Vol.2015-ICS-179 No.5
2015/3/20

ネットワークパターンマイニングにおける包摂関係を用いたアルゴリズムの改善

鷲見 俊貴[1,a)]　武藤 敦子[1]　犬塚 信博[1]

概要：述語論理上で帰納推論を展開するアプローチであり，さまざまな分類問題を解決することができるILP（帰納論理プログラミング）がある．ILP によるパターンマイニングのひとつとして花火節を用いたパターンマイニングがある．これはネットワークのあるノードと隣接するノードの集合から導出される誘導部分グラフであるエゴセントリックネットワークを連結してパターンを生成しパターンマイニングを行っていくものである．しかし，パターンの生成される段階で頻出でない多くの候補パターンが生成される問題があった．生成される候補パターンはエゴネットを連結して生成しているため互いに包摂関係になっていることが多い．そこで包摂関係を用いてアルゴリズム内でのネットワークのマッチング回数を減らす方法を提案する．

1. はじめに

社会ネットワークとは社会構造をネットワークとして考え，人や組織をノード，関係をエッジで表したものである．社会ネットワーク分析はそこから性質などを明らかにするものである．現在，社会ネットワーク分析は web サービスの発達に伴いデータ量が大規模化してきている．大規模なデータの中から有用な情報を見つけ出す手段としてデータマイニングがある．データマイニングによりデータからパターンやルールを見つけ出してそれを有効に活用することができるようになった．データマイニングの中でも複数のデータから知識を発見するものは関係型データマイニング（Multi-Relational Data Mining: MRDM) と言われている．複数のデータベースに対応するために関係型データマイニングは ILP(Inductive Logic Programming：帰納論理プログラミング [1]）で行われる．ILP とは述語論理上で帰納推論を行い知識発見をするアプローチである．社会ネットワークの中からよく現れるつながり方や構造を発見する方法はパターンマイニングといわれている．「ネットワーク」,「信頼」,「規範」を資本と考えることにより，それらが形成，蓄積されると捉える社会関係資本 [2] という概念がある．例えば，友人関係ネットワークから成績が良い学生がどのような人間関係を持っているかや，会社間のネットワークから不況に強い会社はどのような関係性があるのかといった情報が社会ネットワークの中に眠っている可能性がある．パターンマイニングは「社会関係資本」から有益な情報を見つけ出すことを目標としている．ILP によるネットワークパターンマイニングのひとつとして花火節を用いたパターンマイニング [3] がある．これはネットワークのあるノードと隣接するノードの集合から導出される誘導部分グラフであるエゴセントリックネットワークからパターンを生成しパターンマイニングを行っていくものである．この手法は計算量が多く大規模な社会ネットワークに適応するのは未だ現実的ではない．そこで本研究では包摂関係を用いてアルゴリズム内でのネットワークのマッチング回数を減らすことを提案する．また，これまでの手法と提案手法の比較を行う．

[1] 名古屋工業大学
Nagoya Institute of Technology, Gokiso-cho, Showa-ku, Nagoya, 466-8555, Japan
[a)] sumi@nous.nitech.ac.jp

2. 花火節を用いたパターンマイニング

2.1 花火節を用いたパターンマイニング

花火節を用いたパターンマイニングとはネットワークの中に表れる頻出なパターンを発見するためのアルゴリズムである．そのためにノードと隣接するノードの集合から導出される誘導部分グラフであるエゴセントリックネットワークをもとにネットワークパターンである花火節を生成する．花火節のうちネットワークの中に頻出なものだけを組み合わせて効率よく探索を行っていく．以下に詳しく定義する．

2.2 花火節

エゴセントリックネットワークについて定義する．

定義 1 (エゴセントリックネットワーク（エゴネット）) ネットワーク D の中のノード n のエゴネットはノード n と n からの距離が 1 のノードから導出される誘導部分グラフとする．このときの n を中心ノードとする．

ネットワークとマッチングを行うためにエゴネットを変数化する．

定義 2 (エゴネットの変数化) あるエゴネット E_s の変数化したものは E_v である．このとき E_s と E_v は以下の関係で表される．

- E_s=E_v θ となる代入 $\theta = \{v_1/s_1, ..., v_n/s_n\}$ が存在する
- $s_1, ..., s_n$ は全て異なる項である

定義 3 (単位花火節) 変数化したエゴネットを単位花火節と呼ぶ．エゴネットにおいて中心ノードであったノードは単位花火節では中心項とする．

定義 4 (花火節) 単位花火節は花火節である．また，花火節同士を連結したものも同様に花火節である．

例 1 人間ネットワーク N_p（図 1, 表 1）について着目すると，member（person3）のエゴネットについて考える．エゴネットは中心ノード person3 と距離 1 のノード person4, person5 から導出される誘導部分グラフ [friend(person3, person4), friend(person3, person5), friend(person4, person5)] である．エゴネットを変数化することによってできる person3 の単位花火節は [friend(x_1, x_2), friend(x_1, x_3), friend(x_2, x_3)] である．

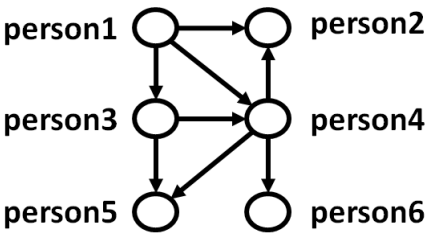

図 1 人間ネットワーク N_p

表 1 人間ネットワーク N_p 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	person3	person4
person5	person3	person5
person6	person4	person2
	person4	person5
	person4	person6

2.3 マッチング

花火節がネットワークの中に頻出に表れているか判別するため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 $G = \{V,E\}$ におけるパターン x の支持度 sup_x は以下の式 (1) で表される．

$$sup_x = \frac{\sum_{v \in V} match(G, x, v)}{|V|} \quad (1)$$

$match(G, x, v)$ はノード v を中心項に代入した時，x θ を満たすような代入 θ が G 上に存在する場合，真となる．ここでの代入方法は θ 包摂とする．θ 包摂はグラフに対して論理的に真となる代入が存在する場合，真となる方法である．

例 2 例 1 の単位花火節 C_3 = [friend(x_1, x_2), friend(x_1, x_3), friend(x_2, x_3)] を人間ネットワーク N_p（図 1, 表 1）にマッチングし，支持度 sup_{c3} を求める．分母は全ノード数 6 である．分子はマッチングに成功したノード数である．C_3 は中心項が person1 の時，代入 [friend(person1, person3), friend(person1, person4), friend(person3, person4)] を満たす．また person3 の時，代入 [friend(person3, person4), friend(person3, person5), friend(person4, person5)] も満たす．よって支持度 sup_{c3}=2/6=1/3 となる．

支持度から頻出であるかを判別するために閾値である最小サポートを設定する．最小サポートを超える支持度を持つ花火節を頻出とする．

2.4 重ね合わせ

頻出な花火節同士を連結してより大きく複雑なパターンマイニングを行う．花火節の連結を行っていくためにパターン木を定義する．

定義 6 (パターン木) 頻出な単位花火節を基本パターンとする．基本パターンをノードとして基本パターンを連結した木をパターン木とする．パターン木によってできる花火節を候補パターンとする．最小サポートを超える支持度を持つ候補パターンは頻出パターンである．候補パターンの根から最深の葉までの最短距離を深度とする．（図 2）

ある頻出パターン A と頻出パターン B の重ね合わせは以下の方法で行う．A の根ノードとエゴネットの辺を除いた部分グラフを A' とする．A' が非連結グラフの場合，複数の連結グラフ $A'_1 \cdots A'_n$ とする．B の最深の葉ノードすべてを除いたノードで構成される誘導部分グラフを B' と

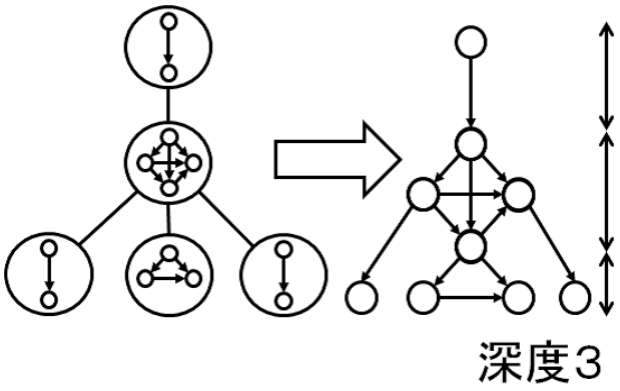

図 2　パターン木

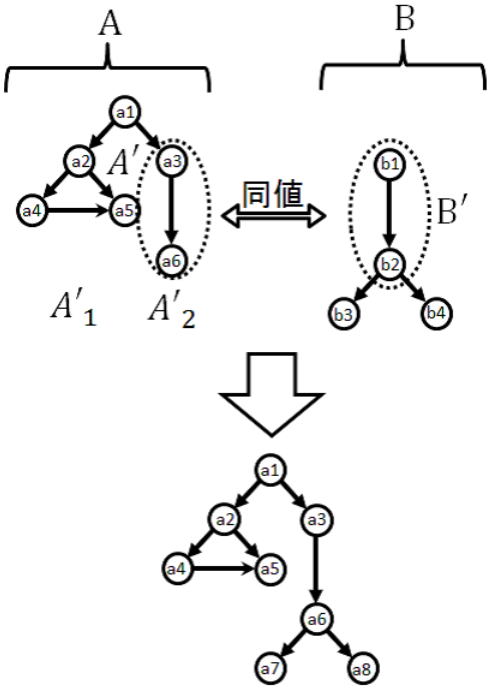

図 3　例 3 の重ね合わせ

する．ある連結グラフ A'_k と誘導部分グラフ B' が同値関係である時，頻出パターン A における A'_k を B' に置換する．

例 3　花火節 A = [friend($a1$, $a2$), friend($a1$, $a3$), friend($a2$, $a4$), friend($a2$, $a5$), friend($a4$, $a5$), friend($a3$, $a6$)] と花火節 B = [friend($b1$, $b2$), friend($b2$, $b3$), friend($b2$, $b4$)] の重ね合わせを考える．A の葉の部分を取り除くと A'_1= [friend($a2$, $a4$), friend($a2$, $a5$), friend($a4$, $a5$)], A'_2= [friend($a3$, $a6$)] となる．B の葉の部分を取り除くと B'=[friend($b1$, $b2$)] となる．この時，A'_2 と B' は同値関係である．そこで A'_2 を B' に置換し，新たな花火節 A_{New} = [friend($a1$, $a2$), friend($a1$, $a3$), friend($a2$, $a4$), friend($a2$, $a5$), friend($a4$, $a5$), friend($a3$, $a6$), friend($a6$, $a7$), friend($a6$, $a8$)] とする．（図 3）

HANABI($\mathbf{G}$, sup_{min}):
input ：ネットワーク $\mathbf{G} = \{\mathbf{V}, \mathbf{E}\}$；
最小サポート sup_{min};
output：頻出花火パターン $Freq$;

1. $\mathbf{E} := \emptyset$; $k := 1$;
2. **for each** $v \in V$ **do** $\mathcal{E} = v$ のエゴネット;
3. $\mathcal{F}_1 := \{x \in \mathbf{E} \mid \sup_x \geq sup_{min}\}$;
4. **while** $\mathcal{F}_k \neq \emptyset$ **do**
5. $\quad \mathcal{C}_{k+1} := SUMPATTERN(\mathcal{F}_k)$;
6. $\quad$ **for each** $c \in \mathcal{C}_{k+1}$ **do**;
7. $\quad\quad \sup_c = SUPPORT(G, c)$;
8. $\quad\quad$ **if** $(\sup_c \geq sup_{min})$ $\mathcal{F}_{k+1} := \mathcal{F}_{k+1} \cup c$;
9. $\quad Freq := Freq \cup \mathcal{F}_{k+1}$; $k := k + 1$;
10. **return** $Freq$;

図 4　花火節を用いたパターンマイニングのアルゴリズム

SUMPATTERN($\mathcal{T}_k$): %重ね合わせ
input ：パターン木の集合 $\mathcal{T}_k$;
output：候補パターン $\mathcal{T}_{k+1}$;

1. $\mathcal{T}_{k+1} := \emptyset$;
2. **for each** $A \in \mathcal{T}_k$ **do**
4. $\quad A' := A$ から根を除いて得られる木の集合;
5. $\quad$ **for each** $A'_i \in A'$ **do**
6. $\quad\quad$ **for each** $B \in \mathcal{T}_k$ **do**
7. $\quad\quad\quad$ **if** A'_i と B から最深の葉を除いた B' が同値
8. $\quad\quad\quad\quad \mathcal{T}_{k+1} := \mathcal{T}_{k+1} \cup$
$\quad\quad\quad\quad\quad \{A$ の A'_i の部分を B' に置換したパターン $\}$;
9. **return** $\mathcal{T}_{k+1}$;

図 5　頻出パターンの重ね合わせ

2.5　花火節を用いたパターンマイニングのアルゴリズム

花火節を用いたパターンマイニングは，はじめに頻出パターンを見つけたいネットワークからエゴネットを抽出し単位花火節を生成する．単位花火節のうち最小サポートを超える支持度をもつものを深度 1 の頻出パターン（基本パターン）とする．深度 k（k $\geq$ 2）の候補パターンは深度 k-1 の重ね合わせによって生成される．候補パターンのうち最小サポートを超える支持度をもつものを深度 k の頻出パターンとする．これを新たな頻出パターンが現れなくなるまで繰り返す．

3.　包含関係を用いた花火節を用いたパターンマイニングの改善

3.1　花火節を用いたパターンマイニングの利点と問題点

花火節を用いたパターンマイニングはアイテムが基本パターン，アイテムセットが連結した花火節の連結となる相関ルールの考え方を参考にしている．また頻出な花火節だけを連結して大きな花火節を生成する方法は Apriori アルゴリズム [4][5] を参考にしている．Apriori アルゴリズムはアイテム数が 1 つである頻出集合を見つけ，その集合に 1 つずつアイテムを追加することで大きな頻出集合を幅優先的に見つけ出すアルゴリズムである．Apriori アルゴリズムでは効率よく頻出アイテム集合を見つけるためにアイテ

ム集合同士の包含関係を利用している．例えばあるアイテム集合 A と A から 1 つアイテムを減らしたアイテム集合 B が存在する場合，A と B の間に A⊃B という関係が成り立つ．この時それぞれのアイテム集合の支持度 sup_A,sup_B 及び最小サポート sup_{min} の間には以下の関係 (2)，(3) が成り立つ．

$$sup_A \leq sup_B \tag{2}$$

$$sup_A > sup_{min} \rightarrow sup_B > sup_{min} \tag{3}$$

式 (3) より，あるアイテム集合が頻出集合であるためには，そのアイテム集合から 1 つアイテムを減らしたアイテム集合も同様に頻出集合でなければならないと言える．Apriori アルゴリズムではアイテム数が k 個の候補となるアイテム集合があるとき，アイテム数が k-1 個になる部分集合を生成し，もし部分集合のうち頻出でないアイテム集合が存在する場合，候補となるアイテム集合も頻出アイテム集合ではないと言えるので支持度を計算することなく候補から除外している．花火節を用いたパターンマイニングではこの考え方は重ね合わせの際に使用している．花火節を重ねあわせる場合，頻出パターン同士を連結している。これは連結した花火節が頻出であるためにはその花火節から 1 つ基本パターンを取り除いたパターンも頻出でなければならないためである．

一方で花火節を用いたパターンマイニングにはアイテムセットの相関ルールマイニングと相違点がある。花火節を用いたパターンマイニングでは同じ基本パターン，つまり同じアイテムを複数使用して連結することができる．また，複数のアイテムからできるアイテムセットはただ一つに定まるのに対し，複数の基本パターンからできる候補パターンは連結の仕方によって複数の候補パターンが考えられる．そのため，Apriori アルゴリズムの考え方を用いても連結によってできる候補パターンは Apriori アルゴリズムに比べて膨大になってしまうという問題がある．

3.2 包摂関係による候補パターンの削減

花火節を用いたパターンマイニングとアイテムセットの相関ルールマイニングの相違点がもう一つある．基本パターンはアイテムと違いネットワークのパターンである．そのため基本パターン同士で包摂関係になっているものが存在する．それにより，基本パターンから構成される花火節もまた包摂関係になっているものが存在する．そこで生成された候補パターン同士を比較することにより候補パターンを削減が可能である．

Apriori アルゴリズムでは効率よく頻出アイテム集合を見つけるためにアイテム集合同士の包含関係を利用している．ただし，この包含関係は先で述べたようにアイテム数が異なるアイテム集合同士で使用していた．花火節を用いたパターンマイニングにおいてもまた深度が異なるパターンで使用しており，これまでは深度が同じ候補パターン同士を比較することはなかった．

HANABI(**G**, sup_{min}):
input : ネットワーク $\mathbf{G} = \{\mathbf{V}, \mathbf{E}\}$;
最小サポート sup_{min};
output : 頻出花火パターン $Freq$;
1. $\mathbf{E} := \emptyset$; $k := 1$;
2. **for each** $v \in V$ **do** $\mathcal{E} = v$ のエゴネット;
3. $\mathcal{F}_1 := \{x \in \mathbf{E} \mid \mathrm{sup}_x \geq sup_{min}\}$;
4. **while** $\mathcal{F}_k \neq \emptyset$ **do**
5. $\quad \mathcal{C}_{k+1} := SUMPATTERN(\mathcal{F}_k)$;
6. $\quad$ **for each** $c \in \mathcal{C}_{k+1}$ **do**;
7. $\quad\quad \mathrm{sup}_c = SUPPORT(G, c)$;
8. $\quad\quad$ **if** $(\mathrm{sup}_c \geq sup_{min})$ $\mathcal{F}_{k+1} := \mathcal{F}_{k+1} \cup c$;
%% 提案手法 %%%%%%%%%%
9. $\quad\quad$ **else for each** $e \in \mathcal{C}_{k+1}$ **do**;
10. $\quad\quad\quad$ **if** $(c \subseteq e)$ $\mathcal{C}_{k+1} := \mathcal{C}_{k+1} - e$;
%%%%%%%%%%%%%%%%%%%%%%
11. $\quad Freq := Freq \cup \mathcal{F}_{k+1}$; $k := k + 1$;
12. **return** $Freq$;

図 6 提案手法の花火節を用いたパターンマイニングのアルゴリズム

3.3 提案手法

候補パターンから頻出パターンを発見するためには候補パターンを実際のネットワークとマッチングを行い，支持度を計算する必要がある．そこでネットワークパターンの支持度が最小サポートを満たさなかった場合，その拡大パターンにあたるパターンも頻出でないといえることから，頻出でない候補パターンをまだネットワークとマッチングを行っていない候補パターンと比較を行い，拡大パターンである候補パターンをネットワークとマッチングすることなく候補から除外する．以下に提案手法の花火節を用いたパターンマイニングのアルゴリズム（図 6）を示す．

3.4 候補パターン同士の比較

包摂関係を明らかにするために候補パターン同士を比較する必要がある．候補パターンも小さなネットワークの一つであり，比較することは容易ではない．そこで単純に小さなネットワーク同士を比較するのではなく，候補パターンは単位花火節同士で構成されていることから，根より順に単位花火節同士を再帰的に比較することで大きな候補パターンの包摂関係を明らかにする．

例 4 あるネットーワークのエゴネットから生成される基本パターンは $basic_x$=[friend(x_1, x_2)], $basic_y$=[friend(y_1, y_2), friend(y_1, y_3), friend(y_2, y_3)], $basic_z$=[friend(z_1, z_2), friend(z_1, z_3), friend(z_1, z_4), friend(z_2, z_3), friend(z_2, z_4), friend(z_3, z_4)] であり，$basic_x \subseteq basic_y \subseteq basic_z$ という関係が成り立っている．（図 7）この時，頻出でなかったパターン A=[friend(a_1, a_2), friend(a_1, a_3), friend(a_2, a_3), friend(a_2, a_4), friend(a_3, a_5), friend(a_3, a_6), friend(a_5, a_6)] と候補

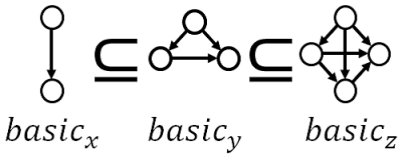

図 7　例 4 の基本パターンの包含関係

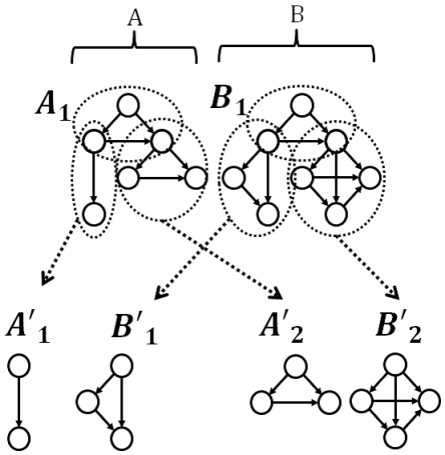

図 8　例 4 の候補パターンの比較

パターン B=[friend(b_1, b_2), friend(b_1, b_3), friend(b_2, b_3), friend(b_2, b_4), friend(b_2, b_5), friend(b_4, b_5), friend(b_3, b_6), friend(b_3, b_7), friend(b_3, b_8), friend(b_6, b_7), friend(b_6, b_8), friend(b_7, b_8)] の比較を行う．A と B の根の部分の基本パターン A_1,B_1 はともに $basic_y$ である．よって $A_1 \subseteq B_1$ である．次に A の根の部分を除いてできる複数の非連結グラフを A'_1,A'_2 とする．同様に B の非連結グラフも B'_1,B'_2 とする．このとき A'_1 と B'_1、A'_2 と B'_2 を再帰的に比較する．A'_1 と B'_1 はそれぞれ基本パターン $basic_x$, $basic_y$ である．これより $A'_1 \subseteq B'_1$ である．また A'_2 と B'_2 はそれぞれ基本パターン $basic_y$，$basic_z$ である．よって $A'_2 \subseteq B'_2$ である．A を構成している基本パターンはすべて B を構成している基本パターンの部分パターンなので $A \subseteq B$ である．よって B も頻出でないといえる．（図 8）

4. 実験

4.1 実験

Zachary の空手クラブネットワーク [6]（図 9）に対して従来法と提案手法のネットワークとのマッチング回数および頻出パターン発見の実行時間の比較を行った．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いる．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５％と１０％で行う．またネットワークのエッジは本来無向辺だが，簡単のため有向辺で行った．

4.2 実験結果

最小サポート５％と 10 ％でパターンマイニングした時のマッチング回数は図 10 と図 11 の通りとなった．候補の数が大幅に増える深度 2, 深度 3, 深度 4 では包摂関係になっているものが多く，多くの候補パターンが頻出でないと識別された．実行時間は表 2 の通りとなった．マッチング回数が減ったことに伴い実行時間が減少した．これはネットワークとのマッチングより候補パターンの比較のほうが簡単であったためだと考える．

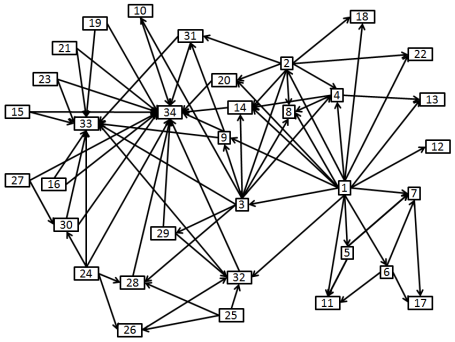

図 9　Zachary の空手クラブネットワーク

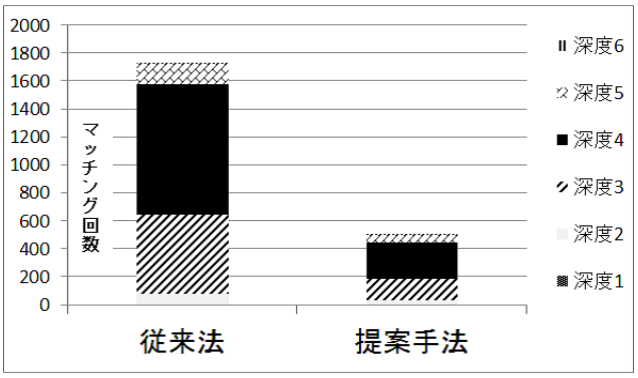

図 10　最小サポート５％でのマッチング回数

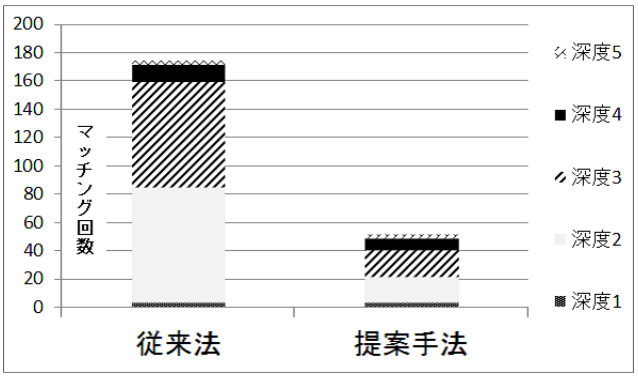

図 11　最小サポート 10 ％でのマッチング回数

表 2　アルゴリズム修正前後の実行時間

(ms)	従来法	提案手法
5 %	5875.3	3797.3
10 %	224.7	48.4

5. まとめ

候補パターン同士を比較することで花火節を用いたパターンマイニングのマッチング回数を減らすことを提案した．提案手法を実装して従来法と比較を行い、マッチング回数の削減及び実行時間が短縮したことを確認した．今後の課題として花火節を用いたパターンマイニングのさらなる改善が必要と考える．特にまだネットワークからエゴネット生成の段階での計算量が多いのでエゴネット抽出の効率化が必要である．そして，花火節により挙げられた頻出パターンから実際の社会ネットワーク分析を行っていきたいと考えている．

参考文献

[1] 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).
[2] 佐藤 寛；援助と社会関係資本 — ソーシャルキャピタル論の可能性—，日本貿易振興会アジア経済研究所（2002）.
[3] Noriaki Nishio, Atsuko Mutoh and Nobuhiro Inuzuka；“On Computing Minimal Generators in Multi-Relational Data Mining with respect to theta-Subsumption”, CEUR Workshop Proceedings (Late Breaking Papers of 22nd Inductive Logic Programming Conference), pp. 50-55（2012）.
[4] 岡田 孝, 元田 浩；相関ルールとその周辺，オペレーションズ・リサーチ：経営の科学（2002）.
[5] 宇野 毅明, 有村 博紀；頻出パターン発見アルゴリズム入門 − アイテム集合からグラフまで −，人工知能学会全国大会 AI レクチャー 「先端 AI」（2008）.
[6] W. W. Zachary；An information flow model for conflict and fission in small groups，*Journal of Anthropological Research 33*, 452-473,（1977）.