3A1-4

イベント列からの頻出多部エピソードの効率的な抽出

Efficient Mining Algorithms for Partite Episodes

河東 孝[*1] 有村 博紀[*2] 平田 耕一[*3]
Takashi Katoh Hiroki Arimura Kouichi Hirata

[*1][*2]北海道大学 大学院情報科学研究科
Graduate School of Information Science and Technology, Hokkaido University

[*3]九州工業大学大学院 情報工学研究院
Department of Artificial Intelligence

In this paper, we study the problem of mining frequent *partite* episodes *efficiently* from an input event sequence. Then, we design algorithm FreqPartite for extracting all of the frequent partite episodes from a given event sequence.

1. はじめに

時間に依存するデータからの頻出パターン抽出は，データマイニングの領域で非常に重要な問題である．このような系列データマイニング問題に対して，マニラらは [10]，入力イベント列から**エピソード**と呼ばれるパターンを抽出する**エピソードマイニング**を導入した．

これまで，エピソードの部分クラス [7, 8, 2, 4, 5, 6, 10] に対して，入力列から頻出エピソードを抽出するさまざまなアルゴリズムが開発されている（図 1）．河東らが開発したアルゴリズム Kpar [6] は，イベント集合の列である**多部エピソード** [6] に対して，入力列からすべての頻出多部エピソードを抽出するアルゴリズムである．アルゴリズム Kpar は，すべての頻出多部エピソードの候補に対して，そのエピソードの頻度を入力列を走査することで計算する．一方，本論文では，多部エピソードの極小出現を利用することで，より**効率よく**頻出多部エピソードを抽出するアルゴリズム FreqPartite を開発する．入力列 $\mathcal{S}$ と，アルファベットサイズ $|\Sigma|$，$\mathcal{S}$ の長さ n に対して，アルゴリズム FreqPartite は，$\mathcal{S}$ からすべての頻出多部エピソードを，1 出力あたり $O(|\Sigma|^3 n)$ 時間と $O(|\Sigma|^2 n)$ 領域で重複なく抽出する．

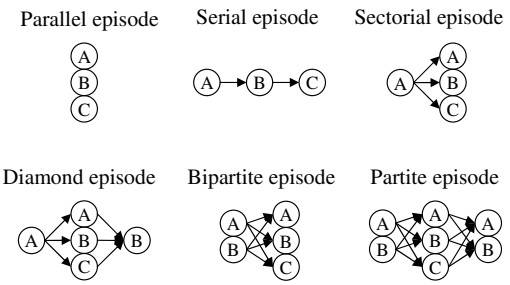

図 1: エピソードの部分クラスの例．並列エピソード (マニラら [10]), および，直列エピソード (マニラら [10]), 扇状エピソード (河東ら [7]), 菱形エピソード (河東ら [8, 4]), 二部エピソード (河東ら [5]), 多部エピソード (河東ら [6]，および，本論文).

2. エピソードマイニング

本章では，頻出エピソードマイニング問題と，以下の議論に必要な概念について述べる．以下では，すべての整数の集合を $\mathbf{Z}$，すべての自然数の集合を $\mathbf{N}$ と書く．集合 S に対して，S の要素数を $|S|$ と書く．

2.1 入力イベント列

有限のアルファベットを $\Sigma = \{1, \ldots, m\}$ $(m \geq 1)$ とする．このとき，$e \in \Sigma$ を **イベント** という．an *event* [*1]．アルファベット Σ 上の **入力イベント列** (**入力列**と略す) $\mathcal{S}$ は，イベントの有限列 $\langle S_1, \ldots, S_n \rangle \in (2^\Sigma)^*$ $(n \geq 0)$ である．このとき，任意の $1 \leq i \leq n$ に対して，集合 $S_i \subseteq \Sigma$ を i 番目の **イベント集合**という．任意の $i < 0$ または $i > n$ に対して，$S_i = \emptyset$ と定義する．入力列 $\mathcal{S}$ に対して，$\mathcal{S}$ の **長さ** $|\mathcal{S}|$ を n と定義し，$\mathcal{S}$ の **大きさ** $||\mathcal{S}||$ を $\sum_{i=1}^{n} |S_i|$ と定義する．

2.2 エピソード

マニラら [10] は，エピソードをイベントの半順序として定義した．本論文では，エピソードをラベル付き非巡回有向グラフとして以下のように定義する．アルファベット Σ 上の**エピソード**は，ラベル付き非巡回有向グラフ $X = (V, E, g)$ である．ここに，V は頂点集合，および，$E \subseteq V \times V$ は有向辺集合，$g : V \to \Sigma$ は頂点からイベントへの写像である．

エピソード $X = (V, E, g)$ に対して，X の**大きさ** $||X||$ を $|V|$ と定義する．頂点集合 V 上の有向辺集合 E に対して，E の**推移閉方** E^+ を有向辺の集合 $E^+ = \{ (u, v) \mid u \text{ から } v \text{ への有向道が存在する} \}$ と定義する．

定義 1 (埋め込み) エピソード $X_i = (V_i, E_i, g_i)$ $(i = 1, 2)$ に対して，以下のような写像 $f : V_1 \to V_2$ が存在するとき，X_1 は X_2 に**埋め込まれている**といい，$X_1 \sqsubseteq X_2$ と書く．(i) f は**頂点のラベルを保存する**，すなわち，任意の $v \in V_1$ に対して $g_1(v) = g_2(f(v))$ を満たす，および，(ii) f は**順序関係を保存**

連絡先: 氏名: 河東 孝, 所属: 北海道大学 大学院情報科学研究科 CS 専攻, 住所: 〒 060-0814 札幌市北区北 14 条西 9 丁目, 場所: 情報科学研究科棟 7 F，情 706 号室, TEL: 011-706-7680, FAX: 011-706-7680, Email: t-katou@ist.hokudai.ac.jp

[*1] マニラら [10] は，要素 $e \in \Sigma$ を **イベント型**，要素 e の出現を **イベント** と定義した．本論文では，これら両方を単に **イベント**という．

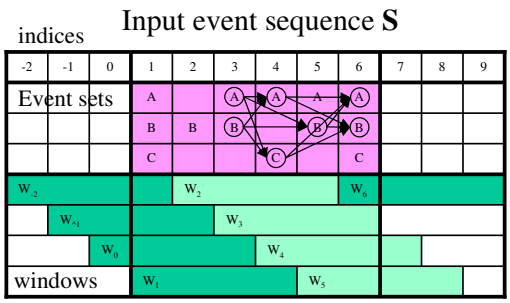

図 2: アルファベット $\Sigma = \{A, B, C\}$ 上の，長さ $n = 6$ の入力列 $\mathcal{S} = (S_1, \ldots, S_6)$ と，$\mathcal{S}$ 中の 4 ウィンドウ．この例では，入力列 $\mathcal{S}$ のウィンドウ W_3 に，多部エピソード $X = \langle\{A, B\}, \{A, B, C\}, \{A, B\}\rangle$ が出現する（埋め込まれている）ことを円と矢線で示す．

する，すなわち，$u \neq v$ を満たす任意の $u, v \in V$ に対して，もし $(u, v) \in E_1$ ならば，$(f(u), f(v)) \in (E_2)^+$ を満たす．このような写像 f を X_1 から X_2 への**埋め込み**という．

入力列 $\mathcal{S} = \langle S_1, \ldots, S_n\rangle \in (2^\Sigma)^*$ に対して，$\mathcal{S}$ の連続した部分列を**ウィンドウ**という．整数 s, t $(s < t)$ に対して，$W(\mathcal{S}, s, t)$ を $\mathcal{S}$ のウィンドウ $\langle S_s \cdots S_{t-1}\rangle \in (2^\Sigma)^*$ とする．このとき，$t - s$ を $W(\mathcal{S}, s, t)$ の**幅**といい，幅 w のウィンドウを w **ウィンドウ**という．

定義 2 (エピソードの出現) エピソード $X = (V, E, g)$ とウィンドウ $W = \langle S_1 \cdots S_w\rangle \in (2^\Sigma)^*$ に対して，以下のような写像 $h : V \to \{1, \ldots, w\}$ が存在するとき，X は W に**出現する**といい，$X \sqsubseteq W$ と書く．(i) h は頂点のラベルを保存する，すなわち，任意の $v \in V$ に対して，$g(v) \in S_{h(x)}$ を満たす，かつ，(ii) h は順序関係を保存する，すなわち，$u \neq v$ を満たす任意の $u, v \in V$ に対して，もし $(u, v) \in E$ ならば $h(u) < h(v)$ を満たす．このような写像 h を X から W への**埋め込み**という．

ウィンドウ幅は正の整数 $1 \leq w \leq n$ である．$W_i = W(\mathcal{S}, i, i + w)$ を $\mathcal{S}$ における i 番目の w ウィンドウといい，$\mathbf{W}_i^{\mathcal{S},w}$ と書く．このとき，任意の $-w + 1 \leq i \leq n$ に対して，$X \sqsubseteq W_i$ を満たすならば，エピソード X は，入力列 $\mathcal{S}$ の位置 i に **出現する** という．この i を $\mathcal{S}$ における X の**出現**，または，**ラベル**という．集合 $\mathbf{W}_{\mathcal{S},w}$ を $\mathcal{S}$ におけるすべて幅 w ウィンドウのラベルの集合 $\mathbf{W}_{\mathcal{S},w} = \{\, i \mid -w + 1 \leq i \leq n \,\}$ と定義する．さらに，エピソード X に対して，$\mathbf{W}_{\mathcal{S},w}(X)$ を $\mathcal{S}$ における X の出現する w ウィンドウのラベルの集合 $\mathbf{W}_{\mathcal{S},w}(X) = \{\, -w + 1 \leq i \leq n \mid X \sqsubseteq W_i \,\}$ と定義する．

整数の組 $[s, t)$ $1 \leq s \leq t$ を**区間**という．入力列 $\mathcal{S}$ とエピソード X に対して，整数 s と t が，以下の条件を満たすとき，$[s, t)$ は X の**極小出現である**という．(i) $X \sqsubseteq W(\mathcal{S}, s, t)$，(ii) $X \not\sqsubseteq W(\mathcal{S}, s + 1, t)$，(iii) $X \not\sqsubseteq W(\mathcal{S}, s, t - 1)$．$\mathcal{S}$ における X のすべての極小出現の集合を極小出現リストという．

2.3 頻出エピソードマイニング問題

エピソードの部分クラスを $\mathcal{C}$ とし，X を $\mathcal{C}$ 中のエピソード，$\mathcal{S}$ を入力列，w (≥ 1) をウィンドウ幅とする．このとき，$\mathcal{S}$ における X の**頻度** $freq_{\mathcal{S},w}(X)$ は，W の出現する w ウィンドウの数として定義される．すなわち，$freq_{\mathcal{S},w}(X) = |\mathbf{W}_{\mathcal{S},w}(X)| = O(|\mathcal{S}|)$ である．**最小頻度**は任意の整数 $\sigma \geq 1$ である．エピソード X が，$freq_{\mathcal{S},w}(X) \geq \sigma$ を満たすとき，X は $\mathcal{S}$ **で** σ **頻出である**という．入力列 $\mathcal{S}$ からすべての頻出多部エピソードを抽出するアルゴリズム FreqPartite.

```
algorithm FreqPartite(S, w, Σ, σ)
入力: 長さ n の入力列 S ∈ (2^Σ)*
ウィンドウ幅 w > 0, アルファベット Σ,
最小頻度 1 ≤ σ ≤ n + k;
出力: 頻出多部エピソード; {
1   T := ∅;
2   foreach ( a ∈ Σ ) do
3       R := ⟨{a}⟩;
4       多部エピソード R の極小出現リスト l_c を
5             S と a から計算する;
6       R の頻度 f を l_c と w から計算する;
7       if (f ≥ σ) then
8           T := T ∪ {(a, l_c)};
9           output R;
10      end if
11  end foreach
12  PartiteRec(⟨⟩, T, w, T, σ);
}
```

図 3: 入力列 $\mathcal{S}$ からすべての頻出多部エピソードを抽出するアルゴリズム FreqPartite.

あるという．入力列 $\mathcal{S}$ に出現するすべての σ 頻出なエピソードの集合を $\mathcal{F}_{\mathcal{S},w,\sigma}$ と書く．

定義 3 頻出エピソードマイニング問題:
クラス $\mathcal{C}$ をエピソードの部分クラスとする．入力列 $\mathcal{S} \in (2^\Sigma)^*$ と，ウィンドウ幅 $w \geq 1$，最小頻度 $\sigma \geq 1$ が与えられたとき，$\mathcal{S}$ にウィンドウ幅 w のもとで出現する，すべての σ 頻出なクラス $\mathcal{C}$ 中のエピソード X を重複なく発見する問題を，頻出エピソードマイニング問題という．

3. 多部エピソード

定義 4 自然数 $k \geq 1$ に対して，Σ 上の k **部エピソード** (または **多部エピソード**) は以下の条件 (i) – (iii) を満たすエピソード $X = (V, E, g)$ である．

(i) 任意の整数 i と j $(1 \leq i < j \leq k)$ に対して，$V = V_1 \cup \cdots \cup V_k$ となる頂点集合 V_i が存在し，$V_i \neq \emptyset$ かつ，$V_i \cap V_j = \emptyset$ を満たす．

(ii) $E = (V_1 \times V_2) \cup \cdots \cup (V_{k-1} \times V_k)$ を満たす．

(iii) 任意の $1 \leq i \leq k$ に対する，集合 V_i の異なる二つの要素 u と v に対して，$g(u) \neq g(v)$ を満たす．

本論文では，このような k 部エピソードを k 項組 $X = \langle A_1, \ldots, A_k\rangle$ として書く，ここに，任意の $1 \leq i \leq k$ に対して，$A_i = \{\, a \mid a = g(v) \text{ for some } v \in V_i \,\}$ はイベントの集合である．このとき，k を X の**長さ**という．さらに，$\langle A_1, ..., (A_k - \{\max A_k\})\rangle$ を X の**頭**といい，$\{\max A_k\}$ を X の**尾**という．

図 2 に，入力列 $\mathcal{S}$ と，$\mathcal{S}$ 中の 4 ウィンドウ，$W_3^{\mathcal{S},4}$ に出現する多部エピソード $\langle\{A, B\}, \{A, B, C\}, \{A, B\}\rangle$ の例を示す．

4. アルゴリズム

本章では，入力列からすべての頻出多部エピソードを抽出するアルゴリズム FreqPartite を設計する．

アルゴリズムは，小さな多部エピソードから大きな多部エピソードへ広がる探索空間を深さ優先探索することで，すべて

```
procedure PartiteRec(P = ⟨A_1, ..., A_k⟩, T, w, I, σ)
入力: 多部エピソードの頭 P,
多部エピソードの尾とその極小出現リストの組の集合 T,
ウィンドウ幅 w > 0,
大きさ 1 の多部エピソードと
その極小出現リストの組の集合 I,
アルファベット Σ;
出力: l ≥ k, および, ||R|| > ||P|| + 1,
B_i = A_i(1 ≤ i < k), B_k ⊇ A_k, を満たす
頻出多部エピソード R = ⟨B_1, ..., B_l⟩;
{
1  foreach ( (a, l_a) ∈ T ) do // P の長さを伸ばす
2      Q := ⟨A_1, ..., (A_k ∪ {a}), ∅⟩; U := ∅;
3      foreach ( (b, l_b) ∈ I ) do
4          R := ⟨A_1, ..., A_k, {a}, {b}⟩;
5          多部エピソード R の極小出現リスト l_c を
6              l_a と l_b から計算する;
7          R の頻度 f を l_c と w から計算する;
8          if ( f ≥ σ) then
9              U := U ∪ {(b, l_c)};
10             output R;
11         end if
12     end foreach
13     PartiteRec (Q, U, w, I, σ);
14 end foreach
15 foreach ( (a, l_a) ∈ T ) do // 集合 A_k を拡張する
16     Q := ⟨A_1, ..., (A_k ∪ {a})⟩; U := ∅;
17     foreach ( (b, l_b) ∈ T such that b > a) do
18         R := ⟨A_1, ..., (A_k ∪ {a, b})⟩;
19         多部エピソード R の 極小出現リスト l_c を
20             l_a と l_b から計算する;
21         R の頻度 f を l_c と w から計算する;
22         if ( f ≥ σ ) then
23             U := U ∪ {(b, l_c)};
24             output R;
25         end if
26     end foreach
27     PartiteRec (Q, U, w, I, σ);
28 end foreach
}
```

図 4: アルゴリズム FreqPartite で使用する再帰関数 PartiteRec.

の頻出多部エピソードを列挙する．探索空間は，多部エピソードの親子関係によって定義される．

定義 5 長さ 0 の多部エピソード $\langle\rangle$ を**根エピソード**と定義する．長さ $i > 1$ の多部エピソード $X = \langle A_1, \ldots, A_i\rangle$ の**親**を以下のように定義する．

$$parent(\langle A_1, \ldots, A_i\rangle) = \begin{cases} \langle A_1, \ldots, A_{i-1}\rangle, & |A_i| = 1 \text{ のとき}, \\ \langle A_1, \ldots, (A_i - \{\max A_i\}\rangle, & |A_i| > 1 \text{ のとき} \end{cases}$$

多部エピソードの親子関係によって，根エピソードを根として，すべての多部エピソードがそれぞれ頂点となる家系木が定義される．

図 3 にアルゴリズム FreqPartite を示す．アルゴリズム FreqPartite は，まず，すべてのアルファベットの要素 a に対して，大きさ 1 の多部エピソード $R = \langle\{a\}\rangle$ を作成し，R の極小出現リストを計算する．次に，R の極小出現リストから R の頻度を計算し，頻出エピソードとその出現リストを再帰関数 PartiteRec にわたす．

図 4 に再帰関数 PartiteRec を示す．再帰関数 PartiteRec は，CHARM [14] と同様に，頭が共通の頻出多部エピソードごとに呼び出される．PartiteRec は，まず，頭が共通の頻出多部エピソードの集合とその極小出現リストを受け取る．このとき，頻出多部エピソードは共通の頭 P と尾に分解され，尾と極小出現リストの組として入力される．PartiteRec は，入力された頻出エピソードそれぞれに対して，その子エピソードを作成する．このとき親の極小出現リストを用いて子の極小出現リストを作成する．次に PartiteRec は，作成した子エピソードのうち頻出なものを出力す，エピソードを頭と尾に分解したのち PartiteRec を再帰呼び出しする．

多部エピソードの定義とエピソードの出現の定義より，以下の補題が成立する．

補題 1 入力列 $\mathcal{S}$ と，多部エピソード $X = \langle A_1, ..., A_k\rangle$ と $Y = \langle\{a\}\rangle$ $(k \geq 1)$，区間 $[s_x, t_x)$ と $[s_y, t_y)$ に対して，$X \sqsubseteq W(\mathcal{S}, s_x, t_x)$ かつ $Y \sqsubseteq W(\mathcal{S}, s_y, t_y)$ ならば，$\langle A_1, ..., A_k, a\rangle \sqsubseteq W(\mathcal{S}, s_x, t_y)$ である．

また，多部エピソードは非並列エピソード [3] なので，以下の補題が成立する．

補題 2 入力イベント列 $\mathcal{S}$ と，多部エピソード $X = \langle A_1, ..., A_{k-1}, B\rangle$ と $X = \langle A_1, ..., A_{k-1}, C\rangle$ $(k > 1)$，区間 $[s_x, t_x)$ と $[s_y, t_y)$ に対して，$X \sqsubseteq W(\mathcal{S}, s_x, t_x)$ かつ $Y \sqsubseteq W(\mathcal{S}, s_y, t_y)$ ならば，$\langle A_1, ..., A_k, (B \cup C)\rangle \sqsubseteq W(\mathcal{S}, \min(s_x, s_y), \max(t_x, t_y))$ である．

したがって，子エピソードの極小出現リストの計算時間は，入力列の長さ n に対して，$O(n)$ である．また，WinCount [15] と同様に，極小出現リストを用いた頻度の計算時間も $O(n)$ である．

定理 1 ウィンドウ幅 w と最小頻度 σ が与えられたとき，アルゴリズム FreqPartite は，入力列 $\mathcal{S}$ からすべての頻出多部エピソードを，出力 1 つあたり $O(|\Sigma|^3 n)$ 時間と，$O(|\Sigma|^2 n)$ 領域で重複なく抽出する．ここに，Σ はアルファベット，および，n は $\mathcal{S}$ の長さである．

5. まとめ

本論文では，頻出多部エピソード問題に対して，多部エピソードの極小出現を利用することで，入力列から，すべての頻出多部エピソードを効率よく抽出するアルゴリズム FreqPartite を設計した．極小出現を利用することで，より一般的なエピソードのマイニングアルゴリズムを設計することは今後の課題である．また，アルゴリズム FreqPartite を細菌検査データ [9] に適用して，医学的に有用なエピソードを抽出することは今後の課題である．

参考文献

[1] R. Agrawal, R. Srikant: Fast algorithms for mining association rules in large databases, *Proc. 20th VLDB*, 487–499, 1994.

[2] T. Katoh, K. Hirata: Mining frequent elliptic episodes from event sequences, *Proc. 5th LLLL*, 46–52, 2007.

[3] T. Katoh, K. Hirata: A simple characterization on serially constructible episodes, *Proc. 12th PAKDD*, LNAI 5012, 600-607, 2008.

[4] T. Katoh, H. Arimura, K. Hirata: A polynomial-delay polynomial-space algorithm for extracting frequent diamond episodes from event sequences *Proc. 13th PAKDD*, LNAI 5476, 172-183, 2009.

[5] T. Katoh, H. Arimura, K. Hirata: Mining frequent bipartite episodes from event sequences *Proc. 12th DS*, LNAI 5808, 136-151, 2009.

[6] T. Katoh, H. Arimura, K. Hirata: Mining frequent k-partite episodes from event sequences *Proc. 6th LLLL*, 43-50, 2009.

[7] T. Katoh, K. Hirata, M. Harao: Mining sectorial episodes from event sequences, *Proc. 10th DS*, LNAI 4265, 137–145, 2006.

[8] T. Katoh, K. Hirata, M. Harao: Mining frequent diamond episodes from event sequences, *Proc. 4th MDAI*, LNAI 4617, 477–488, 2007.

[9] T. Katoh, K. Hirata, M. Harao, S. Yokoyama, K. Matsuoka: Extraction of sectorial episodes representing changes for drug resistant and replacements of bacteria, *Proc. CME'07*, 304–309, 2007.

[10] H. Mannila, H. Toivonen, A. I. Verkamo, Discovery of frequent episodes in event sequences, *Data Mining and Knowledge Discovery* **1**, 259–289, 1997.

[11] J. Pei, H. Wang, J. Liu, K. Wang, J. Wang, P. S.. Yu, Discovering Frequent Closed Partial Orders from Strings, *IEEE TKDE*, 18(11), 1467–1481, 2006.

[12] J. Pei, J. Han, B. Mortazavi-Asi, J. Wang, H. Pinto, Q. Chen, U. Dayal, M.-C. Hsu, Mining sequential patterns by pattern-growth: The PrefixSpan approach, *IEEE Trans. Knowledge and Data Engineering* **16**, 1–17, 2004.

[13] T. Uno, Two general methods to reduce delay and change of enumeration algorithms, *NII Technical Report*, NII-2003-004E, April 2003.

[14] M. J. Zaki, C.-J. Hsiao: CHARM: An efficient algorithm for closed itemset mining, *Proc. 2nd SDM*, 457–473, 2002.

[15] H. Ohtani, T. Kida, T. Uno, H. Arimura: *Efficient serial episode mining with minimal occurrences*, *Proc. 3rd ICUIMC*, 457–464, 2009.