン g に対し，(1) 最右パス上の頂点と最右頂点間に辺を追加する，もしくは (2) 最右パス上の頂点から新たな頂点をともなう辺を追加することで，新たなグラフパターンを得る．この操作を最右拡張と呼ぶ．なお，頂点数 1 のグラフパターンに対し最右拡張を繰り返し適用することで，全てのグラフパターンを漏れなく列挙することが可能である．その一方で，単に最右拡張を適用するだけでは，標準形でないグラフパターンも候補として生成されてしまう．これに対し，「非標準形グラフに対する最右拡張の適用により，標準形グラフが得られることはない」という性質 [Borgelt 05] を利用し，最右拡張により候補として生成されるグラフパ

ターンに対して標準形判定を行うことで，非標準形グラフの列挙，すなわち同型グラフの重複生成を回避する．詳細は [Borgelt 05] を参照されたい．

3.2 高密度クラスタ

本節では，文献 [Washio 05] に従い，定量的アイテム集合における高密度クラスタを導入する．

定量的アイテム集合からなるデータセット D を考える． D 中のトランザクション（定量的アイテム集合） T と，数値属性 p に対し， T の近傍集合を

$$N_{\Delta_p}(T) = \{T' \in D \mid \text{Dist}_p(q, q') \leq \Delta_p\}$$

と定義する．また $|N_{\Delta_p}(T)| \geq \text{MinPts}$ のとき， T を p 上のコアトランザクションと呼ぶ．ここで， $\text{Dist}_p(q, q')$ は，属性 p 上でのトランザクション T と T' の距離， Δ_p 及び MinPts は利用者により与えられる閾値である．

2 つのトランザクション T と T' が，以下のいずれかの条件を満たすとき，両者は属性 p 上で連結していると定義する．

- (1) $T' \in N_{\Delta_p}(T)$ かつ T はコアトランザクション
- (2) $T' \in N_{\Delta_p}(T_1), T_1 \in N_{\Delta_p}(T_2), \dots, T_n \in N_{\Delta_p}(T)$ なる p 上のコアトランザクションの集合 $T_i (1 \leq i \leq n)$ が存在．

また，与えられたパラメータ Δ_p に対し，属性 p 上で互いに連結しているトランザクションの集合を， $(\Delta_p$ 以内で) 隣接しているトランザクション集合と定義する．

$C \subseteq D$ 中のすべてのトランザクションが，属性 p 上で連結しているとき， C は p 上で密である [Washio 05] と定義し， $\text{dense}(C, p)$ と表記する．次に属性集合 S を考える．トランザクション集合 C^S が， S 中のすべての属性で密であるような極大集合であるとき， $\text{dense}(C^S, S)$ と表記し， C^S を高密度クラスタと定義する [Washio 05]．

データセット D と利用者によって与えられた最小支持度 $\sigma (0 < \sigma \leq 1)$ に対し， $|C^S| \geq |D| \times \sigma$ を満たす高密度クラスタ C^S は，一つの定量的アイテム集合に対応する．具体的には， S 中の各属性に対し，その最小，最大値を考えることで，以下の定量的アイテム集合が得られる．

$$\left\{ \langle p_i : [f_i, t_i] \rangle \mid \begin{array}{l} p_i \in S, \\ f_i = \min\{f \mid \langle p_i : [f, t] \rangle \in T \in C^S\}, \\ t_i = \max\{t \mid \langle p_i : [f, t] \rangle \in T \in C^S\} \end{array} \right\}$$

高密度クラスタは、その定義により「もしトランザクション集合 C が属性 S 上で高密度クラスタでないとき、より上位の属性集合 $S' \supset S$ 上で高密度クラスタとなることはない」($\neg \text{dense}(C, S) \rightarrow \forall S' \supset S [\neg \text{dense}(C, S')]$) という属性集合に関する単調性が成り立つ。この性質を利用し QFIMiner と呼ばれる、アイテム集合に対する階層的探索アルゴリズムを定量的アイテム集合発見に拡張したアルゴリズムが提案されている [Washio 05]。

3.3 属性集合の割り当てと頂点数値属性付きグラフの代表元

標準形を満たすラベル付きグラフパターンに対し、高密度クラスタを定量的アイテム集合として割り当ててことで、 AG^V を生成することを考える。その際、3.2 節で導入した QFIMiner を利用して高密度クラスタを抽出することを考えるが、QFIMiner は定量的アイテム集合のデータセット（トランザクション集合）を対象としておりグラフパターンを直接扱うことはできない。そこで、グラフパターンのトランザクション集合への変換と、得られた定量的アイテム集合のグラフパターンへの割り当て方法を導入する。加えて、高密度クラスタ生成において、単純な変換だけでは高密度でない部分まで抽出されてしまうため、隣接トランザクション集合の導出基準の修正も必要となる。さらに、生成された属性付きグラフパターンには同型グラフが存在する場合があるので、同型グラフの重複列挙を避けるため、標準形を定め、標準形でないパターンを排除する必要がある。

以上のように各々の問題に対し本節では、(1) グラフ形式からトランザクション集合への変換、(2) 隣接トランザクション集合の定義の修正、(3) 標準形の考慮、の順で、ラベル付きグラフパターンから属性付きグラフパターンを生成する方法について述べる。

§1 トランザクション集合への変換

グラフパターンからトランザクション集合への変換は、グラフ中の定量的アイテム集合に着目し、ラベル付きグラフパターン p 中の各頂点に対応する頂点数値属性付きグラフ G 中の頂点リストの集合、すなわち全出現 Φ_G^{pV} に基づいて行われる。頂点リスト $o = [v_1, \dots, v_n] \in \Phi_G^{pV}$ に対応する定量的アイテム集合のリスト $A^V(o)$ を

$$\left\{ \begin{array}{l} \langle 1 - p_1^1 : [f_1^1, t_1^1] \rangle, \dots, \langle 1 - p_1^{m_1} : [f_1^{m_1}, t_1^{m_1}] \rangle, \\ \dots, \\ \langle n - p_n^1 : [f_n^1, t_n^1] \rangle, \dots, \langle n - p_n^{m_n} : [f_n^{m_n}, t_n^{m_n}] \rangle \end{array} \right\}$$

と定義する。すなわち、各頂点中の定量的アイテムの属性名 p_i^j に頂点番号 k を付与し（属性名を $k - p_i^j$ に置き換える）、それらを一つの定量的アイテム集合へと変換する。この変換により、同一の数値属性を持つアイテムを区別することができるようになる。

図5 同型な出現の例

§2 隣接トランザクション集合の定義の修正

出現を基にグラフパターンをトランザクション集合へ変換した後、 $\{A^V(o) \mid o \in \Phi_G^{pV}\}$ を対象とした高密度クラスタ抽出を通じ、(属性なしの) ラベル付きグラフ p に対する定量的アイテム集合の付与、すなわち属性付きグラフパターンの生成が実現される。このとき、例えば図4の o_{11}, o_{12}, o_{13} における v_5 のように、複数の出現 (o_{11}, o_{12}, o_{13}) が同じ頂点 (v_5) を含む場合がある。このような状態で、トランザクション集合 $\{A^V(o) \mid o \in \Phi_G^{pV}\}$ に対してクラスタリングを行うと、同一の頂点のアイテムを重複して数えてしまうため、実際には高密度にならない部分まで抽出されるという問題が存在する（同一の頂点のアイテムを複数のトランザクションに与えるために生じる問題である）。そこで、グラフにおける属性 a 上の $o \in \Phi_G^{pV}$ の近傍集合を以下のように定義する。

$$N_{\Delta_a}(o) = \{o' \in \Phi_G^{pV} \mid \text{Dist}_a(o, o') \leq \Delta_a, V_a(o) \neq V_a(o')\}$$

$V_a(o)$ は、 o 中の属性 a に対応するグラフの頂点を表す。つまりグラフにおける属性 a 上の隣接集合は、異なる頂点からなる属性 a を持ったトランザクションに対し、パラメータ Δ_a 以内で隣接しているトランザクションの集合である。トランザクション集合の変換の例を、図4に示す。 P_1 の全出現を表のように定量的アイテム集合トランザクションに変換する。このデータセットを利用し、QFIMiner により定量的アイテム集合を抽出し、 P_1 の各頂点に割り当ててことで属性付きグラフパターン P_2 を得る。

§3 属性付きグラフパターンの標準形

ラベル付きグラフパターンに定量的アイテム集合を割り当てて得られた FAG^V には、ラベル付きグラフ同様、幾何的な形は異なるが、頂点の接続関係が一致する同型グラフが生成される可能性がある。本論文では外側のラベル付きグラフが標準形であるパターンのみに対し出現を用いて高密度クラスタとなる定量的アイテム集合を割り当てて AG^V を生成する。この際、外部のグラフ構造が決まっても割り当て方によって同型なグラフが生成される場合が存在する。これは出現の同型性によるもので、同じグラフパターンでも順序が変わって重複した出現が存在するために起こる問題である。例えば、図5の属性付きグラフ G に対する P の2つの出現 o_1, o_2 は、対応する頂点は異なるが同型のグラフの出現を表している。このように同型な出現群からは同型な AG^V が生成される

図6 属性付きグラフパターン発見における探索空間の例

可能性があるため、得られたパターンから同型なパターン集合に対し、順序をつけることで代表元のみを出力する。例えば、図6の S_{23}, S_{24} は、互いに同型であり、 S_{23} が代表元である。

以下では、属性付きグラフの代表元について議論する。

FAG^Vの外部構造であるラベル付きグラフはgSpanによって標準形を満たすグラフが抽出される。そのため頂点に付与された数値属性の順序さえ考慮すれば、同型の属性付きグラフ間の順序（大小）関係を定めることができる。そこで、本節では以下のように頂点の番号順に、頂点に含まれる定量的アイテム集合を比較することで決定する。

まず、2つの定量的アイテム $x = \langle p^x : [f^x, t^x] \rangle$ と $y = \langle p^y : [f^y, t^y] \rangle$ に対し、その順序関係を以下の様に定義する。

- (1) 辞書順で小さい属性名(p^x と p^y)を持つ定量的アイテムが小さい。
- (2) 属性名が等しい場合は、より小さな“範囲の始点”(f^x, f^y)を持つ定量的アイテムが小さい。
- (3) 範囲の始点も等しい場合は、より小さな“範囲の終点”(t^x, t^y)を持つ定量的アイテムが小さい。
- (4) どの条件も満たさない場合、両者は等しい。

次に、定量的アイテム集合 $X = \{x_1, \dots, x_{|X|}\}$ と $Y = \{y_1, \dots, y_{|Y|}\}$ に対し、以下の順序関係を導入する。

- (1) $|X| < |Y|$ の場合、 X が小さい。
- (2) 条件 $\forall i (1 \leq i \leq j-1) [x_i = y_i]$ かつ $x_j < y_j$ を満たす j が存在する、すなわち、 j 番目の定量的アイテムが小さく、かつそれまでの定量的アイテムは等しい場合、 X が小さい。
- (3) すべての定量的アイテムが等しい場合($\forall i (1 \leq i \leq |X| = |Y|) [x_i = y_i]$)、両者は等しい。
- (4) いずれの条件も満たさない場合、 Y が小さい。

同型である2つの属性付きグラフ P_1 と P_2 に対し、頂点の番号順に頂点中の定量的アイテム集合を並べたリストを $L_1 = [a_1^1, \dots, a_1^{|V_{P_1}|}]$, $L_2 = [a_2^1, \dots, a_2^{|V_{P_2}|}]$ とする。このとき P_1 と P_2 の順序関係は、以下の様に定義される。

- (1) $\forall i (1 \leq i \leq j-1) [a_1^i = a_2^i]$ かつ $a_1^j < a_2^j$ を満たす

j が存在する、すなわち j 番目の定量的アイテム集合が小さく、かつそれまでの定量的アイテム集合が等しい場合、 P_1 が小さい。

- (2) すべての定量的アイテム集合が等しい場合($\forall i (1 \leq i \leq |V_{P_1}| = |V_{P_2}|) [a_1^i = a_2^i]$) 両者は等しい。

- (3) いずれの条件も満たさない場合、 P_2 が小さい。

すなわち $|L_1| = |L_2|$ であるので、頂点の番号順に定量的アイテム集合を比較し順序をつける。

以上の準備の下、属性付きグラフ P が以下の条件を満たすとき、 P は標準形であると定義する。

- (1) $s^V(P)$ はラベル付きグラフにおける標準形[Borgelt 05, Yan 02]である。
- (2) $P = \min\{Q \mid P \subset Q, Q \subset P, s^V(P) = s^V(Q)\}$ である。

すなわち P は、 P と同型かつ同一のラベル付きグラフを持つFAG^Vの中で最小となる。

3.4 FAG^V-gSpan

本節で、FAG^Vを発見するアルゴリズムFAG^V-gSpanを提案する。FAG^V-gSpanは、ラベル付きグラフパターンの列挙アルゴリズムgSpanと、グラフパターンの各頂点に割り当てる定量的アイテム集合を生成する定量的アイテム集合発見アルゴリズムQFIMinerの2つを上手く組み合わせることで、FAG^Vを発見する。FAG^V-gSpanの疑似コードを図7に示す。図7において、 G はマイニングの対象となるAG^V, σ は最小支持度、 $\mathcal{L}$ はラベルの全体集合である。

FAG^V-gSpanは、すべてのラベルに対し、頂点だけからなるグラフパターン p (図7, FAG^V-gSpanの2行目)に対してFAG^V-Enumを繰り返し適用することで、FAG^Vを列挙する。FAG^V-Enumでは、生成された各ラベル付きグラフパターン p に対し、 p が標準形でない(図7, FAG^V-Enumの1行目, $\neg \text{isCanG}(p)$)、もしくは頻出でない($\text{sup}_G^V(p) < \sigma$)場合は、それ以上の展開を止める。また p に対し、前節で導入した高密度クラスタ抽出手法を適用し、 p 中の各頂点へ定量的アイテム集合を割り当てることで、FAG^Vの候補集合 S を得る(2行目)。生成

Algorithm FAG^V-gSpan($G, \sigma, \mathcal{L}$)

```

1: for each  $p \in \mathcal{L}$ 
2:   FAGV-Enum( $p, G, \sigma, \mathcal{L}$ )

```

```

Subroutine FAGV-Enum( $p, G, \sigma, \mathcal{L}$ )
1: if  $\neg \text{isCanG}(p) \vee \text{sup}_G^V(p) < \sigma$  then return
2:  $S := \text{QFI}(\{A^V(o) \mid o \in \Phi_G^V\}, \sigma)$ 
3:  $S := \{s' \in S \mid \text{isCanA}^V(s') \wedge \text{sup}_G^V(s') \geq \sigma\}$ 
4: if  $S = \emptyset$  then return
5: for each  $s' \in S$ 
6:   output  $s'$ 
7:  $P := \text{RME}(p, \mathcal{L})$ 
8: for each  $p' \in P$ 
9:   FAGV-Enum( $p', G, \sigma, \mathcal{L}$ )

```

図 7 FAG^V-gSpan の擬似コード

された候補パターン $s' \in S$ に対し、標準形判定 (3 行目, $\text{isCanA}^V(s')$) と支持度判定 ($\text{sup}_G^V(s') \geq \sigma$) を行うことで、 s' が FAG^V であるかを判定し、FAG^V 集合として S (3 行目) を更新する。 S が空集合である場合 (4 行目)、すなわちラベル付きグラフ p から FAG^V が生成されなかった場合、 p の展開を中止する。各 $s' \in S$ を FAG^V として出力した後 (5–6 行目)、 p に対して gSpan における最右拡張 (7 行目, $\text{RME}(p, \mathcal{L})$) を適用することでより特殊なラベル付きグラフパターンの集合 P を生成し、各 $p' \in P$ に対して再起呼び出しを行うことで、より特殊なパターンの列挙を行う (8–9 行目)。なお、FAG^V-gSpan では、最右拡張のラベル付きグラフパターン列挙に関する完全性 [Yan 02] と頻度の逆単調性により、定量的アイテム集合の割り当てを通じ FAG^V となる可能性のあるグラフパターン (FAG^V の外部構造) はすべて列挙される。加えて、QFIMiner の性質により、割り当てに用いられる、グラフパターンの出現集合中に含まれる高密度クラスタ (内部構造) はすべて生成されることとなる。

FAG^V-gSpan の動作過程の例を図 6 に示す。FAG^V-Enum に与えるラベル付きグラフパターン P_1 を考える。 P_2 と P_3 は P_1 を最右拡張することで得られる (図 7, FAG^V-Enum の 7 行目)。 P_2 は標準形でないので (1 行目, $\neg \text{isCanG}(p)$)、 P_2 の拡張は枝刈りされる。 P_3 に定量的アイテム集合を割り当てることで、FAG^V の候補集合 S_1 が抽出される (2 行目)。 S_1 中に FAG^V が存在するとき (3 行目の条件を満たす s' が存在するとき)、 P_3 を最右拡張することにより新たなパターンを抽出する (7 行目)。 P_3 から FAG^V-Enum を 1 回適応することで、 P_4 を得られる。 P_4 に定量的アイテム集合を割り当てることで、FAG^V の候補集合 S_2 が抽出される (2 行目)。 S_{23} と S_{24} のような同型の AG^V が抽出されたとき、標準形のみが選ばれる (3 行目, $\text{isCanA}^V(s')$)。この例の場合、 S_{23} の

み出力される。一方で、 P_3 に新しい辺と頂点を追加することで、 P_5 が抽出される。今、 P_5 から FAG^V が抽出されなければ、 P_5 の拡張を止める (4 行目)。

4. 辺数値属性付きグラフからのパターン発見

本章では、FAG^V-gSpan でのアイデアを応用し、辺に定量的アイテム集合が付与された辺数値属性付きグラフ (以降、AG^E と略記) からのパターン発見アルゴリズム FAG^E-gSpan を提案する。

4.1 辺数値属性付きグラフとパターン

ラベル集合 $\mathcal{L}$ 及び数値属性集合 $\mathcal{A}$ 上の辺数値属性付きグラフ $G = (V_G, E_G, l_G, a_G^E)$ とは、頂点集合 V_G 及び辺集合 $E_G \subseteq V_G \times V_G$ 、ラベル関数 $l_G: V_G \cup E_G \rightarrow \mathcal{L}$ のラベル付きグラフ $s^E(G) = (V_G, E_G, l_G)$ の各辺に、関数 $a_G^E: E_G \rightarrow \text{pow}(\mathcal{A}) \setminus \{\emptyset\}$ により数値属性集合を割り当てたグラフである。ここで $s^E(G)$ は G の各辺から定量的アイテム集合を取り除いたグラフを表す。

2 つの辺数値属性付きグラフ $P_1 = (V_{P_1}, E_{P_1}, l_{P_1}, a_{P_1}^E)$ と $P_2 = (V_{P_2}, E_{P_2}, l_{P_2}, a_{P_2}^E)$ に対し、

- (1) $\forall v \in V_{P_1} [l_{P_1}(v) = l_{P_2}(f^E(v))]$
- (2) $\forall (u, v) \in E_{P_1} [l_{P_1}((u, v)) = l_{P_2}((f^E(u), f^E(v)))]$
- (3) $\forall (u, v) \in E_{P_1} [a_{P_1}^E((u, v)) \preceq a_{P_2}^E(((f^E(u), f^E(v))))]$

の 3 つの条件を満たす単射関数 $f^E: V_{P_1} \rightarrow V_{P_2}$ が存在するとき、 P_1 を P_2 の部分グラフと定義し、 $P_1 \subset P_2$ と表記する。また、 $P_1 \subset P_2$ かつ $P_2 \subset P_1$ が成り立つとき、 P_1 と P_2 は同型であると定義する。2 つの辺数値属性付きグラフ P と G に対し、 $P \subset G$ であれば、 P を G 中のパターンと呼ぶ。

$P \subset G$ なる 2 つの AG^E、 P 及び G と P と G の間で上記の (1)-(3) を満たす単射関数 f^E に対して、 P 中の各辺に対応する G 中の辺のリスト

$$o^E = [(u_1, v_1), (u_2, v_2), \dots, (u_{|E_P|}, v_{|E_P|})] \\ ((u_i, v_i) \in E_G, (u'_i, v'_i) \in E_P, (u_i, v_i) = (f^E(u'_i), f^E(v'_i)))$$

を G における P の辺出現と呼ぶ。ここで、 P 中の辺番号はラベル付きグラフ $s^E(P)$ が標準形となる訪問順で番号付けられているとし、 o^E 中の i 番目の辺 $o^E(i)$ は、 P の i 番目の辺に対応する G 中の辺を表す。また、 G における P の全辺出現を $\Phi_G^{P^E}$ と表す。

以上の準備の下、 G に対する P の支持度を

$$\text{sup}_G^E(P) = \min_{1 \leq i \leq |E_P|} |\{o^E(i) \mid o^E \in \Phi_G^{P^E}\}| / |E_G|$$

と定義する。つまり、 P の各辺 (u, v) に対し、 (u, v) に対応するすべての出現 $o^E \in \Phi_G^{P^E}$ 上の辺の数を求め、その最小値を P の支持度とする。なおこの定義は、単一のラベル付きグラフにおける支持度 [Bringmann 08] の計算を、頂点に対して行うのではなく、辺に対して適用

Algorithm FAG ^E -gSpan($G, \sigma, \mathcal{L}$)	
1: for each $p \in \mathcal{L}$	
2: $P := RME(p, \mathcal{L})$	
3: for each $p' \in P$	
4: FAG ^E -Enum($p', G, \sigma, \mathcal{L}$)	
Subroutine FAG ^E -Enum($p, G, \sigma, \mathcal{L}$)	
1: if $\neg \text{isCanG}(p) \vee \text{sup}_G^E(p) < \sigma$ then return	
2: $S := \text{QFI}(\{A^E(o) \mid o \in \Phi_G^{P^E}\}, \sigma)$	
3: $S := \{s' \in S \mid \text{isCanA}^E(s') \wedge \text{sup}_G^E(s') \geq \sigma\}$	
4: if $S = \emptyset$ then return	
5: for each $s' \in S$	
6: output s'	
7: $P := RME(p, \mathcal{L})$	
8: for each $p' \in P$	
9: FAG ^E -Enum($p', G, \sigma, \mathcal{L}$)	

図 8 FAG^E-gSpan の擬似コード

しており、グラフの特殊化に関して逆単調性 ($P \subset Q \rightarrow \text{sup}_G^E(P) \geq \text{sup}_G^E(Q)$) が成り立つ。

単一の辺数値属性付きグラフ G と利用者によって与えられた最小支持度 σ ($0 < \sigma \leq 1$) に対し、 $\text{sup}_G^E(P) \geq \sigma$ を満たすと共に、すべての辺に数値属性集合を含み、かつ連結である辺数値属性付きグラフ P を、頻出辺数値属性付きグラフパターン (以後、FAG^E) と定義する。FAG^V と同様、本論文では、与えられた G と σ に対し、FAG^E を抽出する手法について議論する。

4.2 辺数値属性付きグラフの代表元

AG^V 同様、AG^E に対しても、同型グラフが存在する。同型グラフの代表元のみを抽出することで、同型グラフの重複列挙を避けることができる。以下では、AG^E の代表元について議論する。

与えられた AG^E は、縦型探索による訪問順に辺に番号が付与されていることを仮定する。同型の AG^E の順序 (大小) 関係は、この番号順に辺に含まれる定量的アイテム集合を比較することで決定される。

2 つの同型な AG^E P_1 と P_2 に対し、辺の番号順に辺中の定量的アイテム集合を並べたリストを $L_1 = [a_1^1, \dots, a_1^{|\mathcal{E}_{P_1}|}]$, $L_2 = [a_2^1, \dots, a_2^{|\mathcal{E}_{P_2}|}]$ (a^k は k 番目の辺の定量的アイテム集合) とする。このとき P_1 と P_2 の順序関係は、3.3 節の順序関係と同様に求まる。すなわち、リスト内の各要素を前から順に比較していき、異なる要素の比較が行われた場合、小さい要素を含むパターンが小さいとする。また、全ての要素が同一である場合、 P_1 と P_2 は等しいとする。

この順序関係を基に、辺数値属性付きグラフ P が 2 つの条件、(1) ラベル付きグラフにおいて標準形 [Borgelt

表 1 実験で使用するデータセット

データ	H_1	H_2	H_3	D_4	D_5	D_6	D_7
頂点数	122	76	78	1755	6505	3009	1725
辺の数	176	125	111	2647	8716	3442	1940

05] である、かつ (2) P と同型かつ同一のラベル付きグラフを持つ AG^E の中で最小である、を満たすとき、 P を同型グラフの代表元であると定義する。

4.3 FAG^E-gSpan

本節では、FAG^E を発見するアルゴリズム FAG^E-gSpan を提案する。FAG^V-gSpan 同様、FAG^E-gSpan は、gSpan[Yan 02] によるグラフパターンの列挙と QFIMiner[Washio 05] により抽出された定量的アイテム集合の割り当てを繰り返すことで、FAG^E を抽出する。FAG^E-gSpan の擬似コードを図 8 に示す。

FAG^E-gSpan では、頂点のみからなるラベル付きグラフパターン p に対し新たな辺と頂点を追加し (図 8, FAG^E-gSpan の 2 行目, $RME(p, \mathcal{L})$)、辺を持つグラフパターン p' を生成する。すべての p' に対して FAG^E-Enum を繰り返し呼び出すことで、FAG^E を列挙することが可能である。FAG^E-Enum では、(属性がついていない) ラベル付きグラフパターン p に対し、標準形でない (図 8, FAG^E-Enum の 1 行目, $\neg \text{isCanG}(p)$)、もしくは頻出でない ($\text{sup}_G^E(p) < \sigma$) 場合は、それ以降の拡張を止める。枝刈りされなかったパターン p に対し、 p 中の各辺に定量的アイテム集合を割り当てることで、FAG^E の候補集合を得る (2 行目)。ここで $A^E(o)$ は、各辺出現 $o \in \Phi_G^{P^E}$ に対応する定量的アイテム集合のリストである。生成された FAG^E の候補集合中の各パターン $s' \in S$ に対し、標準形判定 (3 行目, $\text{isCanA}^E(s')$) と支持度判定 ($\text{sup}_G^E(s') \geq \sigma$) を行うことで、 s' が FAG^E であるかを判定し、FAG^E 集合 S を更新する。FAG^E である各 $s' \in S$ を出力した後、 p に対して最右拡張を適用し、FAG^E-Enum を呼び出すことで、より特殊な FAG^E を列挙する。

5. 実 験

提案手法の有効性を評価するために、Java 言語を用いて FAG^V-gSpan 及び FAG^E-gSpan を実装し、Windows マシン (CPU: Intel(R) Xeon(R) CPU X5260 3.3GHz, メインメモリ: 32GB) 上で通信履歴データ (haggle[Scott 06]*¹) 及び共著者ネットワークデータ (DBLP[Ley 02]*²) を用いて評価実験を行った。

通信履歴データは、モバイル機器を頂点、通信が行われたモバイル機器間を辺で結ぶことにより、図 1(iii) で示すように動的グラフとして表すことができる。このデータは、全体で約 3 日分 (250,000 秒) のデータを有して

*1 <http://crawdad.cs.dartmouth.edu/cambridge/haggle/imote/intel>
 *2 <http://www.informatik.uni-trier.de/~ley/db/>

表 2 通信履歴データに対する FAG^V-Span の実験結果

Δ_p	σ	time			pat.			cand.		
		H_1	H_2	H_3	H_1	H_2	H_3	H_1	H_2	H_3
0.05	0.3	<0.1	<0.1	0.1	0	1	2	6	11	24
	0.15	0.3	0.4	1	2	2	6	31	25	84
	0.1	369.3	38.7	4.5	55	13	6	803	437	113
0.10	0.3	<0.1	0.2	0.9	0	1	3	6	11	26
	0.15	1.3	1.1	4.3	4	2	4	53	25	74
	0.1	1409.2	-	-	81	-	-	1115	-	-

Δ_p : 属性 p の許容範囲. σ : 最小支持度. time: 実行時間 [秒].

pat.: 発見されたパターン数. cand.: 生成された候補パターン数.

いるが、実験では、一つの区間の時間幅を (1)40,000 秒、(2)20,000 秒及び (3)50,000 秒と変化させ、図 1(iii) に示したような動的グラフをそれぞれ作成した。各動的グラフにおいて、(1) 時刻 190,000 秒から 230,000 秒 (時間幅 40,000 秒) に対応するグラフを H_1 、(2) 時刻 120,000 秒から 140,000 秒 (時間幅 20,000 秒) に対応するグラフを H_2 、(3) 時刻 200,000 秒から 250,000 秒 (時間幅 50,000 秒) に対応するグラフを H_3 として準備した。また各頂点のラベルとして (概ねデバイスの種類に対応する) 各機器の MAC アドレスの先頭 3 バイトを与えた。なお、辺ラベルは考慮していない。

一方共著ネットワークは、著者を頂点に、共著論文があるときに、著者間に辺を結んだグラフである。本論文では、DBLP に含まれる 2004 年、1990 年、1985 年、1980 年の article からなるそれぞれのネットワーク D_4, D_5, D_6, D_7 を準備した。

5.1 FAG^V-gSpan の有効性の評価

第一の実験では、通信履歴データ (haggle) を用いて、 H_1, H_2, H_3 の各グラフに対し、グラフの各頂点に 2 種の頂点に関する中心性 (次数中心性及び近傍中心性) [Barabasi 02] を割り当てた AG^V を入力グラフとした。すなわち、各頂点は頂点の中心性を表現する 2 つの定量的アイテムを持つ。以下に、頂点集合 V_G 、辺集合 E_G からなるグラフ G における頂点 v_i の次数中心性 $c_{v_i}^d$ 、及び近傍中心性 $c_{v_i}^c$ の定義を与える。

$$c_{v_i}^d = \frac{|\{e | e \in V_E \text{ s.t. } e = (v_i, u) \vee e = (u, v_i)\}|}{|V_G| - 1}$$

$$c_{v_i}^c = \frac{|V_G| - 1}{\sum_{j \in \{1, \dots, |V_G|\}} d_{ij}}$$

ここで d_{ij} は v_i から v_j への最短パスの長さとする。また、グラフ G が非連結であるときは、各弱連結成分毎に中心性を計算した。

実験は、最小支持度 σ と許容範囲 Δ_p を変化させて、実行時間 (time)、発見されたパターン数 (pat.)、そして生成した候補パターン数 (cand.) を測定した。パラメータは $MinPts = 3$ と設定する。実験結果を表 2 に示す。表中の ‘-’ はメモリオーバーフローにより結果が得られな

図 9 通信履歴データから抽出されたパターンの例

かったことを示す。また候補パターンとは、FAG^V に加え、最右拡張により生成されたが最小支持度条件を満たさないラベル付きグラフパターン、標準形を満たさなかったラベル付きグラフパターン、出現集合から高密度クラスタが得られなかったラベル付きグラフパターン、標準形を満たす形で高密度クラスタが割り当てられなかったラベル付きグラフパターン、高密度クラスタを割り当てたが定量的アイテム集合を持たない頂点を含む数値属性付きグラフパターン、及び属性付きグラフの標準形を満たさなかった数値属性付きグラフパターンを表す。 σ の減少に伴い、実行時間、発見されたパターン数、候補パターン数は増加する。

第二の実験では、各頂点中の定量的アイテムの数の影響を評価する。 H_i ($i = \{1, 2, 3\}$) を含む各動的グラフを用い、 H_i より過去のグラフの頂点の中心性を、 H_i の対応する頂点に割り当てることで、複数の数値属性を持つグラフを準備する。つまり、 H_i の各頂点には、過去のある時間幅におけるグラフの対応する頂点の中心性に相当する定量的アイテム集合が割り当てられる。

この実験では、 $MinPts = 3$ 、 $\sigma = 0.15$ 、 $\Delta_p = 0.05$ と設定した。実験結果を表 3 に示す。

実験結果より、属性数の増加に伴い、発見されるパターン数も増加していることが分かる。さらに属性数が多くなるほど、発見されるパターン数に比べ、候補数が非常に多くなる傾向がある。得られたパターンの例を図 9 に示す。なお、図中における各定量的アイテムの属性名は、最初の数値が時間幅の id を、続く記号が中心性の種類 (c: 近傍中心性, d: 次数中心性) をそれぞれ表す。例え

表 3 FAG^V-gSpan に対する属性数の影響

A	time			pat.			cand.		
	H ₁	H ₂	H ₃	H ₁	H ₂	H ₃	H ₁	H ₂	H ₃
2	0.3	0.3	1.0	2	2	6	31	25	84
4	1.0	1.6	245.9	3	5	6	37	65	92
6	1.5	4.4	283.2	14	14	33	81	77	181
8	1.6	5.1	490.3	23	25	71	93	134	293
10	2.6	13.3	4,672.3	50	48	478	142	222	1370

|A|: グラフ中の各頂点の属性数 .
time: 実行時間 [秒] . pat.: 発見されたパターンの数.
cand.: 生成された候補パターン数 .

ば、定量的アイテム $\langle 1c: [0.54, 0.69] \rangle$ は、第一時間幅における近傍中心性が 0.54 から 0.69 の範囲であることを表す。得られたパターン P_1 は閉路を含み、かつ各頂点は同じ定量的アイテム集合を含んでいるため、ある時間幅における強い関係を持ったモバイル機器の集合を表していると考えられる。一方 P_2 は、2 つの頂点が複数の異なる時間幅に対する定量的アイテムを含んでいるので、 P_2 は、動的グラフの時間的変化に関係したパターンであると考えられる。これらの例が示す通り、提案手法では複数種の数値属性を含むパターンが生成される。パターンの外部構造に加え、各頂点に含まれる属性の意味や関係、またその値を考察することで、マイニング対象に関するより深い知見が得られることが期待できる。

次に、FAG^V-gSpan により、「豊かな情報を持つパターン」、すなわち複数種の属性を含むパターンの獲得がどの程度達成できているのかを定量的に考察する。実験 2 でも利用した、 H_3 を元に生成された各頂点に 10 種 (2 種類 $\times$ 5 区間) の定量的アイテム集合を持つグラフを対象に、 $\sigma = 0.15$, $\Delta_p = 0.05$ の条件で、得られた各パターンが持つ属性種数とパターンの関係を表 5 にまとめた。FAG^V-gSpan は極大な定量的アイテム集合を割り当てるため、属性の種類数が多いパターンが数多く抽出されている。このように複数の属性がお互い関連性を持って存在する場合もあるため、属性を独立して利用するのではなく同時に扱うことは対象が持つ情報を効率良く利用していると言える。既存のラベル付きグラフマイナーでは 1 種類のラベルに対するグラフパターンを抽出することを目的としている。そのため複数の異なる属性を持つグラフから属性間の関連性を分析するには、それぞれの属性に対するパターンを抽出し、属性の組み合わせ毎に得られたパターンを比較していく方法が考えられる。しかしながらこの方法では、複数の属性が必ずしも同時に現れるかどうかを判断するのは難しい。一方で FAG^V-gSpan は異なる属性すべてに対し高密度な定量的アイテム集合を持ったパターンを抽出することを目的とし、既存の手法では困難であった複数の異なる属性を同時に扱うことができる。

表 4 共著者ネットワークデータに対する FAG^V-Span の実験結果

T	σ	time		pat.		cand.	
		D_4	D_5	D_4	D_5	D_4	D_5
20	0.3	0.6	6.0	0	0	0	0
	0.15	1.5	21.0	16	19	46	57
	0.1	1.5	22.1	20	19	61	58
10	0.3	5.4	33.6	17	9	21	20
	0.15	136.6	339.7	54	55	143	187
	0.1	678.3	512.2	204	79	574	430

|T|: トピック数. σ : 最小支持度. time: 実行時間 [秒].
pat.: 発見されたパターン数.
cand.: 生成された候補パターン数.

表 5 発見されたパターンの頂点数・属性種数毎の集計

$ V \setminus A $	1	2	3	4	5	6	7	8	9
1	1	0	1	2	1	4	3	1	0
2	1	0	7	26	36	40	36	20	1
3	0	2	4	15	68	103	43	8	1
4	0	3	6	8	23	14	0	0	0

|V|: パターンの頂点数 .
|A|: パターンに含まれる数値属性種数

これらの結果から、FAG^V-gSpan によって、従来のアルゴリズムで取得できなかった複雑なパターンを発見できる、ということを確認できる。

第三の実験では、共著ネットワーク (DBLP) を利用して、FAG^V-gSpan の規模に対する頑健性を示す。各頂点には、同一のラベルを与えるとともに、論文タイトル (複数の論文がある場合は、すべてを連結したもの) に対してトピックモデル (LDA) [Blei 03, Griffiths 04] を適用することにより得られるトピック分布を定量的アイテム集合として割り当てる。ただし、この分布における確率値が 0.1 に満たないトピックは省略している。なお、トピック分布について仮定する対称ディリクレ事前分布の超パラメタ α の値は、50/トピック数とした [Griffiths 04]。実験は、表 1 の D_4, D_5 を使用し、トピック数 (頂点に割り当てられる最大属性数) と最小支持度 σ を変化させて、実行時間 (単位は秒)、発見されたパターン数、そして生成した候補パターン数を測定した。なお、 $\Delta_p = 0.05$, $MinPts$ は頂点数 $|V|$ の 0.5% ($MinPts = |V| \times 0.005$) と設定した。実験結果を表 4 に示す。

実験結果より、比較的大きいグラフに対しても現実的な実行時間でパターンが抽出されていることが分かる。また表 2 同様、 σ の減少に伴い、実行時間、発見されるパターン数、そして候補パターン数の増加が認められる。一方、属性数の減少に伴い、実行時間、発見されるパターン数、および候補パターン数が増加する。これは、トピック数を大きくすることにより、抽象度が低く識別性の高い数値属性が得られ、結果としてパターンが発見されにくくなったことによるものだと考えられる。

表 6 共著者ネットワークデータにおける FAG^E -gSpan の実験結果

σ	time				pat.				cand.			
	D_4	D_5	D_6	D_7	D_4	D_5	D_6	D_7	D_4	D_5	D_6	D_7
0.3	1.5	27.0	2.4	0.8	0	2	1	0	1	4	3	1
0.15	9.0	346.7	18.4	3.0	19	27	22	12	35	80	36	27
0.1	39.6	3650.7	34.8	5.3	39	62	35	22	87	224	75	60

 σ : 最小支持度. time: 実行時間 [秒].

pat.: 発見されたパターン数. cand.: 生成された候補パターン数

図 10 ‘preprocessing’ と ‘refineMinMax’ 手法によって抽出されたパターンの例

5.2 FAG^E -gSpan の有用性の評価

FAG^E -gSpan の有用性を評価するために、辺に数値属性が割り当てられたグラフとして、共著ネットワーク (DBLP) を利用した実験を行った。グラフの各頂点及び辺には同一のラベルを与えると同時に、各辺に対し、辺で結ばれている著者同士の共著論文タイトルを連結した文字列 (複数論文がある場合は、全てを連結したもの) に対してトピック分布を推定し、それを定量的アイテム集合として割り当てた。この際、先ほどの実験同様、この分布における確率値が 0.1 に満たないトピックは定量的アイテム集合から排除している。実験は、表 1 の D_4, D_5, D_6, D_7 を使用し、最小支持度 σ を変化させて、実行時間 (単位は秒)、発見されたパターン数、そして生成した候補パターン数を測定した。パラメータは、 $\Delta_p = 0.05$, $MinPts$ は各グラフの頂点数 $|V|$ の 0.5% ($MinPts = |V| \times 0.005$) とする。また、トピック数は 10 とし、前節と同様に超パラメータ α の値を設定した。実験結果を表 6 に示す。

実験結果より、 σ が減少すると、 FAG^E -gSpan の実験同様、実行時間、発見されるパターン数、候補パターン数は増加するということが分かる。この実験により FAG^E -gSpan もある程度大きなデータに対しても現実的な時間で結果を得ることができることが示された。

5.3 FAG^V -gSpan を利用したマイニング方法の比較

FAG^V -gSpan はパターンの構造 (ラベル付きグラフ) を決定した後で、グラフパターンの出現によるクラスタリングを行うことで定量的アイテム集合を取得するが、前処理として定量的アイテム集合を用意しておく方法も考えられる。そこで、 FAG^V の定量的アイテム集合を取得するために以下の 3 つの手法を準備し、 FAG^V -gSpan を改良して、得られる FAG^V の数や実行時間を比較する。

(1) ‘preprocessing’: 前処理として、グラフ中のすべて

表 7 前処理手法の比較結果

$ A $	time[sec]				pat.			
	W_1	W_2	W_3	W_4	W_1	W_2	W_3	W_4
2	1.0	0	0	0	6	0	0	0
4	245.9	0	0	0.1	6	2	2	1
6	283.2	<0.1	<0.1	0.2	33	5	5	3
8	490.3	<0.1	<0.1	0.2	71	5	5	3
10	4,672.3	<0.1	<0.1	0.2	478	14	14	13

 $|A|$: グラフ中の各頂点の属性数.

time: 実行時間 [秒]. pat.: 発見されたパターン数.

 W_1 : FAG^V -gSpan. W_2 : ‘preprocessing’. W_3 : ‘refineMinMax’. W_4 : ‘re-clustering’.

の頂点を利用して、極大な定量的アイテム集合パターンの集合を生成する。入力グラフに付与された定量的アイテム集合の代わりに、前処理で得られたパターンを対応する頂点の定量的アイテム集合として割り当てる。この際、定量的アイテム集合パターンが割り当てられなかった頂点とその頂点を一端とする辺をグラフから削除する。このように再構成されたグラフから、ラベル付きグラフパターンを列挙していく。得られたグラフパターンに対し、出現によるクラスタリングを行わず、前処理で得られた定量的アイテム集合パターンをそのまま定量的アイテム集合として利用し、 FAG^V を生成する。

- (2) ‘refineMinMax’: ‘preprocessing’ 同様の方法で FAG^V を生成する。その後、パターンの出現をもとに、各頂点の定量的アイテム集合の最小値と最大値を求め、パターン中の定量的アイテム集合の区間幅を訂正する。
- (3) ‘re-clustering’: ‘preprocessing’ によるグラフパターン抽出後、出現を使用して、再度クラスタリングを実行し、高密度クラスターの定量的アイテム集合を割り当てた FAG^V を抽出する。

比較実験は、グラフデータ D_3 ($MinPts = 3$, $\sigma = 0.15$, $\Delta_p = 0.05$) を用いて行った。実験結果を表 7 に示す。実験結果より、 FAG^V -gSpan は最も多く実行時間がかかっていることが分かる。これはグラフ構造が抽出されるたびに、繰り返しクラスタリングを行う必要があるからである。一方、‘preprocessing’, ‘refineMinMax’, ‘re-clustering’ によって得られるパターン数は数少ない。これは、これらの手法が特徴的なパターンを十分に抽出できなかったことを表している。表 7 において、‘preprocessing’

と ‘refineMinMax’ 実験の結果は同じである。しかし、図 10 に示すように、得られたパターンは異なっている。‘reclustering’ のパターン数は ‘preprocessing’ より少ない。これは、‘preprocessing’ (前処理) によって抽出されたアイテム集合中に、高密度クラスタが存在しなかったからだと考えられる。以上の考察より、 FAG^V -gSpan は、頂点数値属性付きグラフから頂点の内部情報を考慮した特徴的なパターンを抽出するのに、実用的であると言える。

6. 関 連 研 究

近年、多くのグラフマイニング手法が提案されている [Inokuchi 00, Kuramochi 01, Kuramochi 05, Yan 02]。しかしながら、これらの手法のほとんどは、頂点や辺の内部情報を考慮しないため、 FAG -gSpan によって抽出される定量的アイテム集合が付与された特徴的なグラフパターンを発見することができない。一方で、頂点がアイテム集合からなるグラフを対象としたパターン発見手法がいくつか提案されている [Chen 04, 山本 08]。DAG-Miner [Chen 04] は、頂点にアイテム集合を含む有向非巡回グラフ (DAG) データベースから、頻出パターンを発見する手法である。FMG [山本 08] は、頂点に複数の構造を持った複合構造グラフデータベースから、頻出パターンを発見する手法である。これら手法は、頂点にアイテム集合や系列データを含むグラフを扱っているが、数値情報が付与されたグラフを直接的に扱うことはできない。一方、頂点に数値属性を含むグラフを対象とした手法としては、頂点に座標情報を持つグラフデータベースからの頻出パターン発見手法 gFSG [Kuramochi 02] が提案されている。また同様に、3 次元グラフを対象とした手法も提案されている [Arimura 07]。CoPaM [Moser 09] は、特徴空間において類似した頂点集合からなる高密度部分グラフを抽出する手法である。一方、 FAG^V -gSpan は高密度クラスタからなる定量的アイテム集合を付与した頻出グラフパターンを抽出する。また、筆者の知る限り、辺に複数の情報を持つグラフを対象としたパターン発見手法は提案されておらず、その点で、 FAG^E -gSpan の新規性は高いと考えている。

7. ま と め

本論文では、グラフ構造の列挙と定量的アイテム集合発見の技術を有機的に結合し、頂点や辺に属性が付与されたグラフを対象とした頻出パターン発見手法 FAG^V -gSpan および FAG^E -gSpan を提案した。またいくつかの実験により提案手法の有効性を検証した。

今後の課題としては、提案手法の理論的な解析や、より大規模データを用いた評価実験があげられる。特に大規模データを用いた実験では、メモリーオーバーフローなどを回避するとともに、より現実的なパターン発見に焦

点をあてるために、発見すべきパターンに対する制約の導入や、支持度の高いパターンから優先的に拡張する [Ke 09] などの探索の制御が必要となると考えられる。

ところで、 FAG -gSpan では、グラフマイナーと定量的アイテム集合のクラスタリングを独立に行っている。このことに着目した今後の課題のひとつとして、他のグラフマイナーとクラスタリング手法の組み合わせでパターン抽出を行い、得られたパターンを比較することを試みることが挙げられる。また、時系列を考慮した動的グラフへの応用 [Borgwardt 06] やグラフデータベースへの拡張なども今後の重要な課題である。

謝 辞

有益なご指摘とご指導を賜りました査読者の方々に深く感謝いたします。本研究の一部は、文部科学省科学研究費補助金 (若手研究 (B): 課題番号 21700168 及び基盤研究 (B): 課題番号 20300038) による。

◇ 参 考 文 献 ◇

- [Arimura 07] Arimura, H., Uno, T., and Shimozone, S.: Time and Space Efficient Discovery of Maximal Geometric Graphs, *Proceedings of the 10th International Conference on Discovery Science*, pp. 42-55 (2007)
- [Avis 93] Avis, D. and Fukuda, K.: Reverse Search for Enumeration, *Discrete Applied Mathematics*, Vol. 65, pp. 21-46 (1993)
- [Barabasi 02] Barabasi, A.-L.: *Linked: The New Science of Networks*, Basic Books (2002)
- [Blei 03] Blei, D. M., Ng, A. Y., and Jordan, M. I.: Latent dirichlet allocation, *Journal of Machine Learning Research*, Vol. 3, pp. 993-1022 (2003)
- [Borgelt 05] Borgelt, C.: On Canonical Forms for Frequent Graph Mining, *Working Notes of the Third International ECML/PKDD-Workshop on Mining Graphs*, pp. 1-12 (2005)
- [Borgwardt 06] Borgwardt, K. M., Kriegel, H.-P., and Wackerreuther, P.: Pattern Mining in Frequent Dynamic Subgraphs, *Proceedings of the 6th International Conference on Data Mining*, pp. 818-822 (2006)
- [Bringmann 08] Bringmann, B. and Nijssen, S.: What Is Frequent in a Single Graph?, *Proceedings of the 12th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 858-863 (2008)
- [Chen 04] Chen, Y.-L., Kao, H.-P. and Ko, M.-T.: Mining DAG Patterns from DAG Databases, *Proceedings of the 5th International Conference on Advances in Web-Age Information Management*, pp. 579-588 (2004)
- [Griffiths 04] Griffiths, T. L. and Steyvers, M.: Finding Scientific Topics, *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 101, pp. 5228-5235 (2004).
- [Inokuchi 00] Inokuchi, A., Washio, T., and Motoda, H.: An Apriori-based Algorithm for Mining Frequent Substructures from Graph Data, *Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery*, pp. 13-23 (2000)
- [Ke 09] Ke, Y., Cheng, J., and Yu, J. X.: Top-k Correlative Graph Mining, *Proceedings of 9th the SIAM International Conference on Data Mining*, pp. 1038-1049 (2009)
- [Kuramochi 01] Kuramochi, M. and Karypis, G.: Frequent Subgraph Discovery, *Proceedings of the 2001 IEEE International Conference on Data Mining*, pp. 313-320 (2001)
- [Kuramochi 02] Kuramochi, M. and Karypis, G.: Discovering Frequent Geometric Subgraphs, *Proceedings of the 2002 IEEE International Conference on Data Mining*, pp. 258-265 (2002)
- [Kuramochi 05] Kuramochi, M. and Karypis, G.: Finding Frequent Patterns in a Large Sparse Graph, *Data Mining and Knowledge Dis-*

- covery, Vol. 11, No. 3, pp. 243–271 (2005)
- [Lewis 09] Lewis, T. G.: Network Science: Theory and Applications, WILEY (2009)
- [Ley 02] Ley, M.: The DBLP Computer Science Bibliography: Evolution, Research Issues, Perspectives, *Proceedings of the 9th International Symposium on String Processing and Information Retrieval*, pp. 1–10 (2002)
- [Moser 09] Moser, F., Colak, R., Rafiey, A. and Ester, M.: Mining Cohesive Patterns from Graphs with Feature Vectors, *Proceedings of the 9th SIAM Conference on Data Mining*, pp. 593–604 (2009)
- [Newman 03] Newman, M. E. J.: The Structure and Function of Complex Networks, *SIAM Review*, Vol. 45, pp. 167–256 (2003)
- [Scott 06] Scott, J., Gass, R., Crowcroft, J., Hui, P., Diot, C. and Chaintreau, A.: CRAWDAD trace cambridge/haggle/imote/intel (v. 2006-01-31) (2006)
- [Washio 03] Washio, T. and Motoda, H.: State of the Art of Graph-based Data Mining, *SIGKDD Explorations*, Vol. 5, No. 1, pp. 59–68 (2003)
- [Washio 04] Washio, T., De Raedt, L., and Kok, J. N.: Advances in Mining Graphs, Trees and Sequences: Preface *Fundamenta Informaticae*, Vol. 66, No. 1-2, pp. 5–8 (2004)
- [Washio 05] Washio, T., Mitsunaga, Y., and Motoda, H.: Mining Quantitative Frequent Itemsets Using Adaptive Density-Based Subspace Clustering, *Proceedings of the 5th IEEE International Conference on Data Mining*, pp. 793–796 (2005)
- [山本 08] 山本 翼, 尾崎 知伸, 大川 剛直: 構造データ集合からなるグラフデータベースからの頻出パターン発見, 情報処理学会論文誌: データベース, Vol.1, No.1, pp.26–35 (2008)
- [Yan 02] Yan, X. and Han, J.: gSpan: Graph-Based Substructure Pattern Mining, *Proceedings of the 2002 IEEE International Conference on Data Mining*, pp. 721–724 (2002)
- [安田 01] 安田 雪: 実践ネットワーク分析 - 関係を解く理論と技法, 新曜社 (2001)

〔担当委員: 矢田 勝俊〕

2010 年 9 月 8 日 受理

著者紹介

三好 裕樹

2009 年神戸大学工学部情報知能工学科卒業。現在, 神戸大学大学院工学研究科所属。構造データマイニングに関する研究を行っている。

尾崎 知伸 (正会員)

1996 年慶應義塾大学総合政策学部卒業。1998 年同大学大学院政策・メディア研究科前期修士課程修了。2001 年同研究科博士課程単位取得退学。慶應義塾大学講師。神戸大学助手, 助教を経て, 2010 年より大阪大学サイバーメディアセンター特任講師。博士 (政策・メディア)。帰納論理プログラミング, 構造データマイニングなどの研究に従事。

江口 浩二 (正会員)

神戸大学大学院システム情報学研究科准教授。国立情報学研究所客員准教授。博士 (工学)。情報検索, Web 情報処理, データマイニングの研究に従事。

大川 剛直 (正会員)

1986 年大阪大学工学部通信工学科卒業。1988 年同大学大学院工学研究科通信工学専攻博士前期課程修了。大阪大学助手, 講師, 助教授を経て, 2005 年神戸大学大学院自然科学研究科教授。現在, 神戸大学大学院システム情報学研究科教授。博士 (工学)。知的データ処理, バイオインフォマティクスなどの研究に従事。IEEE, 情報処理学会, 電子情報通信学会, 電気学会などの会員。