平成 25 年度　修士論文概要	
開いた構造のパターンマイニングにおける 論理的手法に関する研究	犬塚・武藤研究室　山崎 誠治
	学籍番号　No. 24417612

1　はじめに

近年, **社会ネットワーク分析**の研究が活発に行われている. 社会ネットワーク分析 (Social Network Analysis: SNA) とは, 行為者の関係性に着目して現象をとらえようとする方法論である. 社会ネットワークを示したデータベースもソーシャルネットワーキングサービスなどの発展に伴い大規模化してきている.

大規模なデータベースから知識を発見する手法として**データマイニング**がある. 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) の枠組みで行われてきた.

MRDM において代表的な手法として WARMR[1] がある. これはパターンの完全探索が可能だが計算時間が大きい. これに対し, 基本パターンの組み合わせによりパターンを探索する MAPIX[2] が考案された. しかし, 社会ネットワークのようなデータベースに対応していなかった.

社会ネットワークから知識を効率的に探索する手法として花火節を用いたパターンマイニング [3] がある. この手法では, 社会ネットワークからのパターンマイニングを高速に行うことが可能である.

しかし, この手法では社会ネットワーク固有の特徴について考慮されていなかった. 本研究では, 花火節に距離の概念を導入したマッチングを行い, 社会ネットワークにより適したマイニングを行えるようにする.

2　ILP におけるパターンマイニング

パターンマイニングにおけるデータベースはターゲットテーブルと呼ばれるテーブルをただ 1 つだけ持ち, そのリテラルをターゲットと呼ぶ. マイニングはターゲットに関する知識を探索する. また, リテラルの引数には入力 (+)/出力 (-) の情報を与え, 探索空間を制御する.

パターンは節の形式で表され後件がターゲット, 前件がそれ以外のリテラルの連言で構成される. パターンマイニングはパターンとデータベースのマッチングを行い, 頻出なものを枚挙する.

マッチングはパターンとデータベースのマッチング以外にパターン同士のマッチングがある. パターン同士のマッチングでは 2 つのパターンが同値かどうかを判定し, その強さを分解能と呼ぶとする.

よく扱われるマッチング方法として θ 包摂がある.

定義 1　θ 包摂

2 つの節 S_1, S_2 に対して, $S_1\theta \subseteq S_2$ となる代入 θ が存在するとき, S_1 は S_2 を包摂する *($S_1 \succeq S_2$)* という. □

もう一つのマッチング方法として OI 包摂がある.

定義 2　OI 包摂

2 つの節 S_1, S_2 に対して, 以下の条件を満たす代入 $\theta = \{A_1/B_1, A_2/B_2, ..., A_n/B_n\}$ が存在するとき, S_1 は S_2 を包摂する *($S_1 \sqsupseteq S_2$)* という.

- $S_1\theta \subseteq S_2$
- $B_i \neq B_j (i \neq j)$

□

θ 包摂において節 S とデータベース D は $S \succeq N$ のとき S は N にマッチするとし, 2 つの節 S_1, S_2 に関して $S_1 \succeq S_2$ かつ $S_1 \preceq S_2$ のとき S_1 と S_2 は同値とみなす. OI 包摂においても同様である.

3　SNA におけるパターンマイニング

SNA におけるパターンマイニングでは, 社会ネットワーク上に頻出な構造を枚挙することでネットワークの分析に役立てる. 社会ネットワークを示すデータベースは事例が独立していない. このような各事例が独立していない構造を開いた構造と呼ぶ.

本研究では, 対象とするデータベースは以下のような性質をもつものとする.

- 開いた構造を持つ
- データベース内の項の型は全て同じ
- ターゲットテーブルに現れる項の集合を $Term_{target}$, ターゲットテーブル以外のテーブルに現れる項の集合を $Term_{others}$ とすると
$$Term_{target} \supseteq Term_{others}$$
となる
- データベース内に現れるリテラルの引数は 1 入力 1 出力または, 1 入力のみ

4　花火節

開いた構造を持つデータベースに対して, 社会ネットワーク分析の考え方であるエゴセントリックネットワーク [4] から着想し, パターンマイニングを行う.

ILP 分野におけるエゴセントリックネットワークと類似した構造である近傍について述べる. あるターゲット e の近傍 N とは e と, 以下のいずれかを満たすリテラルの集合である.

- 全ての入力引数が e の基礎項であるリテラル
  このときの出力項を近傍基礎項と呼ぶ.
- 全ての引数が e の近傍基礎項であるリテラル

この近傍を変数化したものを基本パターンとし, 単位花火節と呼ぶ.

この単位花火節どうしを連結することで複雑なパターンのマイニングを行う.

花火節はこの単位花火節とそれを連結してできる節である.

花火節の連結に関する情報を保存しておくために, パターン木を用いる. パターン木のノードは単位花火節を示し, エッジはその連結の情報を示している.

またパターン木の深さがkのパターンを(k+1)-shellと呼ぶものとする.

5　距離に基づくマッチング

花火節と社会ネットワークのマッチングにおいて, 変数深度と項の深度を用いたマッチングを考える.

5.1　変数深度を考慮したマッチング

変数深度とはパターンのターゲットの変数からある変数にたどり着くのに最低でいくつの連鎖が必要かを示しており, 変数 v の深度を $d_v(v)$ と表す. 変数深度を用いたマッチングは以下のようになる.

定義 3 変数深度を考慮したマッチング
ある節 C とデータベース D において, 以下の条件を満たす代入 $\theta = \{A_1/B_1, A_2/B_2, ..., , A_n/B_n\}$ が存在するとき C は D にマッチするという.

- $C\,\theta \subseteq D$
- $d_v(A_i) \neq d_v(A_j)$ ならば $B_i \neq B_j$

□

このマッチングにおいてパターン同士のマッチングは θ 包摂を用いるとする.

5.2　項の深度を考慮したマッチング

項の深度とは社会ネットワーク上の注目している項からある項にたどり着くのに最低でいくつの辺を通るのかを示しており, 項 e から項 t への深度を $d_t(e,t)$ と表す.

定義 4 項の深度を考慮したマッチング
ある節 C とデータベース D において, 以下の条件を満たす代入 $\theta = \{A_1/B_1, A_2/B_2, ..., , A_n/B_n\}$ が存在するとき C は D にマッチするという.

- $C\,\theta \subseteq D$
- $d_v(A_i) = d_t(e, B_i)$
  このとき e は C のヘッドに代入された値である.

□

このマッチングにおいてパターン同士のマッチングは θ 包摂と OI 包摂の 2 種類を用いるとする.

5.3　マッチングの比較

θ 包摂, OI 包摂, 変数深度を考慮したマッチング (変数深度を考慮), θ 包摂を用いた項の深度を考慮したマッチング (θ 包摂+項の深度), OI 包摂を用いた項の深度を考慮したマッチング (OI 包摂+項の深度) の性質の比較を行う. 各マッチングの分解能の強さは

$$\begin{matrix}\theta\text{ 包摂} \\ \text{変数深度を考慮} \\ \theta\text{ 包摂}+\text{項の深度}\end{matrix} \quad < \quad \begin{matrix}\text{OI 包摂} \\ \text{OI 包摂}+\text{項の深度}\end{matrix}$$

である. また, ネットワークとのマッチングにおける枚挙されるパターン数の関係は

$$\text{OI 包摂}+\text{項の深度} \quad < \quad \begin{matrix}\text{OI 包摂} \\ \theta\text{ 包摂}+\text{項の深度}\end{matrix}$$

$$< \quad \text{変数深度を考慮} \quad < \quad \theta\text{ 包摂}$$

となる.

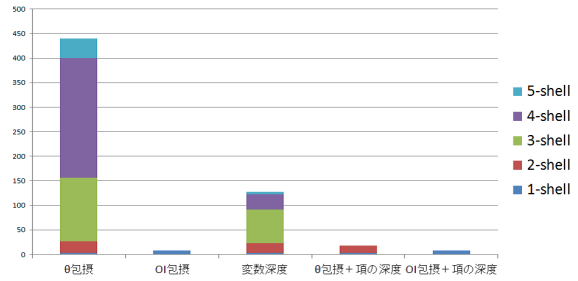

図 1: 実験結果 1

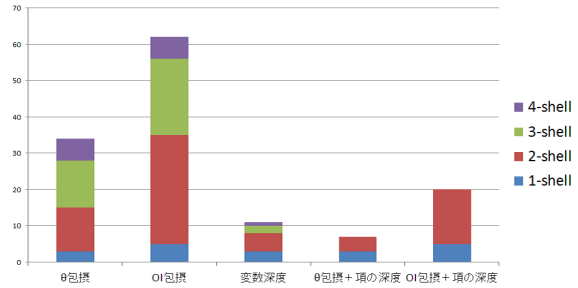

図 2: 実験結果 2

6　実験

実験では Zachary の空手クラブネットワーク [5] に対して各マッチング方法を適用した. ノード数は 34 で最小サポートは５％と１０％で行った. またネットワークのエッジは本来は無向辺だが, 簡単のため有向辺で行った. 図 1 は最小サポート 5 %, 図 2 は最小サポート 10 %において枚挙されたパターン数である. なお, 5 %において OI 包摂と OI 包摂+項の深度は 2-shell の候補集合を生成する際に組み合わせ爆発がおき, 探索が終了しなかった.

7　おわりに

本研究では花火節に対して距離に基づくマッチングを提案した. また, Zachary の空手クラブネットワークに適応した. 今後の課題として, より大きな規模のデータベースへの適応, 見つかったパターンの意味的な考察などがあげられる.

参考文献

[1] L. Dehaspe H. Toivonen: Discoveryof frequent datalog patterns, Data Mining and Knowledge Discovery, Vol. 3 No. 1 , pp7-36, 1999.

[2] Y. Nakano N. Inuzuka: Multi-relational pattern mining based-on combination of properties with preserving their structure in examples, ILP ' 2010, Vol. 6489 of LNCS, pp. 181189. 2011.

[3] N. Nishio, N. Inuzuka: Multi-Relational Pattern Mining in Open-Ended Object Domains. ILP ' 2013.

[4] N. Inuzuka, S. Takeuchi H. Matsushima: Pattern Mining on Ego-Centric Networks of Friendship Networks, Knowledge-Based and Intelligent Information and Engineering Systems, LNCS, 6884, Springer, pp.89-97, 2011.

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