3I4-5

ネットワーク構造のパターンマイニングにおける誘導部分グラフ同型に基づくパターンマッチング

A Pattern Matching using Induced Subgraph Isomorphism for Network Pattern Mining

森 遼太 Ryota Mori　武藤 敦子 Atsuko Mutoh　森山 甲一 Koichi Moriyama　犬塚 信博 Nobuhiro Inuzuka

名古屋工業大学 大学院工学研究科 情報工学専攻
Department of Computer Science and Engineering, Graduate School of Engineering, Nagoya Institute of Technology

Hanabi is a pattern mining algorithm from networks. However, a pattern matching in Hanabi did not consider features of networks. Accordingly, a pattern matching using distance was proposed. This paper proposes a pattern matching using induced subgraph isomorphism. Then we compared the proposed pattern matching with other five matchings. We implemented Hanabi algorithm with all the matchings and results of an experiment are reported.

1. はじめに

大規模データの中から有用な知識を見つけ出すデータマイニングが注目されている。中でも複数のテーブルから構成されるデータベースを対象とするものを**関係型データマイニング** (Multi-Relational Data Mining: MRDM) と呼び、**帰納論理プログラミング** (Inductive Logic Programming: ILP) の枠組みで研究されてきた。ILP とは、述語論理の豊かな表現力を利用して帰納推論を行うアプローチである。

MRDM においては、1 つの事例に関するデータがデータベースの中で他の事例のデータと区別され、独立している状態のものを扱うことが通例である。分子構造を扱う場合等がこの場合であり、グラフマイニングでも同様な扱いがされる。他方で、社会ネットワーク分析等でパターンを扱う場合は対象間のデータが切れ目を持っていない。これを**開いた構造**と呼び、前者をこれと区別して**閉じた構造**という。

開いた構造、特に社会ネットワークを扱うアルゴリズムとして **Hanabi**[西尾 13] が提案されている。事例をサンプリングし、そこから基本パターンをボトムアップに得、これを組み合わせる **Mapix** アルゴリズム [Motoyama 06] をネットワークマイニングに拡張している。しかし、この手法ではマッチングにおいてネットワーク特有の特徴を十分に考慮していない。そのため、山崎らは距離の概念を導入したマッチング [山崎 13] を提案した。本論文では、距離ではなくノード同士の繋がり方に着目した誘導部分グラフ同型に基づくマッチングを提案する。誘導部分グラフ同型に基づくマッチングは、グラフマイニングでは既に考えられているが、ネットワークマイニングでは考えられていない。そこで、ネットワークマイニングに適用し、他マッチングと比較する。

本論文では、2 節で Hanabi のアルゴリズムについて、3 節で従来のマッチングについて述べ、4 節で誘導部分グラフ同型に基づくマッチングを提案する。また、5 節で各マッチングの定義による比較、及び実験を行い評価する。最後に 6 節でまとめを行う。

2. Hanabi

Hanabi は、ILP の枠組みでネットワーク構造のパターンマイニングを行うアルゴリズムである。Hanabi では、ネットワークからノードをサンプリングし、その近傍からボトムアップに基本パターンを生成する。そして、アプリオリ性に基づいてそれらを重ね合わせることでパターンを枚挙する。

Hanabi には対象の集合に関する**関係データベース**、つまりリテラルの集合が与えられる。関係データベースは、単項の述語 1 つと 2 項の述語 1 つ以上を含み、単項の述語を**目標述語**、そのリテラルを**目標リテラル**という。2 項の述語は対象間のネットワークを表現し、片方の引数に項を与えると別の引数が決まる関係を持つ。前者を**入力引数**、後者を**出力引数**といい、これが指定されているとする [古川 01]。

定義 1 (近傍) データベース D のある目標リテラル e の**近傍**とは、(1) 全ての入力引数が e の基礎項であるリテラル (その出力項を近傍基礎項と呼ぶ) (2) 全ての引数が e の近傍基礎項であるリテラル、によって決まるリテラル集合である。

定義 2 (変数化) 基礎リテラル集合 C_g に対して、次の条件を満たすリテラル集合 C_v を C_g の変数化という。(1) C_v は基礎項を含まない (2) $C_g = C_v\theta$ を満たす代入 $\theta = \{v_1/t_1, ..., v_n/t_n\}$ ($t_1, ..., t_n$ は全て異なる項である) が存在する。

目標 e の近傍 N を本体、e を頭部とする節 ($e \leftarrow N$) を変数化したものを**単位花火節**という。単位花火節 C の頭部を head(C)、本体を body(C) と書く。これが基本パターンである。C の本体の項のうち出力引数に現れる変数を**連結変数**という。

例 1 (近傍の変数化) 表 1 のデータベースで、member を目標述語とする。目標リテラル e = member(person1) の近傍 N_e は、次のとおり (member(person2) 等も含むが省略)。

N_e = {friend(person1,person2), friend(person1,person3),
friend(person1,person4), friend(person2,person4),
friend(person3,person4)}

これからできる単位花火節は次の通りである。

member(x_1) ← friend(x_1, x_2), friend(x_1, x_3), friend(x_1, x_4),
friend(x_2, x_4), friend(x_3, x_4).

定義 3 (花火節) 次で定義するものが、**花火節**である。(1) 単位花火節は花火節である。(2) 花火節 C_1, C_2 に対して、C=C_1∪body(C_2)$\{A/B\}$ として得られる C も花火節である。ここで、A は head(C_2) の唯一の変数、B は body(C_1) の 1 つ

連絡先: 森 遼太，名古屋工業大学大学院工学研究科情報工学専攻，愛知県名古屋市昭和区御器所町，r.mori.163@nitech.jp

表 1: 友人関係ネットワークに関するデータベース

member(X)	friend(X, Y)	
person1	person1	person2
person2	person1	person3
person3	person1	person4
person4	person2	person4
person5	person3	person4
person6	person3	person6
person7	person4	person5
	...	...

表 2: ネットワークパターン枚挙アルゴリズム Hanabi

HANABI($\mathbf{r}$, $\mathbf{t}$, $\sup_{\min}$):
input : データベース D; 目標リテラル $\mathbf{t}$; 最低支持度 $\sup_{\min}$;
output : 頻出花火パターン $Freq$;
1. $\mathcal{U} := \emptyset$; $k := 1$;
2. **for each** $e \in \mathbf{t}$ **do** $\mathcal{U} := \mathcal{U} \cup e$ の花火アイテム;
3. $\mathcal{F}_1 := \{S \in \mathcal{U} \mid sup_s \geq \sup_{\min}\}$;
4. **while** $\mathcal{F}_k \neq \emptyset$ **do**
5. $\mathcal{C}_{k+1} :=$ SUPERPOSITIONSHELL($\mathcal{F}_k$のパター ン木);
6. $\mathcal{F}_{k+1} := \{CS \in \mathcal{C}_{k+1} \mid sup_{cs} \geq \sup_{\min}\}$;
7. $Freq := Freq \cup \mathcal{F}_{k+1}$; $k := k + 1$;
8. **return** $Freq$;

の**連結変数**である。
(2) によって得られた花火節 C の連結変数は、C_1 の連結変数のうち A 以外のものおよび C_2 の連結変数である。

次に、連結の情報を保存するためパターン木を定義する。

定義 4 (パターン木) データベース D に関するすべての単位花火節の集合$\mathcal{U}$ と任意の要素を表す記号 χ の和集合 $V = \mathcal{U} \cup \{\chi\}$ を頂点集合とし、辺 E $\subseteq$ V $\times$ V、および割当関数 $l : V \to S$ に対する順序木 $T = ((V, l), E)$ を**パターン木**と呼ぶ。ここで、順序は連結変数のどれで連結したかが分かるようにつけられている。

深さが k のパターンを (k+1)-shell と呼ぶ。次に、パターンの**支持度**を定義する。

定義 5 (支持度) ネットワークを表すデータベース D におけるパターン (花火節)C の支持度は次の sup_c である。

$$sup_c = \frac{\sum_{t \in T} match(D, C, t)}{|T|}$$

T は D の目標リテラルの集合である。$match$ については、3 節で検討する。

支持度が最小サポートを超えるパターンを**頻出パターン**とする。Hanabi では、アプリオリ性に基づいて頻出パターンのみを連結していくことで効率よくパターンを枚挙する。Hanabi のアルゴリズムを表 2、表 3 に示す。

3. 従来のパターンマッチング

枚挙されるパターンは、マッチングによって大きく変化する。代表的なマッチングとして θ **包摂**と **OI 包摂**がある。

表 3: 重ね合わせによるパターン木の成長

SUPERPOSITIONSHELL($\mathcal{T}_k$):
input : パターン木の集合 $\mathcal{T}_k$;
output : 重ね合わせ $\mathcal{T}_{k+1}$;
1. $\mathcal{T}_{k+1} := \emptyset$;
2. **for each** $T \in \mathcal{T}_k$ **do**
3. $Sub^T :=$ T からルートを除いて得られる木の集合;
4. **for each** $Sub_i^T \in Sub^T$ **do**
5. **for each** $T_j \in \mathcal{T}_k$ **do**
6. **if** Sub_i^T と最深の葉を除いた T_j が同型 **then**
7. $\mathcal{T}_{k+1} := \mathcal{T}_{k+1} \cup \{T_j$ と Sub_j^T の重ね合わせ $\}$;
8. **return** $\mathcal{T}_{k+1}$;

定義 6 (θ 包摂) 花火節 C とデータベース D について、$t = \text{head}(C)\theta$, $\text{body}(C)\theta\rho \subseteq D$ となる代入 θ,ρ が存在するとき、C は D を θ 包摂するといい、$match_\theta(D, C, t)$ と表す。

定義 7 (OI 包摂) 花火節 C とデータベース D について、(1)t = head(C)θ, body(C)$\theta\rho \subseteq D$(2)$b_i \neq b_j$(i $\neq$ j) を満たす代入 $\theta = \{A_1/b_1\}$, $\rho = \{A_2/b_2, \ldots, A_n/b_n\}$ が存在するとき、C は D を OI 包摂するといい、$match_{OI}(D, C, t)$ と表す。

θ 包摂と OI 包摂は ILP でよく使われるが、ネットワーク特有の構造を考慮しているとは言えない。山崎らはパターンのネットワーク上での広がりを反映するよう、**距離に基づくマッチング**を提案した。

定義 8 (変数深度) 節 C に現れる変数 v に対して、次のように再帰的に定義される $d_{var}(v)$ を v の**変数深度**という。(1)v が C の head(C) である場合 $d_{var}(v)$=0、(2) それ以外の場合 $d_{var}(v) = (min_{u \in U_v}\ d_{var}(u))$ +1。U_v は C の本体のリテラルのうち v を出力として含むものに含まれる入力変数の集合である。

変数深度は、花火節の側で定義される。これを用いて花火節とネットワークのマッチングを以下のように行う。

定義 9 (変数深度を考慮したマッチング) 花火節 C とデータベース D において、(1)t = head(C)θ, body(C)$\theta\rho \subseteq D$(2)$d_{var}(A_i) \neq d_{var}(A_j)$ ならば $b_i \neq b_j$ を満たす代入 $\theta = \{A_1/b_1\}$, $\rho = \{A_2/b_2, \ldots, A_n/b_n\}$ が存在するとき C は D にマッチするといい、$match_{vd}(D, C, t)$ と表す。

$match_{vd}$ では、ネットワーク上の 1 つのノードを変数深度の異なる 2 つの変数にマッチさせない。次に、ネットワーク上での距離を項の深度として定義する。

定義 10 (項の深度) データベース D のある目標リテラルの基礎項 e と任意の基礎項 t について、e から t への**項の深度** $d_{term}(e, t)$ は次のとおりである。(1)$t = e$ ならば、$d_{term}(e, t)$ = 0 (2)$d_{term}(e, y)$ が有限の値を持つ y を入力に、t を出力に持つリテラルが存在するならば、$d_{term}(e, t) = min_y d_{term}(e, y)$ +1 (3) その他の場合、$d_{term}(e, t) = \infty$

これを用いて、花火節とネットワークのマッチングを以下のように行う。

定義 11 (項の深度を考慮したマッチング) 花火節 C とデータベース D において、(1)t = head(C)θ, body(C)$\theta\rho \subseteq$

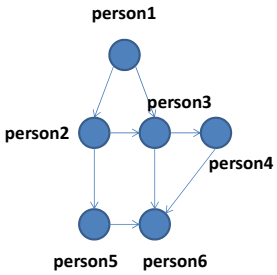

図 1: 友人関係ネット

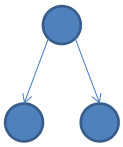
図 2: パターン

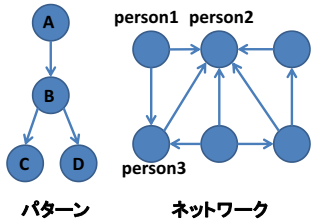

図 3: マッチングの例 1

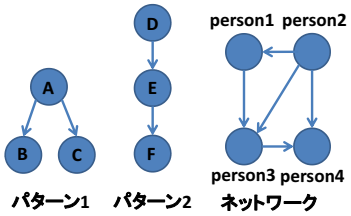

図 4: マッチングの例 2

$D(2)d_{var}(A_i) = d_{term}(e, b_i)$($e$ は C の頭部に代入された値) を満たす代入 $\theta = \{A_1/b_1\}$, $\rho = \{A_2/b_2, \ldots, A_n/b_n\}$ が存在するとき C は D にマッチするといい、$match_{td\theta}(D, C, t)$ と表す。また、(1)(2) に加え (3)$b_i \neq b_j$(i ≠ j) を満たす代入が存在するとき、$match_{tdOI}(D, C, t)$ と表す。

4. 提案手法

山崎らの提案した距離に基づくマッチングにより、θ 包摂、OI 包摂の問題点はいくらか解消された。しかし、まだ不十分な面が考えられる。例えば図 1 の person1 に注目したとき、図 2 がマッチングするのは本当に適切なのだろうか。部分グラフとして存在しているのは確かだが、person1, person2, person3 の間の繋がりに注目すると図 2 では不足がある。

本稿では、ネットワーク上でのノード同士の繋がり方に着目し、**誘導部分グラフ同型に基づくマッチング**を提案する。

定義 12 (誘導部分グラフ同型に基づくマッチング) 花火節 C とデータベース D において、次の条件を満たす代入 $\theta = \{A_1/b_1\}$, $\rho = \{A_2/b_2, \ldots, A_n/b_n\}$ が存在するとき C は D にマッチするといい、$match_{is}(D, C, t)$ と表す。(1)t = head$(C)\theta$, body$(C)\theta\rho \subseteq D$ (2)$b_i \neq b_j$(i ≠ j) (3)D のリテラルの内、$\{b_1, ..., b_n\}$ 以外に項を持たないものの集合を D'(誘導部分グラフを表す) としたとき $D' = C$。

5. 提案手法の評価

$match_{is}$ がその他のマッチングと比較してどの程度マッチングしやすいか、定義に基づいて確認する。また、実際のマッチングの状況を確認するため実験を行う。

5.1 マッチングの比較

$match_\theta$、$match_{OI}$、$match_{vd}$、$match_{td\theta}$、$match_{tdOI}$、$match_{is}$、について枚挙される頻出パターンの数、つまり**マッチングしやすいかしにくいか**、を比較する。

定義 13 (マッチングのしやすさ) 2 つのマッチング M_1、M_2 に対し、任意のパターン P とネットワーク上のノード N が M_1 でマッチングするならば M_2 でマッチングするとき、$M_2 \geq M_1$ と表す。$M_2 \geq M_1$ かつ $M_2 \not\geq M_1$ のとき、$M_2 > M_1$ と書く。

定理 1 $match_{OI} < match_{vd} < match_\theta$

定義より $\leq$ は明らかである。$\not\leq$ の反例は容易に見つかる。

定理 2 $match_{td\theta} < match_{vd}$

証明 $match_{td\theta}$ は、パターン上の変数深度とネットワーク上の項の深度が一致するような代入しか許さないので、パターン上の変数深度が異なるノードにネットワーク上の同じノードが代入されることはない。つまり、$match_{td\theta}$ でマッチングするならば $match_{vd}$ でマッチングする、といえる。逆に、$match_{vd}$ でマッチングするが、$match_{td\theta}$ でマッチングしない場合が存在するかを考えると、図 3 のような場合が存在する。$match_{vd}$ の場合 person1 にマッチングするが、$match_{td\theta}$ ではマッチングしない。したがってマッチングのしやすさは、$match_{td\theta} < match_{vd}$ といえる。

定理 3 $match_{td\theta}$ と $match_{OI}$ の間で、マッチングのしやすさは比較できない。

証明 図 4 のパターンとネットワークを例に挙げて反例を示す。パターン 1 が person1 にマッチングするかどうかを調べると、$match_{td\theta}$ の場合マッチングするが、$match_{OI}$ の場合マッチングしない。また、パターン 2 が person2 にマッチングするかどうかを調べると、$match_{OI}$ の場合マッチングするが、$match_{td\theta}$ の場合マッチングしない。したがって、この 2 つのマッチング方法については一方的にどちらかの方がマッチングしやすいとはいえない。

定理 4 $match_{tdOI} < match_{OI}$

$match_{tdOI}$ は $match_{OI}$ に条件が追加されたものである。

定理 5 $match_{tdOI} < match_{td\theta}$

$match_{tdOI}$ と $match_{td\theta}$ は、OI 包摂か θ 包摂かの違いであるから、これらの関係がそのまま適用できる。
定理 1 から定理 5 をまとめると次のとおりである。

$$match_{tdOI} < \begin{matrix} match_{td\theta} \\ match_{OI} \end{matrix} < match_{vd} < match_\theta$$

最後に、提案した $match_{is}$ について検討する。

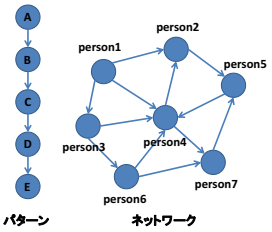

図 5: マッチングの例 3

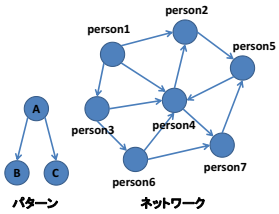

図 6: マッチングの例 4

定理 6 $match_{is} < match_{OI}$

$match_{is}$ は、$match_{OI}$ に条件が追加されたものである。

定理 7 $match_{is}$ と、$match_{td\theta}$ 及び $match_{tdOI}$ の間で、マッチングのしやすさは比較できない。

証明 図 5 と図 6 を例に挙げて反例を示す。図 5 で、パターンの person1 へのマッチングを考えると、$match_{is}$ の場合 θ = {A/person1}, ρ = {B/person3, C/person6, D/person7, E/person5} となりマッチングするが、$match_{td\theta}$ 及び $match_{tdOI}$ ではマッチングしない。また、図 6 でパターンの person1 へのマッチングを考えると、$match_{td\theta}$ 及び $match_{tdOI}$ ではマッチングするが、$match_{is}$ ではマッチングしない。したがって、$match_{is}$ は $match_{td\theta}$ 及び $match_{tdOI}$ と比較できない。

5.2 実験

定義に基づいた考察では、誘導部分グラフ同型と項の深度を考慮したマッチングを比較できなかった。そこで、各種マッチングを Hanabi において使用し、マイニングされる頻出パターンの数を確認する。また、計算量も比較する。

実験では Zachary の空手クラブネットワーク (図 7) [Zachary 77] に対して各マッチングを用いて、枚挙される頻出パターンの数、実行時間を比較をした。このネットワークはある空手クラブに所属するメンバーの交友関係を示しており、本来は無向辺であるが、有向辺とした。Zachary の空手クラブネットワークのノード数は 34 で、最小サポートは 8%とした。

枚挙された頻出パターンの数を表 4 に、実行時間を表 5 に示す。OI 包摂によるマッチングを行った際、組み合わせ爆発が起こりマイニングが完了しなかった。

実験結果から誘導部分グラフ同型に基づくマッチングは、OI 包摂+項の深度よりマイニングされるパターンの数が少なくなる傾向があると考えられる。また他手法に比べ実行時間が増えることが分かった。誘導部分グラフを抽出するコストが大きく、実行時間を削減する方法を検討する必要がある。

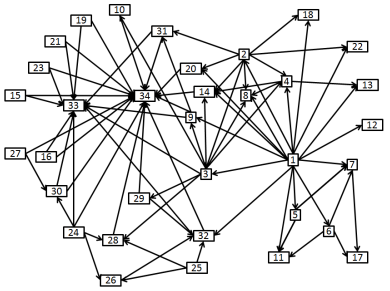

図 7: Zachary の空手クラブネットワーク

表 4: 最小サポート 8%:頻出パターン

	θ	変数深度	θ +項	OI	OI+項	誘導
1-shell	3	3	3	5	5	5
2-shell	12	7	5	62	31	18
3-shell	29	7	0	-	0	0
4-shell	14	1	0	-	0	0
5-shell	0	0	0	-	0	0
6-shell	0	0	0	-	0	0

6. まとめ

本研究では、ネットワーク上でのノード間の繋がりに着目した誘導部分グラフ同型に基づくマッチングの提案を行った。また各マッチングの性質や、枚挙されるパターン数、実行時間についての比較を行った。今後の課題としては、別のネットワークに対し同様の実験を行い、項の深度を考慮したマッチングと誘導部分グラフ同型に基づくマッチングの性質の違いをさらに詳しく調べること、大規模なネットワークに適用する際、候補パターンの数が組み合わせ爆発を起こしてしまう問題の解消及び実データへの適用、が挙げられる。

参考文献

[古川 01] 古川 康一, 尾崎 知伸, 植野 研. 帰納論理プログラミング. 共立出版, 2001.

[Motoyama 06] J. Motoyama, S. Urazawa, T. Nakano, and N. Inuzuka. A mining algorithm using property items extracted from sampled examples. In ILP ' 2006, Vol. 4455 of LNCS, pp. 335-350. Springer, 2007.

[西尾 13] 西尾 典晃, 犬塚 信博. 開いた構造を持つ事例を対象とした関係的知識発見. 第 75 回情処全国大会, 2013.

[山崎 13] 山崎 誠治, 西尾 典晃, 武藤 敦子, 犬塚 信博. 開いた構造のパターンマイニングにおける距離に基づくパターンマッチング. WiNF2013, 2013.

[Zachary 77] W. W. Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research 33*, 452-473, 1977.

表 5: 最小サポート 8%:実行時間

	θ	変数深度	θ +項	OI	OI+項	誘導
(ms)	296	219	219	-	251692	398208