情報処理学会論文誌　データベース　Vol. 3　No. 2　1–12 (June 2010)

# 内部および外部重みを考慮した頻出部分グラフマイニング

信田正樹[†1]　尾崎知伸[†2]　大川剛直[†3]

近年，グラフデータの増大にともない，そこから何らかの意味のあるパターンや情報を発見するグラフマイニング手法に関する研究がさかんに行われている．本論文では，グラフマイニングの 1 つの発展として，グラフ自身およびグラフの各構成要素に対し，その重要性や信頼性，意義などを表す重みが付与された，外部および内部の重み付きグラフからのパターン発見について議論する．具体的には，重みに着目したパターンの重要性尺度として，一般重み付き頻度（GWF），および制約付き重み付き頻度（CWF）の 2 つを考案するとともに，GWF および CWF に関して高い重要性を示す部分グラフを発見する効率的なアルゴリズム，GWF-mine および CWF-mine をそれぞれ提案する．

# Weighted Frequent Subgraph Mining in Weighted Graph Databases

Masaki Shinoda,[†1] Tomonobu Ozaki[†2] and Takenao Ohkawa[†3]

Recently, graph-structured data is becoming popular in many application domains, and several studies on graph mining have been performed for discovering useful knowledge from graph databases. In this paper, in order to realize more precise knowledge discovery in graph databases, we focus on pattern discovery problems from externally and internally weighted graphs where external weight represents a degree of importance and reliability of a graph itself and internal weight reflects utility and significance of each component in a graph. By using external and internal weights, we propose two importance measures named (1) general weighted frequency (GWF) and (2) external weighted frequency under the constraint of internal weight (CWF), and develop efficient algorithms GWF-mine and CWF-mine for extracting subgraphs having high value of these measures. Experimental results by using synthetic and real world datasets show the effectiveness of the proposed framework.

## 1. はじめに

グラフは，複雑な構造を表現するのに適したデータ表現であり，近年，社会ネットワーク分析やバイオインフォマティクス分野などを中心に幅広く利用されている．また，蓄積された大量のグラフデータから有益な知識や知見を得るために，頻出パターン発見を中心にグラフマイニング[5),21)] に関する研究が積極的に行われている．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，対象のより柔軟な表現手段として重み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．

例として，グラフマイニングによるウェブアクセスログ分析問題について考えてみよう．この場合，各ユーザセッションは，ウェブページを頂点，その遷移を辺とするグラフとして表現される．また追加的な情報として，各ページの総閲覧時間やページ間の遷移回数を考えることができる．これらの情報を適切に正規化し，前者は頂点，後者は辺に対する重要性と見なすことで，各セッションは，単なるグラフではなく，重み付き（重要性付き）グラフとして表現される．本論文では，頂点もしくは辺に付与される重み（重要性）を内部重みと呼ぶ．一方，グラフ自身の重要性として，グラフの外部重みを考えることも可能である．たとえば，ウェブアクセスログ分析問題では，新しいセッションは最近の状況を表しているので古いセッションと比較して重要であるなど，セッションの生成時期をその重要性（重み）として考えることができる．

このほかにも，重み付きグラフによって自然に表現できるデータの例として，生物学ネットワークやソフトウェアの挙動を表すコールグラフなどが考えられる．生物学ネットワークデータは，実際の生物学的な実験による方法に加え，計算機シミュレーションなどによって決定されることもあると考えられる．したがって，その生成過程の差異を外部重みとして表現することで，データの信憑性を表現することが可能である．また内部重みについても，化学的生物学的性質に基づいてその重要性を決定することが可能である．一方，ソフトウェ

†1 神戸大学工学部
Faculty of Engineering, Kobe University
†2 大阪大学サイバーメディアセンター
Cybermedia Center, Osaka University
†3 神戸大学大学院システム情報学研究科
Graduate School of System Informatics, Kobe University

ア解析におけるコールグラフもまた，重み付きグラフにより自然に表現される例の 1 つである．コールグラフとは，各関数（モジュール）を頂点，その呼び出し関係を辺とするグラフであるが，たとえば，その呼び出し回数などを内部重み，ソフトウェアの規模や信頼性，有用性を外部重みとしてとらえることが可能である．

以上の例からも分かるように，外部および内部重みは，対象とするデータをモデル化するうえで非常に重要な役割を果たす．したがって，これらの重みを積極的に利用することで，より精密かつ有意味な知識やパターンの発見を期待することができる．

本論文ではこれらのことに着目し，グラフマイニングの 1 つの拡張として，外部および内部重み付きグラフデータベースからのパターン発見について議論する．より詳細には，内部重みはグラフ中の各構成要素の重要性を，外部重みはグラフ自身の重要性をそれぞれ表すと仮定し，これらの重みに関する頻度の組合せとしてパターンに対する 2 つの頻度，(1) 一般重み付き頻度および，(2) 制約付き重み付き頻度を考案する．また，これらの頻度に対し高い値を示すパターンを発見する新たなデータマイニング問題を設定するとともに，そのデータマイニング問題に対する効率的なアルゴリズム GWF-mine および CWF-mine を提案する．なお詳しくは後述するが，一般に頻出パターン発見に対する効率的なアルゴリズム構築において，逆単調性と呼ばれる性質は非常に重要となる．しかし，本論文で提案する 2 つの頻度は逆単調性を満たさない．そこで GWF-mine および CWF-mine では，頻度そのものではなく，その上界値を用いることでこの問題の解決を図っている．

以下に本論文の構成を示す．まず 2 章では，準備としていくつかの記法と定義を導入する．また，本論文におけるデータマイニング問題の形式的な定義を与える．次に 3 章で，提起したデータマイニング問題に対するアルゴリズム GWF-mine および CWF-mine を提案し，その詳細を説明する．4 章で関連研究について述べる．5 章で実験結果について説明した後，最後に 6 章で結論と今後の課題を述べる．

## 2. 準　　備

本章では，基本的な記法と定義を与えた後，本論文で扱うデータマイニング問題を導入する．また，定義などの直感的理解を助けるため，**図 1** に示す例を用いる．

**重み付きグラフ**を $g = (V_g, E_g, l_g, w_g, ew(g))$ と表記する．ここで，$V_g$ は頂点集合，$E_g \subseteq V_g \times V_g$ は辺集合，$l_g : V_g \cup E_g \to \mathcal{L}$ は頂点および辺にラベル集合 $\mathcal{L}$ 上のラベルを割り当てるラベル関数である．また，$ew(g) \in \mathcal{R}^+$ は $d$ の**外部重み**を表す非負の実数，$w_g : V_g \cup E_g \to \mathcal{R}^+$ は $g$ 中の各頂点と辺に**内部重み**である非負の実数を割り当てる関数である．$g$ に対しその内部重みの総計を，$iw(g) = \sum_{ve \in V_g \cup E_g} w_g(ve)$ と表記する．図 1 中のグラフ $g1$，$g2$，$g3$ は重み付きグラフの例である．なお，括弧内の数値が内部重みを表す．また簡略化のために，辺ラベルは考慮せず，さらにその内部重みは 0 とする．

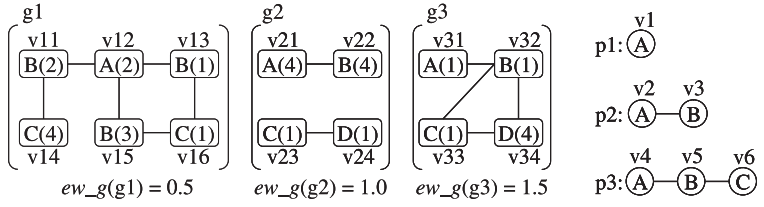



**図 1**　重み付きグラフ（g1–g3）およびラベル付きグラフ（p1–p3）の例
Fig. 1　Examples of weighted graphs (g1–g3) and labeled graphs (p1–p3).

重み付きグラフデータベース $D$ から抽出するパターンとして，頂点集合 $V_p$，辺集合 $E_p$，ラベル関数 $l_p : V_p \cup E_p \to \mathcal{L}$ からなる**ラベル付きグラフパターン** $p = (V_p, E_p, l_p)$ を考える．重み付きグラフ $g \in D$ とラベル付きグラフパターン $p$ に対し，以下の条件 (1) と (2) を満たす単射関数 $f : V_p \to V_g$ が存在するとき，$p$ は $g$ の部分グラフであるといい，$p \preceq g$ と表記する．(1) $\forall u \in V_p\,[\,l_p(u) = l_g(f(u))\,]$，(2) $\forall (u,v) \in E_p\,[\,\exists (f(u), f(v)) \in E_g\,(\,l_p(u,v) = l_g(f(u), f(v))\,)\,]$．なお，2 つのパターン $p$ と $q$ に対して上記の条件を満たす単射関数が存在するときも，同様に $p \preceq q$ と表記する．たとえば図 1 において，2 つのグラフ $p2 = (\{v2, v3\}, \{(v2, v3)\}, \{v2 \to A, v3 \to B\})$，$p3 = (\{v4, v5, v6\}, \{(v4, v5), (v5, v6)\}, \{v4 \to A, v5 \to B, v6 \to C\})$ に対して，関数 $\{v2 \to v4, v3 \to v5\}$ が上記の条件を満たすので $p2 \preceq p3$ が成り立つ．

パターン $p$ と重み付きグラフ $g$ に関して，上記の条件を満たす単射関数 $f$ の集合を $F(p, g)$ と表記する．$f \in F(p, g)$ に関する $g$ 中の $p$ の**出現**を，重み付き部分グラフ $e(p, g, f) = (\{f(u) \in V_g \,|\, u \in V_p\}, \{(f(u), f(v)) \in E_g \,|\, (u, v) \in E_p\}, l_g, w_g, ew(g))$ と定義する．また，$g$ における $p$ の出現の全体集合を $E(p, g) = \{e(p, g, f) \,|\, f \in F(p, g)\}$ と定義する．図 1 において，$E(p2, p3)$ はただ 1 つのグラフ $(\{v4, v5\}, \{(v4, v5)\}, \{v4 \to A, v5 \to B, v6 \to C\})$ を含む集合である．

ここで，重み付きグラフデータベース $D = \{g_1, \cdots, g_n\}$ における，ラベル付き部分グラフパターン $p$ の頻度に関する一連の定義を導入する．まず文献 15) などと同様，外部重みを考慮した頻度として，**外部重み付き頻度**（External Weighted Frequency）を定義する．以降これを EWF と略記する．

**定義 1**　$D$ における $p$ の EWF は，$D$ 中のグラフの総外部重みに対する $p \preceq g$ なるグラフ $g \in D$ の総外部重みの割合であり，以下のように定義される．

$$ewf(p, D) = \sum_{g \in D,\, p \preceq g} ew(g) \,/\, \sum_{d \in D} ew(d)$$

なお，その定義により $0 \leq ewf(p, D) \leq 1$ となる．

次に，内部重みを考慮した頻度である，**内部重み付き頻度**（Internal Weighted Frequency）を提案する．以降これを IWF と略記する．

**定義 2**　$D$ における $p$ の IWF を，$D$ 中のグラフの総内部重みに対する $p \preceq g$ なるグラフ $g \in D$ の最重出現の総内部重みの割合と定義する．定義より，$0 \leq iwf(p, D) \leq 1$ である．

$$iwf(p, D) = \sum_{g \in D,\, p \preceq g} max\_iw(p, g) \,/\, \sum_{d \in D} iw(d), \text{ where } max\_iw(p, g) = \max_{e \in E(p,g)} iw(e)$$

ここで $max\_iw(p, g)$ は，$g$ 中の $p$ の最重出現の内部重みを表す．本論文では，IWF 計算に最重出現を採用しているが，これは，(1) EWF 同様，出現数に依存せず出現の有無を基準とし，(2) より大きな重みを持つものが重要であると考えるためである．また，各出現の内部重みの合計などを利用することも考えられるが，(3) 小さなパターンに対して（不当に）高い評価を与えることを避けたいことに加え，単一グラフにおける頻度定義に関する議論[2),10)] とも密接に関連し，(4) 出現の重なりをどう考えるかということに対し適切な意味づけが必ずしも容易ではない，ということも最重出現を採用する大きな理由である．どのような定義を採用するかは応用の目的に依存する．本論文で提案する枠組みは，IWF の定義の変更に対し比較的容易に対応できると考えているが，応用に対して適切な IWF 基準自体を開発することは，今後の大きな課題である．

図 1 を対象に，EWF と IWF の計算例を示す．$D = \{g1, g2, g3\}$ としたとき，$p3 \preceq g1$，$p3 \npreceq g2$，$p3 \preceq g3$ より，$ewf(p3, D) = (0.5 + 1.5)/(0.5 + 1.0 + 1.5) \simeq 0.67$ である．一方 $max\_iw(p3, g1) = \max\{8, 6, 4\}$，$max\_iw(p3, g3) = \max\{3\}$ より，$iwf(p3, D) = (8 + 3)/(13 + 10 + 7) \simeq 0.37$ となる．

本論文では，重み付きグラフにおける部分グラフパターンの重要性尺度として，外部重みと内部重みを組み合わせて得られる 2 種類の頻度を提案する．

第 1 の尺度は EWF と IWF の加重平均である．

**定義 3**　$D$ 中の $p$ の**一般重み付き頻度**（General Weighted Frequency，以降 GWF）を，次のように定義する．$GWF(p, D, \lambda) = \lambda ewf(p, D) + (1 - \lambda) iwf(p, D)$.

$\lambda$ は，EWF と IWF の影響の割合を調整するするパラメータである．$\lambda = 1$ のとき，GWF は EWF と等しくなる（$GWF(p, D, 1) = ewf(p, D)$）．一方 $\lambda = 0$ のとき，GWF は IWF と等しくなる（$GWF(p, D, 0) = iwf(p, D)$）．また，たとえば図 1 において $\lambda = 0.5$ とすると，$GWF(p3, \{g1, g2, g3\}, 0.5) = 0.5 \times (2/3) + 0.5 \times (11/30) \simeq 0.52$ となる．

パターンに対し，EWF は重要なグラフに現れていることを表し，IWF は重要な部分として現れていることを意味する．GWF において $\lambda$ を用いてこれらの加重平均をとることで，片方が多少小さくても他方が大きければ，特徴的なパターンとして扱うことが可能となる．たとえば 1 章で述べたグラフマイニングによるウェブアクセスログ解析では，EWF は大きくないが IWF が大きい場合として，「最近の閲覧者は必ずしも多くはないが，じっくりとした閲覧をともなうページの遷移」など，数だけではとらえにくい傾向などが把握しやすくなると考えられる．一方 IWF は大きくないが EWF が大きい場合として，「閲覧時間は長くはないが，多くの利用者が最近よく利用するページ」なども扱えるようになると期待できる．

第 2 の尺度は，内部重みの制約付き EWF である．これは，パターンに対する内部重みが閾値以上であるグラフを対象とした外部重み付き頻度である．

**定義 4**　$\theta \geq 0$ を制約としたとき，$D$ 中の $p$ の**制約付き重み付き頻度**（External Weighted Frequency under the constraint of Internal Weight，以降 CWF）を次のように定義する．

$$CWF(p, D, \theta) = \sum_{g \in D,\, p \preceq g,\, max\_iw(p,g) \geq \theta} ew(g) \,/\, \sum_{d \in D} ew(d).$$

図 1 において $\theta = 4$ とすると，$max\_iw(p3, g1) = 8$ および $max\_iw(p3, g3) = 3$ より，$CWF(p3, \{g1, g2, g3\}, 4) = ew(g1)/(ew(g1) + ew(g2) + ew(g3)) = 0.5/3 \simeq 0.17$ となる．

一般的に，小さい部分グラフパターンほど EWF の値は大きくなる．しかし，小さいパターンはそれほど重要でないと考えることもできる．そこで，内部重みを制約として用いることで，小さいパターンに不当に高い評価を与えることを避けることができる．たとえばウェブアクセスログ解析の例では，使い慣れたサイトにおいてメニューページ階層だけをたどる部分は閲覧時間が短くなると考えられるが，そのような部分だけからなるパターンを排除することができ，より重要かつ本質的な部分に焦点を当てたパターンの発見が期待できる．

上記の準備の下，本論文で対象とするデータマイニング問題の形式的な定義を与える．なお，入力データである重み付きグラフは非連結グラフでもよいが，抽出される部分グラフパターンは連結グラフでなければならない．

**問題 1　一般重み付き頻出部分グラフマイニング問題**（GWF-mining）とは，重み付きグ

ラフデータベース $D$，パラメータ $\lambda$（$0 \leq \lambda \leq 1$），最小頻度閾値 $\sigma$（$0 < \sigma \leq 1$）が与えられたとき，一般重み付き頻出部分グラフ，すなわち $GWF(p, D, \lambda) \geq \sigma$ を満たすすべての連結部分グラフパターン $p$ を発見することである．

**問題 2　制約付き重み付き頻出部分グラフマイニング問題**（CWF-mining）とは，重み付きグラフデータベース $D$，制約 $\theta$（$\geq 0$），最小頻度閾値 $\sigma$（$0 < \sigma \leq 1$）が与えられたとき，制約付き重み付き頻出部分グラフ，すなわち $CWF(p, D, \theta) \geq \sigma$ を満たすようなすべての連結部分グラフパターン $p$ を発見することである．

## 3. 重み付き頻出部分グラフの発見

本章では，まず部分グラフパターンを列挙する方法について概要を述べる．その後，この方法を用いて，GWF-mining および CWF-mining を解決するためのアルゴリズム GWF-mine および CWF-mine を提案する．

### 3.1 部分グラフの列挙

これまでに，部分グラフデータベースに含まれるすべての頻出部分グラフパターンを重複なく列挙する手法がいくつか提案されている．本論文ではこのうち，コードワードを用いた標準形判定と，最右拡張をともなう逆探索に基づく手法[22),23)] を採用する．これらの手法では，各部分グラフパターンをコードワードと呼ばれる頂点の順序と辺の接続関係を表す文字列で表現する．また，部分グラフパターンに対して辺を 1 つ追加することで新たなパターンを生成し，その操作を繰り返すことですべてのパターンを列挙する．その際，部分グラフパターン $p$ に対して辺を追加することで得られるパターン $q$ のコードワードは，$p$ のコードワードに，追加された辺に関する文字列を追加したものとなる．**図 2** に，コードワードの例を示す．図中において，たとえば $p11$ は $p10$ から生成されるが，そのコードワードは，$p10$ の末尾に新たに追加された辺の情報を加えることで構成される．

コードワードは，その生成過程（辺の追加順序）に依存して決まるので，たとえば図 2 中の $p11$ と $p11_a$，$p11_b$ のように，同型グラフどうしであっても異なるコードワードを持つ．このとき，同型グラフ中で最小のコードワードを持つパターンのみを標準形グラフとして採用することで，同型パターンの重複列挙を回避する．標準形グラフの判定は，部分グラフパターン自身を使って行われる．具体的には，部分グラフパターンにおける辺の接続関係を考慮し，自身のコードワードよりも小さいコードワードを生成するような辺の追加順序がないかを確認する．詳細は文献 1)，22) を参照されたい．

標準形グラフは，最右拡張[22)] に基づき，縦型探索に従い列挙される．**図 3**（右）に探索空間の例を示す．最右拡張とは，一定の条件下で部分グラフパターン $p$ からより特殊なパターン $q$ を得るための技術である．最右拡張を用いることで，明らかに標準形とはなりえないパターンの列挙が抑制され，不要な標準形判定を避けることができる．また，各部分グラフパターンはコードワードの大小関係 $<_{\text{lex}}$ 順に列挙される．すなわち，$p <_{\text{lex}} q$ なる 2 つの部分グラフ $p$，$q$ に対しては，$p$ が $q$ の先に列挙される．たとえば図 2 において，$p11 <_{\text{lex}} p12 <_{\text{lex}} p13 <_{\text{lex}} p14$ であるので，これらの部分グラフは $p11$，$p12$，$p13$，$p14$

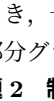
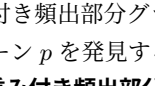
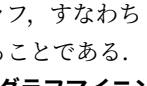
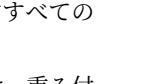
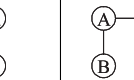
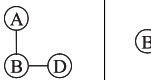
| $p10$ | $p11$ | $p12$ | $p13$ | $p14$ | $p11_a$ | $p11_b$ |
|---|---|---|---|---|---|---|
|  |  |  |  |  |  |  |
| (0,1,A,-,B) | (0,1,A,-,B)<br>(0,2,A,-,C) | (0,1,A,-,B)<br>(1,2,B,-,D) | (0,1,B,-,C) | (0,1,B,-,C)<br>(0,2,B,-,C)<br>(2,3,C,-,D) | (0,1,B,-,A)<br>(1,2,A,-,C) | (0,1,C,-,A)<br>(1,2,A,-,B) |

**図 2**　ラベル付きグラフパターンのコードワード
Fig. 2　Labeled graphs and those code words.

**Algorithm** Find_Freq( $D$, $\sigma$ )
1: **for each** $l \in \mathcal{L}$ in order of $<_{\text{lex}}$
2:　$p$ be a subgraph of one vertex labeled $l$.
3:　Freq-enum( $p$, $D$, $\sigma$ )

**Procedure** Freq-enum( $p$, $D$, $\sigma$ )
1: **if** ¬isCanonical( $p$ ) **then return**
2: **if** frequency of $p$ is less than $\sigma$ **then return**
3: **output** $p$
4: scan $D$ once, find every edge $e$ s.t.
5:　$p$ can be rightmost extended to $p \cdot e$
6: **for each** $e \in E$ in order of $<_{\text{lex}}$
7:　Freq-enum( $p \cdot e$, $D$, $\sigma$ )

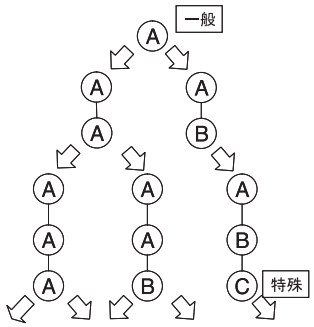



**図 3**　頻出部分グラフマイニングの擬似コード（左）とその探索空間の一部（右）
Fig. 3　Pseudo code for mining frequent subgraphs (left) and a part of search space (right).

の順で列挙される．

図 3（左）に，閾値以上の頻度を持つすべての部分グラフを列挙するアルゴリズムの擬似コードを示す．図中において，$D$ はデータベース，$\mathcal{L}$ は $D$ 中のラベル集合，$\sigma$（$0 < \sigma \leq 1$）は最小頻度閾値を表す．また $p \cdot e$ は，部分グラフ $p$ に 1 つの辺 $e$ を追加することによって得られるラベル付きグラフを表す．関数 'isCanonical' は $p$ が標準形グラフであれば真を返す．ところで，最右拡張を前提としたとき，標準形でないグラフから標準形グラフが得られることはない[22]．よって $p$ が標準形でなければ，$p$ の拡張を止める（Freq-enum の 1 行目）．また，$p \preceq q$ であるような部分グラフ $p$ と $q$ に関しては，$\{d \in D \mid p \preceq d\} \supseteq \{d \in D \mid q \preceq d\}$ の関係が成り立つので，パターンの頻度は逆単調性を満たす．よって，パターンの頻度が $\sigma$ 未満なら，$p$ の拡張は枝刈りされる（Freq-enum の 2 行目）．

### 3.2　一般重み付き頻出部分グラフの列挙

本節では，GWF-mining 問題を解決するためのアルゴリズム，GWF-mine を提案する．

$p \preceq q$ なる部分グラフパターン $p$ と $q$ に関して，$\{d \in D \mid p \preceq d\} \supseteq \{d \in D \mid q \preceq d\}$ の関係が成り立つので，EWF は逆単調性すなわち，$\forall q \succeq p[ewf(p, D) \geq ewf(q, D)]$ を満たす．一方，後に例を示すとおり，IWF は逆単調性を満たさないので，EWF と IWF の組合せである GWF も逆単調性を満たさない．より形式的には，たとえ $p \preceq q$ であっても，パターン $p$ と $q$ に関して $iwf(p, D) \geq iwf(q, D)$ および $GWF(p, D, \lambda) \geq GWF(q, D, \lambda)$ が成り立つとは限らない．図 1 における $p1 \preceq p2 \preceq p3$ の 3 つのパターンを用いて例を示す．$\lambda = 0.5$，$D = \{g1, g2, g3\}$ とすると，$(iwf(p1, D) = 7/30) < (iwf(p2, D) = 15/30) > (iwf(p3, D) = 11/30)$，$(GWF(p1, D, 0.5) = 37/60) < (GWF(p2, D, 0.5) = 45/60) > (GWF(p3, D, 0.5) = 31/60)$ である．

GWF は逆単調性を満たさないので，効率的なマイニングのためには別の枝刈り基準が必要となる．$p \preceq q$ なる部分グラフパターン $q$ の GWF の最大値を，$max\_GWF(p, D, \lambda) = \max_{p \preceq q}\{GWF(q, D, \lambda)\}$ と表記する．$max\_GWF(p, D, \lambda)$ が $\sigma$ より小さいなら $p \preceq q$ なるすべての部分グラフ $p$ を枝刈りすることができる．しかし，直接 $max\_GWF(p, D, \lambda)$ を計算することは $q \succeq p$ であるすべての部分グラフを列挙することに等しい．それゆえ，$max\_GWF(p, D, \lambda)$ の代わりに，枝刈り基準として GWF の上界値 $up\_GWF(p, D, \lambda)$ を採用する．

以下で，$up\_GWF(p, D, \lambda)$ に関する定義と補題を示す．$p$ と $q$ は $p \preceq q$ であるようなラベル付きグラフであり，$d = (V_d, E_d, l_d, w_d, ew(d))$ は $D$ 中の重み付きグラフである．

**定義 5**　以下の条件 (1)〜(4) で定義される，$d$ における $p$ の出現 $e = (V_e, E_e, l_d, w_d, ew(d))$ $\in E(p, d)$ を含む最大連結（重み付き）部分グラフを，$x(e, d) = (V_x, E_x, l_d, w_d, ew(d))$ と表記する．(1) $V_e \subseteq V_x \subseteq V_d$，(2) $E_e \subseteq E_x \subseteq E_d$，(3) $x(e, d)$ は連結，(4) $V_x \subseteq V'_x \subseteq V_d$ かつ $E_x \subseteq E'_x \subseteq E_d$ であるような連結部分グラフ $x' = (V'_x, E'_x, l_d, w_d, ew(d)) \neq x(e, d)$ は存在しない．たとえば，図 1 において，$g2$ 中の出現 $e = (\{v21\}, \{\}, l_{g2}, w_{g2}, 1)$ の最大連結部分グラフは $x(e, d) = (\{v21, v22\}, \{(v21, v22)\}, l_{g2}, w_{g2}, 1)$ である．

**定義 6**　$d$ 中の $p$ の内部重みの上界値を，出現に関する $d$ における最大連結部分グラフの内部重みの最大値，すなわち $up\_iw(p, d) = \max_{e \in E(p,d)} iw(x(e, d))$ と定義する．

$d$ における $p$ の内部重みの上界値に関して，次に示す補題が成り立つ．

**補題 1**　部分グラフパターンを $p$，重み付きグラフを $d$ とすると，不等式 $up\_iw(p, d) \geq \max_{p \preceq q} max\_iw(q, d)$ が成り立つ．

**証明**　集合 $X(p, d) = \{x(e, d) \mid e \in E(p, d)\}$ を考える．連結部分グラフ $q$ に対し，

$\forall (V_q, E_q, l_d, w_d, ew(d)) \in E(q, d)\, \exists (V_p, E_p, l_d, w_d, ew(d)) \in X(p, d)\, [V_q \subseteq V_p \wedge E_q \subseteq E_p]$

が成り立つ．□

**定義 7**　データベース $D$ における $p$ の IWF の上界値を以下のように定義する．

$up\_iwf(p, D) = \sum_{d \in D,\, p \preceq d} up\_iw(p, d) \,/\, \sum_{d \in D} iw(d)$.

**補題 2**　$up\_iwf(p, D)$ の値は，$p \preceq q$ なる部分グラフパターン $q$ の IWF の最大値以上である．すなわち，$up\_iwf(p, D) \geq \max_{p \preceq q} iwf(q, D)$ が成り立つ．

**証明**　$p \preceq q$ であるので，$\{d \in D \mid p \preceq d\} \supseteq \{d \in D \mid q \preceq d\}$ の関係が成り立つ．補題 1 より，$\forall d \in D[q \preceq d \rightarrow up\_iw(p, d) \geq max\_iw(q, d)]$ を得る．□

**定義 8**　EWF と IWF の上界値の組合せにより，GWF の上界値を次のように定義する．

$up\_GWF(p, D, \lambda) = \lambda ewf(p, D) + (1 - \lambda) up\_iwf(p, D)$.

$up\_GWF(p, D, \lambda)$ の上界値は，次の補題により枝刈り基準として用いることができる．

**補題 3**　不等式 $up\_GWF(p, D, \lambda) \geq max\_GWF(p, D, \lambda) = \max_{p \preceq q}\{GWF(q, D, \lambda)\}$ が成り立つ．

**証明**　$p \preceq q$ なる部分グラフパターン $p$ と $q$ に関しては，$\{d \in D \mid p \preceq d\} \supseteq \{d \in D \mid q \preceq d\}$ の関係が成り立つので，EWF は逆単調性を満たす．よって，$ewf(p, D) \geq ewf(q, D)$ が成り立つ．また補題 2 により，$up\_iwf(p, D) \geq iwf(q, D)$ を得る．□

上記の準備の下，GWF-mining 問題を解決するためのアルゴリズム，GWF-mine の擬似コードを**図 4** に示す．GWF-mine は，$up\_GWF(p, D, \lambda)$ に基づく枝刈りを列挙アルゴリズムに組み込むことによって得られる．また前述したように，部分グラフパターンは $<_{\mathrm{lex}}$ 順に列挙されるので，ラベル $l$ を持つパターン $p$ が列挙された後（GWF-mine の 3 行目）は，

**Algorithm** GWF-mine( $D$, $\sigma$, $\lambda$ )
1: **for each** $l \in \mathcal{L}$ in order of $<_{\mathrm{lex}}$
2:　$p$ be a subgraph of one vertex labeled $l$.
3:　GWF-enum( $p$, $D$, $\sigma$, $\lambda$ )
4:　delete all vertices $v$ whose label is $l$
5:　　and all edges connected to $v$ from $D$.

**Procedure** GWF-enum( $p$, $D$, $\sigma$, $\lambda$ )
1: **if** ¬isCanonical( $p$ ) **then return**
2: **if** $up\_GWF(p, D, \lambda) < \sigma$ **then return**
3: **if** $GWF(p, D, \lambda) \geq \sigma$ **then output** $p$
4: scan $D$ once, find every edge $e$ s.t.
5:　$p$ can be rightmost extended to $p \cdot e$
6: **for each** $e \in E$ in order of $<_{\mathrm{lex}}$
7:　GWF-enum( $p \cdot e$, $D$, $\sigma$, $\lambda$ )

**図 4**　GWF-mine の擬似コード
Fig. 4　Pseudo code of GWF-mine.

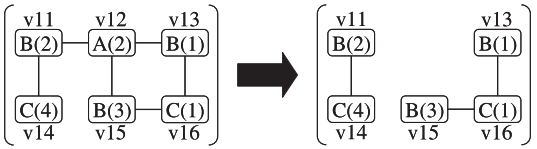



**図 5**　GWF-mine における頂点および辺の削除例
Fig. 5　An example of deleting vertices and edges in GWF-mine.

$<_{\mathrm{lex}}$ で $l$ 以前のラベルを持つ頂点は，その後のパターン発見で利用されない．したがって，そのような頂点 $v$ と $v$ に接続している辺を $D$ から削除する（GWF-mine の 4–5 行目）．この削除によって，$D$ 中の重み付きグラフの中には非連結グラフになるものもある．このとき結果として，削除の後生成される部分グラフパターン $q$ の上界値 $up\_GWF(q, D, \lambda)$ は減少することになる．たとえば，**図 5** において，ラベル $A$ を持つ部分グラフの列挙が終わったとすると，頂点 $v12$ および頂点 $v12$ に接続する辺が削除される．このとき，ラベル $B$ を持つ単一頂点の部分グラフパターン $p$ の上界値は，削除前の $up\_iw(p, g) = (2+2+1+3+4+1) = 13$ から，削除後の $up\_iw(p, g) = \max\{2+4, 1+3+1\} = 6$ に減少する．これにより，枝刈りがより促進されることが期待される．

GWF-mine に関して以下の定理が成り立つ．

**定理 1**　GWF-mine は GWF-mining 問題を解決する．すなわち，重み付きグラフデータベース $\mathcal{D}$，パラメータ $\lambda$，最小頻度閾値 $\sigma$ に対し，GWF-mine は重複なしに $GWF(p, D, \lambda) \geq \sigma$ であるようなすべての連結部分グラフ $p$ を発見することができる．

**証明**　最右拡張による完全な列挙と補題 3 による安全な枝刈りから導出される．　□

### 3.3　制約付き重み付き頻出部分グラフの列挙

GWF と同様に，CWF も逆単調性を満たさない．たとえば図 1 において，$\theta = 4$, $D = \{g1, g2, g3\}$ とすると，$CWF(p1, D, \theta) = 1.0/3$, $CWF(p2, D, \theta) = (0.5 + 1.0)/3$, $CWF(p3, D, \theta) = 0.5/3$ を得る．よって，GWF-mine と同様に CWF-mine においても，効率的な探索のために，CWF の上界値に基づく枝刈りを考える．

**定義 9**　部分グラフパターンを $p$，データベースを $D$，制約を $\theta \geq 0$ とし，CWF の上界値を $up\_CWF(p, D, \theta) = \sum_{d \in D,\, p \preceq d,\, up\_iw(p,d) \geq \theta} ew(d) \,/\, \sum_{d \in D} ew(d)$ と定義する．

上界値 $up\_CWF(p, D, \theta)$ は，制約として $max\_iw(p, d) \geq \theta$ の代わりに $up\_iw(p, d) \geq \theta$ を採用しており，その点のみが $CWF(p, D, \theta)$ と異なる．

$up\_GWF(p, D, \theta)$ と同様に，$up\_CWF(p, D, \theta) \geq \max_{p \preceq q}\{CWF(q, D, \theta)\}$ が成り立てば，$up\_CWF(p, D, \theta)$ を枝刈り基準として用いることができる．次に示す補題により，$up\_CWF(p, D, \theta)$ が安全な枝刈り基準であるということが保証される．

**補題 4**　不等式 $up\_CWF(p, D, \theta) \geq \max_{p \preceq q}\{CWF(q, D, \theta)\}$ が成り立つ．

**証明**　$p \preceq q$ と補題 1 により，$\{d \in D \mid p \preceq d, up\_iw(p, d) \geq \theta\} \supseteq \{d \in D \mid q \preceq d, max\_iw(q, d) \geq \theta\}\}$ が導出される．　□

**図 6** に，CWF-mining 問題を解決するためのアルゴリズム，CWF-mine の擬似コードを示す．この手法の基本構造は GWF-mine と同じである．$GWF(p, D, \lambda)$ および $up\_GWF(p, D, \lambda)$ の代わりに，CWF-mine は $CWF(p, D, \theta)$ および $up\_CWF(p, D, \theta)$ を採用する．さらに，ラベル $l$ の頂点を持つ部分グラフパターン $p$ を探索する前に，$iw(x) < \theta$ であるようなすべての最大連結部分グラフ $x$ を削除する．より形式的に記述すると，データベース $D$ から部分グラフの集合 $\{x(e, d) \mid d \in D, e \in E(p, d), iw(x(e, d)) < \theta\}$ を削除する（CWF-mine の 3–4 行目）．この削除により，$D$ 中の拡張可能な辺を探索するコストを削減できる．また，削除された部分グラフは制約 $\theta$ を満たすことはなく，EWF の計算に関与しないので，この削除はアルゴリズムの完全性を崩さない．

**Algorithm** CWF-mine( $D$, $\sigma$, $\theta$ )
1: **for each** $l \in \mathcal{L}$ in order of $<_{\text{lex}}$
2:　$p$ be a subgraph of one vertex labeled $l$.
3:　delete all subgraphs $x(e,d)$ from $D$ s.t.
4:　　$d \in D, e \in E(p,d)$ and $iw(x(e,d)) < \theta$
5:　CWF-enum( $p$, $D$, $\sigma$, $\theta$ )
6:　delete all vertices $v$ whose label is $l$
7:　　and all edges connected to $v$ from $D$.

**Procedure** CWF-enum( $p$, $D$, $\sigma$, $\theta$ )
1: **if** ¬isCanonical( $p$ ) **then return**
2: **if** $up\_CWF(p,D,\theta) < \sigma$ **then return**
3: **if** $CWF(p,D,\theta) \geq \sigma$ **then output** $p$
4: scan $D$ once, find every edge $e$ s.t.
5:　$p$ can be rightmost extended to $p \cdot e$
6: **for each** $e \in E$ in order of $<_{\text{lex}}$
7:　CWF-enum( $p \cdot e$, $D$, $\sigma$, $\theta$ )

**図 6**　CWF-mine の擬似コード
Fig. 6　Pseudo code of CWF-mine.

CWF-mine に関して次の定理が導出される．

**定理 2**　CWF-mine は CWF-mining 問題を解決する．すなわち，重み付きグラフのデータベース $\mathcal{D}$，制約 $\theta$，最小頻度閾値 $\sigma$ に対し，CWF-mine は重複なしに $CWF(p,D,\theta) \geq \sigma$ であるようなすべての連結部分グラフ $p$ を発見する．

**証明**　重複なしの完全な列挙方法と補題 4 による安全な枝刈りから導出される．　□

ここで，GWF-mine および CWF-mine の時間計算量について簡単に考察する．先述したとおり，両提案手法とも文献 22) による列挙手法に新たな枝刈り基準を導入したものであり，部分グラフ同型判定時に IWF 計算を埋め込むことで，その計算量はこれらの手法と同等となると考えられる．なお部分グラフ同型判定は NP-完全であるため，文献 22) では，部分グラフ同型判定数と標準形判定のための自己同型判定数に着目し，その計算量を $\mathcal{O}(kFS + rF)$ としている．ここで，$F$ は頻出グラフパターン数，$S$ はデータベースに含まれるデータ数，$k$ はある頻出グラフパターンとデータベース中のあるグラフとの間に存在する同型部分グラフの最大数であり，これにより部分グラフ同型判定数は $\mathcal{O}(kFS)$ となる．一方 $r$ は，探索において生成される，ある部分グラフパターンに対する同型グラフの最大数であり，これにより標準形判定のための自己同型判定数は $\mathcal{O}(rF)$ となる．詳細は文献 22) などを参照されたい．

## 4. 関 連 研 究

ラベル付きグラフデータベースからのパターンマイニング[5),21)] に関する研究は数多く行われている．そしてその成果として，文献 8)，9)，14)，22) のような効率的なグラフマイナが開発された．しかしながら，一部を除いてそれらの多くは重みをまったく考慮していない．重み付き部分グラフ発見の枠組みは文献 15) によって提案され，それはさらなるグラフマイニング問題に適用された[4),17)]．しかし，これらの研究は外部重みのみを考慮しているので，本論文の提案手法とは大きく異なっている．一方，重み付き有向グラフにおける重み付き頻出パターン発見は，文献 12) によって提案された．この研究は，重み付きグラフを対象とするという観点においては本論文の提案手法に関連があるが，単一グラフからの経路パターン発見に焦点を当てており，本論文の対象とは異なる．なお筆者の知る限り，グラフを対象とした外部および内部重みを同時に考慮するパターン発見手法は前例がなく，その面からも提案手法の有用性は高いと考えている．

内部重みに関連して，提案手法はユーティリティ（効用）を考慮したアイテム集合マイニング[6),7),11),24)–26)] の研究分野に深く関わりがある．この枠組みでは，トランザクション中の各アイテムは各々の効用値を持ち，その効用値は抽出されるべきパターンに関する基準として用いられている．ユーティリティを考慮したアイテム集合マイニングの分野における，‘一般ユーティリティを考慮したアイテム集合発見’[20)] および ‘ユーティリティ制約付き頻出アイテム集合発見’[16)] において，頻度とユーティリティの両方が同時に考慮されている．前者は評価基準として重み付き合計 $\lambda\text{sup}(p) + (1-\lambda)\text{util}(p)$ を採用している．ここで，‘sup’ および ‘util’ はそれぞれアイテム集合の頻度とユーティリティである．後者はアイテム集合 $p$ を評価するために集合 $\{t \in T \mid p \subseteq t, \text{util}(p,t) \geq \theta\}$ を考えている．ここで，$T$ はトランザクションデータベースであり，$\text{util}(p,t)$ は $t$ 中の $p$ のユーティリティ値である．本論文の課題である GWF-mining および CWF-mining はそれぞれこれらの枠組みを外部および内部重み付きグラフへと拡張したものと見なすことができる．

CWF-mining 問題では，内部重みを制約として扱うので，制約付きグラフマイニング[19),27)] と関連が深い．特に，データベースから今後の拡張に使われない部分を削除する操作（CWF-

mine の 3–4 行目）は，データ空間の枝刈り[27] の特別なケースと見なすことができる．

## 5. 実　　験

提案手法の有用性を評価するため Java 言語を用いて GWF-mine および CWF-mine を実装し，Windows マシン（CPU: Xeon 3.33 GHz，主記憶 32 GB）上で実験を行った．実験には**表 1** に示すデータセットを用いた．以下に，各データセットの詳細を示す．

$G_E$：グラフ生成器[3] を使用して生成された合成のデータセット．指数分布に従い，外部および内部重みを付与した．

$G_D$：内部重みを除いて $G_E$ と同一のデータセット．各頂点にその次数を内部重みとして付与した．また辺の重みは 0 とした．

$G_{DI}$：内部重みを除いて $G_E$ と同一のデータセット．各頂点にその次数の逆数を内部重みとして付与した．また辺の重みは 0 とした．

$MIT_1$：Reality mining データ[13] より生成した，モバイル機器通信のグラフのデータセット．頂点はモバイル機器，辺がモバイル機器どうしの通信を表す．6 時間を単位に 1 つのグラフを生成した．内部重みとして辺に通信回数を付与した．また頂点の重みは 0 とした．一方各グラフに対し，古いデータから新しいデータの順に，(0.75–1.0) の範囲で外部重みを付与した．

$MIT_2$：外部重みを除いて $MIT_1$ と同一のデータセット．このデータセットでは，通信の発生時間に応じ 0–6 時の範囲は 0.5，12–18 時の範囲は 1.5，それ以外の時間は 1 という外部重みを付与した．

$MUT$：化合物の化学構造を表すグラフ[18] のデータセット．頂点には内部重みとして電荷を正規化したものを付与した．一方，各グラフには外部重みとして突然変異性の対数を正規化したものを付与した．辺の重みは 0 である．

**表 1**　実験に用いたデータセット
Table 1　Data sets used in experiments.

| | $\|\mathcal{D}\|$ | $V$ | $E$ | $W_I$ | $W_E$ |
|---|---|---|---|---|---|
| $G_E$ | 10,000 | 11.6 | 20.5 | 32.05 | 0.98 |
| $G_D$ | 10,000 | 11.6 | 20.5 | 40.99 | 0.98 |
| $G_{DI}$ | 10,000 | 11.6 | 20.5 | 4.15 | 0.98 |
| $MIT_1$ | 1,140 | 47.9 | 34.9 | 44.75 | 0.88 |
| $MIT_2$ | 1,140 | 47.9 | 34.9 | 44.75 | 1.00 |
| $MUT$ | 230 | 25.6 | 27.4 | 11.22 | 0.48 |

$|\mathcal{D}|$：グラフ数．
$V$：1 グラフにおける頂点の平均数．
$E$：1 グラフにおける辺の平均数．
$W_I$：データセットにおけるグラフの平均の内部重み．
$W_E$：データセットにおけるグラフの平均の外部重み．

実験結果を**表 2**，**表 3** に示す．ここで，列 '$Pat$' は抽出されたパターン数（単位は千個），列 '$Time$' は実行時間（単位は秒），列 '$Cand$' は生成されたパターンの候補数（単位は千個）を表す．また，表中の '0.00' は 10 未満を表す．実験結果から，すべての場合において，両提案手法がある程度の規模のデータセットに対して適用可能であることが分かる．

次に GWF-mine の結果を考察する．表 2 において，最小頻度閾値 $\sigma$ が下がるにつれて，抽出されるパターン数および実行時間が増加している．また，$\lambda$ が減少するにつれて，パターン数が減少している．一方，パターン数の減少に対して候補数の減り方が緩やかな傾向があり，結果として，実行時間の減少も緩やかである．これらの結果より，外部重みより，内部重みに関する条件の方がより結果に影響を与えやすいことが考えられる．なお，$MIT_1$

**表 2**　GWF-mine 実験結果
Table 2　Experimental results of GWF-mine.

| | | $Pat$ | $Time$ | $Cand.$ | $Pat$ | $Time$ | $Cand.$ | $Pat$ | $Time$ | $Cand.$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $G_E$ | $\lambda$ | $\sigma = 0.015$ | | | $\sigma = 0.01$ | | | $\sigma = 0.005$ | | |
| | 0.7 | 1.55 | 15.69 | 3.67 | 3.77 | 20.56 | 10.31 | 15.86 | 36.72 | 38.73 |
| | 0.5 | 1.09 | 15.30 | 3.51 | 2.17 | 20.59 | 9.87 | 10.81 | 36.11 | 37.65 |
| | 0.3 | 0.59 | 15.09 | 3.35 | 1.24 | 20.08 | 9.46 | 5.94 | 36.14 | 36.88 |
| $G_D$ | $\lambda$ | $\sigma = 0.015$ | | | $\sigma = 0.01$ | | | $\sigma = 0.005$ | | |
| | 0.7 | 1.76 | 15.74 | 3.83 | 4.56 | 20.95 | 10.89 | 18.36 | 36.97 | 40.58 |
| | 0.5 | 1.35 | 15.91 | 3.80 | 3.11 | 20.77 | 10.84 | 14.44 | 37.58 | 40.61 |
| | 0.3 | 0.96 | 15.45 | 3.82 | 2.04 | 20.64 | 10.78 | 10.60 | 37.55 | 40.85 |
| $G_{DI}$ | $\lambda$ | $\sigma = 0.015$ | | | $\sigma = 0.01$ | | | $\sigma = 0.005$ | | |
| | 0.7 | 1.58 | 15.03 | 3.43 | 3.86 | 20.24 | 9.79 | 16.15 | 35.05 | 36.86 |
| | 0.5 | 1.13 | 14.73 | 3.18 | 2.27 | 19.61 | 8.99 | 11.18 | 34.31 | 34.57 |
| | 0.3 | 0.67 | 14.89 | 2.96 | 1.35 | 19.41 | 8.27 | 6.50 | 33.22 | 32.44 |
| $MIT_1$ | $\lambda$ | $\sigma = 0.10$ | | | $\sigma = 0.05$ | | | $\sigma = 0.03$ | | |
| | 0.7 | 0.13 | 14.78 | 0.24 | 0.37 | 19.02 | 0.73 | 0.79 | 29.03 | 1.71 |
| | 0.5 | 0.08 | 14.80 | 0.17 | 0.24 | 16.98 | 0.55 | 0.55 | 23.42 | 1.34 |
| | 0.3 | 0.04 | 15.80 | 0.12 | 0.14 | 17.50 | 0.40 | 0.32 | 20.69 | 1.03 |
| $MIT_2$ | $\lambda$ | $\sigma = 0.10$ | | | $\sigma = 0.05$ | | | $\sigma = 0.03$ | | |
| | 0.7 | 0.13 | 14.52 | 0.23 | 0.36 | 18.00 | 0.69 | 0.80 | 28.55 | 1.69 |
| | 0.5 | 0.08 | 14.39 | 0.17 | 0.24 | 16.92 | 0.53 | 0.53 | 23.81 | 1.35 |
| | 0.3 | 0.04 | 15.19 | 0.12 | 0.14 | 15.94 | 0.40 | 0.31 | 20.78 | 1.01 |
| $MUT$ | $\lambda$ | $\sigma = 0.15$ | | | $\sigma = 0.10$ | | | $\sigma = 0.05$ | | |
| | 0.7 | 541.0 | 950.4 | 2,014.0 | 2,113.1 | 2,083.4 | 7,225.8 | 14,029.8 | 7,465.8 | 46,567.9 |
| | 0.5 | 300.9 | 946.4 | 1,955.1 | 1,512.1 | 1,990.5 | 7,039.9 | 11,136.0 | 11,447.6 | 64,587.4 |
| | 0.3 | 117.6 | 924.2 | 1,908.2 | 1,003.2 | 2,097.6 | 6,813.7 | 8,649.9 | 18,409.2 | 85,931.4 |

$Pat$：抽出されたパターン数（千個）．　$Time$：実行時間（秒）．　$Cand.$：候補数（千個）．

**表 3**　CWF-mine の実験結果
Table 3　Experimental results of CWF-mine.

| | | $Pat$ | $Time$ | $Cand.$ | $Pat$ | $Time$ | $Cand.$ | $Pat$ | $Time$ | $Cand.$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $G_E$ | $\theta$ | $\sigma = 0.015$ | | | $\sigma = 0.01$ | | | $\sigma = 0.005$ | | |
| | 5 | 1.19 | 15.45 | 3.88 | 4.16 | 21.00 | 10.98 | 20.18 | 37.13 | 40.72 |
| | 4 | 1.97 | 15.55 | 3.88 | 5.88 | 20.91 | 11.00 | 23.21 | 37.33 | 40.75 |
| | 3 | 2.45 | 15.70 | 3.89 | 6.76 | 20.84 | 11.03 | 24.53 | 37.11 | 40.76 |
| $G_D$ | $\theta$ | $\sigma = 0.015$ | | | $\sigma = 0.01$ | | | $\sigma = 0.005$ | | |
| | 20 | 0.00 | 15.28 | 3.62 | 0.00 | 20.58 | 10.43 | 2.18 | 36.67 | 39.83 |
| | 16 | 0.20 | 15.83 | 3.82 | 1.51 | 20.91 | 10.83 | 15.70 | 36.67 | 40.58 |
| | 12 | 1.90 | 15.52 | 3.91 | 6.28 | 20.66 | 11.01 | 24.42 | 36.70 | 40.82 |
| $G_{DI}$ | $\theta$ | $\sigma = 0.015$ | | | $\sigma = 0.01$ | | | $\sigma = 0.005$ | | |
| | 1.2 | 0.00 | 15.64 | 3.82 | 0.00 | 20.94 | 10.81 | 0.27 | 36.45 | 40.42 |
| | 0.8 | 1.25 | 15.67 | 3.91 | 5.12 | 21.03 | 11.02 | 22.18 | 37.27 | 40.83 |
| | 0.4 | 2.65 | 15.73 | 3.93 | 7.10 | 20.86 | 11.05 | 25.08 | 37.30 | 40.86 |
| $MIT_1$ | $\theta$ | $\sigma = 0.10$ | | | $\sigma = 0.05$ | | | $\sigma = 0.03$ | | |
| | 7 | 0.00 | 15.84 | 0.21 | 0.07 | 18.59 | 0.81 | 0.33 | 37.33 | 2.02 |
| | 5 | 0.04 | 14.95 | 0.28 | 0.25 | 19.66 | 0.93 | 0.80 | 38.27 | 2.25 |
| | 3 | 0.13 | 18.66 | 0.33 | 0.48 | 19.70 | 0.99 | 1.16 | 48.58 | 2.34 |
| $MIT_2$ | $\theta$ | $\sigma = 0.10$ | | | $\sigma = 0.05$ | | | $\sigma = 0.03$ | | |
| | 7 | 0.00 | 14.63 | 0.20 | 0.08 | 18.69 | 0.79 | 0.38 | 30.55 | 1.99 |
| | 5 | 0.04 | 17.74 | 0.27 | 0.26 | 19.86 | 0.92 | 0.83 | 37.63 | 2.18 |
| | 3 | 0.13 | 16.88 | 0.31 | 0.47 | 19.67 | 0.98 | 1.15 | 40.56 | 2.27 |
| $MUT$ | $\theta$ | $\sigma = 0.15$ | | | $\sigma = 0.10$ | | | $\sigma = 0.05$ | | |
| | 14 | 0.0 | 279.1 | 234.3 | 0.0 | 851.2 | 1,708.2 | 0.0 | 6,141.5 | 28,344.3 |
| | 10 | 0.0 | 909.6 | 2,073.2 | 10.4 | 2,047.7 | 7,616.4 | 1,491.6 | 7,669.8 | 45,580.7 |
| | 7 | 748.5 | 936.9 | 2,096.9 | 2,896.2 | 1,994.5 | 7,643.6 | 18,750.7 | 7,690.8 | 45,618.6 |

$Pat$：抽出されたパターン数（千個）．　$Time$：実行時間（秒）．　$Cand.$：候補数（千個）．

と $MIT_2$ は外部重みを変えただけである．全体的に得られるパターン数が少なかったこともあるが，類似の傾向が見られる．

次に CWF-mine の結果について考察する．表 3 において，最小頻度閾値 $\sigma$ が下がるに従って，抽出されるパターン数および実行時間が増加している．内部重みの制約が緩くなるに従って抽出されるパターン数は増加するが，$MUT$ 以外では，実行時間はほとんど変わらない．$MUT$ において，$\theta = 10$ から $\theta = 14$ にかけて，候補数が大幅に減少しており，この間で多くの枝刈りが行われたと考えられる．一方，$\theta = 7$ から $\theta = 10$ にかけてはそれほど枝刈りが行われておらず，より効率的な枝刈り基準を考える必要がある．

重みを考慮しない場合との違いを検証するために，生成されたパターン集合の比較を行った．第 1 の比較として，$MIT_2$ を対象に最小頻度閾値 $\sigma = 0.05$ における頻度上位 $K$ 位の部分グラフパターンの大きさ（辺数）の平均を算出した．結果を**表 4** に示す．なお，CWF-mine の $\lambda = 7$ における $K = 100$ は，この条件で得られた全 77 パターンの平均である．実験結果より，特に CWF-mine において，gSpan より大きなパターンが得られていることが分かる．このことより，小さなパターンに対して不当に高い評価を与えることを避けるという CWF の意図が適切に機能していることが分かる．一方 GWF-mine に関しても，多少ではあるが得られるパターンサイズの増大が確認される．また若干ではあるが，IWF を重視した場合の方がサイズが大きくなる傾向にあることが分かる．

**表 4**　上位 $K$ パターンの平均辺数（$MIT_2$，$\sigma = 0.05$）
Table 4　Average numbers of edges in Top-$K$ patterns ($MIT_2$, $\sigma = 0.05$).

| | $\lambda/\theta$ | $K$ | | |
|---|---|---|---|---|
| | | 10 | 50 | 100 |
| GWF | 0.7 | 1.50 | 2.30 | 2.85 |
| -mine | 0.5 | 1.50 | 2.34 | 2.89 |
| | 0.3 | 1.50 | 2.44 | 2.99 |
| CWF | 7 | 6.30 | 6.52 | 6.65 |
| -mine | 5 | 4.70 | 5.00 | 5.19 |
| | 3 | 3.20 | 3.52 | 3.87 |
| gSpan | | 1.50 | 2.22 | 2.79 |

**表 5**　gSpan との Jaccard 係数（$MUT$）
Table 5　Jaccard similarity coefficient ($MUT$).

| GWF-mine | $\lambda$ | $\sigma = 0.15$ | $\sigma = 0.10$ | $\sigma = 0.05$ |
|---|---|---|---|---|
| | 0.7 | 0.59 | 0.65 | 0.70 |
| | 0.5 | 0.83 | 0.85 | 0.81 |
| | 0.3 | 0.37 | 0.72 | 0.80 |
| CWF-mine | $\theta$ | $\sigma = 0.15$ | $\sigma = 0.10$ | $\sigma = 0.05$ |
| | 14 | 0.00 | 0.00 | 0.00 |
| | 10 | 0.00 | 0.00 | 0.03 |
| | 7 | 0.18 | 0.35 | 0.47 |

第 2 の比較として，データセット $MUT$ を対象に，頻出部分グラフ発見アルゴリズム gSpan [22] によって得られるパターン集合との Jaccard 係数を算出した．Jaccard 係数とは，集合間の類似性を示す尺度であり，集合 $X$ と $Y$ に対し $\frac{1X \cap Y1}{1X \cup Y1}$ と定義される．定義から分かるように，値が小さいほど，集合間の類似性が低い．結果を**表 5** に示す．この結果より，重みを考慮しない場合と考慮する場合とで，得られるパターン集合が大きく異なることが分

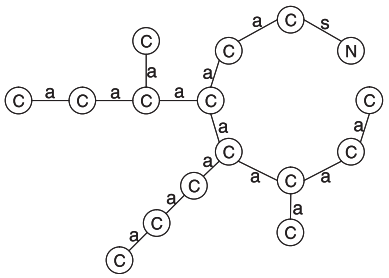


**図 7**　抽出されたパターンの例
Fig. 7　An example of extracted pattern.

かる．また最小頻度閾値 $\sigma$ が増加するにつれて，Jaccard 係数の値も減少する．重みを考慮しない場合，一般に，小さな部分グラフほど頻度が大きい．したがって，最小頻度閾値を大きくすると，多くの小さな部分グラフが列挙されることになる．一方重みを考慮した場合，小さな部分グラフが必ずしも高い評価値を持つとは限らない．よって，最小頻度閾値の増加にともない，Jaccard 係数が減少したのだと考えられる．

ところで，$MUT$ を対象とした場合，GWF-mine の $\lambda = 0.3$ においては，得られた部分グラフの集合は，重みを考慮しない場合に得られる部分グラフの集合に包含されていた．このことから，重みを考慮することで，より特徴的なものだけが残された可能性が示唆される．また**図 7** に，$\sigma = 0.15$，$\lambda = 0.7$ において $MUT$ から抽出された部分グラフの例を示す．ここで，辺ラベル 's' は single bond，'a' は aromatic bond を表す．なおこの部分グラフは，重みを考慮しない場合，同頻度閾値では得られなかったものであり，提案手法により，既存手法とは異なるパターンの獲得が可能であることを表す 1 つの例となっている．

## 6. 結　　論

本論文では，重み付きグラフデータベースからの頻出部分グラフ発見に関して，新たなデータマイニング問題点 GWF-mining および CWF-mining を設定し，これらを効率的に解決する手法 GWF-mine および CWF-mine を提案した．また実験を通じ，その有用性を確認した．

今後の課題としては，(1) GWF-mining と CWF-mining との詳細な比較，(2) 得られた部分グラフ自体の検証，(3) より大規模なデータを用いた実験，(4) 非負の重みの導入，(5) IWF 計算における最重出現以外の利用などがあげられる．また今回は重要性の基準として，EWF と IWF を組み合わせた GWF および CWF を採用したが，その他の基準として，EWF と IWF の相乗平均や調和平均なども考えられる．加えて，EWF および IWF それぞれに対して閾値を設定するということも考えられる．今後は，重み付きデータベースにおける重要性の定義に関しても検討を行っていきたいと考えている．

**謝辞**　有益なご指摘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 および基盤研究（B）：課題番号 20300038）による．

## 参 考 文 献

1) Borgelt, C.: On canonical forms for frequent graph mining, *Working Notes of the 3rd International ECML/PKDD-Workshop on Mining Graphs, Trees and Sequences* (*MGTS-05*), pp.1–12 (2005).
2) Bringmann, B. and Nijssen, S.: What is frequent in a single graph?, *Proc. 12th Pacific-Asia Conference on Knowledge Discovery and Data Mining* (*PAKDD2008*), pp.858–863 (2008).
3) Cheng, J., Ke, Y. and Ng, W.: Graphgen: A graph synthetic generator (2006). http://www.cse.ust.hk/graphgen/
4) Chiappa, S., Saigo, H. and Tsuda, K.: A Bayesian approach to graph regression with relevant subgraph selection, *Proc. 9th SIAM International Conference on Data Mining* (*SDM2009*), pp.295–304 (2009).
5) Cook, D.J. and Holder, B.L. (Eds.): *Mining Graph Data*, Wiley-Interscience (2005).
6) Erwin, A., Gopalan, R.P. and Achuthan, N.R.: CTU-mine: An efficient high utility itemset mining algorithm using the pattern growth approach, *Proc. 7th IEEE International Conference on Computer and Information Technology* (*CIT'07*), pp.71–76 (2007).
7) Erwin, A., Gopalan, R.P. and Achuthan, N.R.: Efficient mining of high utility itemsets from large datasets, *Proc. 12th Pacific-Asia Conference on Knowledge Discovery and Data Mining* (*PAKDD2008*), pp.554–561 (2008).
8) Inokuchi, A., Washio, T. and Motoda, H.: Complete mining of frequent patterns from graphs: Mining graph data, *Machine Learning*, Vol.50, pp.321–354 (2003).
9) Kuramochi, M. and Karypis, G.: Frequent subgraph discovery, *Proc. 2001 IEEE International Conference on Data Mining* (*ICDM'01*), pp.313–320 (2001).
10) Kuramochi, M. and Karypis, G.: Finding Frequent Patterns in a Large Sparse Graph, *Data Mining and Knowledge Discovery*, Vol.11, No.3, pp.213–321 (2005).
11) Liu, Y., Liao, W.-K. and Choudhary, A.: A fast high utility itemsets mining algo-

rithm, *Proc. 1st international workshop on Utility-based data mining* (*UBDM'05*), pp.90–99 (2005).
12) Lee, S.D. and Park, H.C.: Mining weighted frequent patterns from path traversals on weighted graph, *International Journal of Computer Science and Network Security*, Vol.7, No.4, pp.140–148 (2007).
13) MIT Media Lab.: Reality mining. http://reality.media.mit.edu/
14) Nijssen, S. and Kok, J.: A quickstart in frequent structure mining can make a difference, *Proc. 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (*KDD2004*), pp.647–652 (2004).
15) Nowozin, S., Tsuda, K., Uno, T., Kudo, T. and Bakir, G.H.: Weighted substructure mining for image analysis, *2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (*CVPR 2007*), pp.1–8 (2007).
16) Podpečan, V., Lavrač, N. and Kononenko, I.: A fast algorithm for mining utility-frequent itemsets, *Proc. International Workshop on Constraint-based Mining and Learning* (*CMILE'07*), pp.9–20 (2007).
17) Saigo, H. and Tsuda, K.: Iterative subgraph mining for principal component analysis, *Proc. 8th IEEE International Conference on Data Mining* (*ICDM'08*), pp.1007–1012 (2008).
18) Srinivasan, A., Muggleton, S., Sternberg, M.J.E. and King, R.D.: Theories for Mutagenicity: A Study in First-Order and Feature-Based Induction, *Artificial Intelligence*, Vol.85, No.1-2, pp.277–299 (1996).
19) Wang, C., Zhu, Y., Wu, T., Wang, W. and Shi, B.: Constraint-Based Graph Mining in Large Database, *Proc. 7th Asia-Pacific Web Conference* (*APWeb 2005*), pp.133–144 (2005).
20) Wang, J., Liu, Y., Zhou, L., Shi, Y. and Zhu, X.: Pushing frequency constraint to utility mining model, *Proc. 7th international conference on Computational Science* (*ICCS'07*), pp.685–692 (2007).
21) Washio, T. and Motoda, H.: State of the art of graph-based data mining, *SIGKDD Explorations*, Vol.5, No.1, pp.59–68 (2003).
22) Yan, X. and Han, J.: gSpan: Graph-based substructure pattern mining, *Proc. 2002 IEEE International Conference on Data Mining* (*ICDM'02*), pp.721–724, (Expanded Version: UIUC Technical Report, UIUCDCS-R-2002-2296) (2002).
23) Yan, X. and Han, J.: CloseGraph: Mining closed frequent graph patterns, *Proc. 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.286–295 (2003).
24) Yao, H., Hamilton, H. and Geng, L.: A unified framework for utility-based measures for mining itemsets, *Proc. 2nd Workshop on Utility-Based Data Mining*, pp.28–37 (2006).
25) Yao, H., Hamilton, H.J. and Butz, C.J.: A foundational approach to mining itemset utilities from databases, *Proc. 3rd SIAM International Conference on Data Mining*, pp.482–486 (2004).
26) Yeh, J.-S., Li, Y.-C. and Chang, C.-C.: Two-phase algorithms for a novel utility-frequent mining model, *Emerging Technologies in Knowledge Discovery and Data Mining, PAKDD 2007 International Workshops, Revised Selected Papers*, pp.433–444 (2007).
27) Zhu, F., Yan, X., Han, J. and Yu, P.S.: gPrune: A constraint Pushing Framework for Graph Pattern Mining, *Proc. 11th Pacific-Asia Conference on Knowledge Discovery and Data Mining* (*PAKDD2007*), pp.388–400 (2007).

(平成 21 年 12 月 19 日受付)
(平成 22 年 4 月 7 日採録)

(担当編集委員　戸田 浩之)

**信田　正樹**

1987 年生．2010 年神戸大学工学部情報知能工学科卒業．在学中，グラフマイニングに関する研究に従事．

**尾崎　知伸**

1973 年生．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．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等の研究に従事．人工知能学会会員．

**大川　剛直**（正会員）

1963 年生．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等の研究に従事．IEEE，人工知能学会，電子情報通信学会，電気学会等の各会員．