DEIM2022 D21-4

行動時間帯に偏りのある長時間エピソード抽出における探索候補枝刈り手法

安井　壱陽† 　新谷隆彦 †† 　大森　匡 †† 　藤田秀之 ††

† †† 電気通信大学大学院情報理工学研究科　〒 182–8585 東京都調布市調布ヶ丘 1 丁目 5–1
E-mail: †y2031145@edu.cc.uec.ac.jp, ††{shintani,omori,fujita}@is.uec.ac.jp

あらまし　我々は, 長期間の生活に関するライフログとして, いつからいつまでどのような状態であったかを表す運動状態データをリストバンド型加速度センサを用いて収集し, そのデータから長時間を占めた行動に相当する長時間エピソードを抽出する研究を進めている. 報告者は, 行動した時間帯の偏りに着目した長時間エピソードを検討してきた. 行動時間帯が偏った長時間エピソードの抽出は, 従来の長時間エピソードマイニングより探索候補が多くなるため, 継続時間のカウント処理負荷が高くなる. そこで本研究では, 成長させたエピソードとエピソードを構成するイベントタイプのとり得る時間帯継続時間を考慮して探索候補を枝刈りする手法を提案し, 実データを用いて評価実験を行った.

キーワード　エピソードマイニング, 行動時間帯, 探索候補枝刈り, 時系列データ, ライフログ

1　背景と目的

近年のスマートフォンなどの小型端末の普及により, 人の生活に関するデータであるライフログを収集することが容易になった. ライフログの活用に関する研究 [1] として, スマートフォンの操作履歴などのライフログから日常的な利用行動を抽出するモバイルマイニング [2] などの研究が進められている. また, イベントシーケンスデータに対するデータマイニング技術として, エピソードマイニング [3] がある. エピソードマイニングでは, シーケンスデータによく現れたイベントの順序パターンである頻出エピソードを抽出する. これをライフログに適用した場合, 頻繁に行われた行動に相当するエピソードを抽出することができる. しかし, 人の行動の中には頻度は少ないが長い時間費やした行動が存在する. そこで我々は, ユーティリティマイニング [4] の一つとして長時間エピソードマイニング [5] の研究を進めてきた. 長時間エピソードマイニングは継続時間をエピソードの評価値として考慮することで, 長い時間費やした行動に相当する長時間エピソードを抽出する.

人の行動には特定の時間によく行う行動がある. しかし, エピソードマイニングと長時間エピソードマイニングは全日における頻度や継続時間によってエピソードを評価するため, 特定の時間によく行う行動は考慮されてこなかった. そこで報告者は, 特定の時間によく行われた行動に相当するエピソードとして, 行動時間帯に偏りのある長時間エピソードを検討してきた [6]. 行動時間帯の偏りとは, エピソードが 2 つの時刻から表される区間にどのくらい偏って発生したかを表す指標であり, これが高いほど所定の行動時間帯によく発生したエピソードであることを表している.

行動時間帯に偏りのある長時間エピソードを抽出するためには, 所定の行動時間帯における継続時間を考慮する必要がある. 1 日の中で限定された時間帯内にどのくらい行われていたかでエピソードを評価するため, 抽出の条件として従来の長時間エピソードマイニングより閾値を小さくする必要がある. したがって, 探索候補が多くなるため頻度や継続時間のカウント処理の負荷が高くなる.

本研究では, 行動時間帯に偏りのある長時間エピソード抽出において探索候補数を削減する手法を提案する. 行動時間帯において, 成長させたエピソードがとり得る継続時間の最大値を求め, 探索候補を枝刈りする. さらに, エピソードを構成するイベントタイプについても, 行動時間帯においてとり得る継続時間の最大値を求め, 冗長なイベントタイプを枝刈りする. 報告者が収集したライフログを用いて評価実験を行い, 提案手法の探索候補数の削減により処理負荷を低減できることを示す.

2　長時間エピソード

2.1　定　　義

長時間エピソードの定義 [5] を示す. 処理対象データはシーケンスデータ $S = \langle e_1, e_2, ..., e_n \rangle$ であり, イベント e を開始日時順に並べたリストである. イベント $e = (\epsilon, t_s, t_e)$ はイベントタイプ $\epsilon \in E$, 開始日時 t_s, 終了日時 t_e の組からなり, ϵ が t_s から t_e まで行われたことを表す. E はイベントタイプの集合を表す. ここで, シーケンスデータ上で複数のイベントが同時に発生することはない. つまり, シーケンスデータ上のイベント (ϵ, t_s, t_e) に対し, $t_s < t'_e$ かつ $t_e > t'_s$ となるイベント (ϵ', t'_s, t'_e) は存在しない. イベントの例を表 1 に示す. 各行がイベントを表している. これらのイベントを開始日時の順に並べたリストが 2021 年 5 月 20 日 18 時から 2021 年 5 月 21 日 0 時 58 分までのシーケンスデータである. 例えば 1 行目のイベントは, イベントタイプ ϵ_2 が 2021 年 5 月 20 日 18 時から同日 18 時 54 分まで 54 分間継続して行われたことを表す.

エピソード $\alpha = \langle a_1\text{-}a_2\text{-}\cdots\text{-}a_k \rangle$ はイベントタイプのリストである. a_i $(1 \leqq i \leqq k)$ は $a_i \in E$ を満たす. エピ

表 1: イベントの例

イベントタイプ	開始時刻	終了時刻
ϵ_2	2021-5-20 18:00	2021-5-20 18:54
ϵ_1	2021-5-20 18:54	2021-5-20 19:01
ϵ_4	2021-5-20 19:01	2021-5-20 19:25
ϵ_2	2021-5-20 19:25	2021-5-20 19:39
ϵ_1	2021-5-20 19:39	2021-5-20 20:53
ϵ_2	2021-5-20 20:53	2021-5-20 21:48
ϵ_4	2021-5-20 21:48	2021-5-20 22:01
ϵ_1	2021-5-20 22:01	2021-5-20 23:21
ϵ_3	2021-5-20 23:21	2021-5-21 00:01
ϵ_4	2021-5-21 00:01	2021-5-21 00:58

ソード α を構成する各イベントタイプを含むイベントがシーケンスデータ S に同一の順序で現れたとき, エピソードがシーケンスデータ上に発生したと表現する. つまり, $S = \langle(\epsilon_1, t_{s_1}, t_{e_1}), (\epsilon_2, t_{s_2}, t_{e_2}), \cdots, (\epsilon_n, t_{s_n}, t_{e_n})\rangle$ において, α の各イベントタイプ $a_i (1 \leqq i \leqq k)$ について $\epsilon_{j_i} = a_i (1 \leqq j_1 < j_2 < \cdots j_k \leqq n, \quad 1 \leqq i \leqq k)$ となる部分シーケンス $\langle(\epsilon_{j_1}, t_{s_{j_1}}, t_{e_{j_1}}), (\epsilon_{j_2}, t_{s_{j_2}}, t_{e_{j_2}}), \cdots, (\epsilon_{j_k}, t_{s_{j_k}}, t_{e_{j_k}})\rangle$ が存在するとき, α が S に発生したとする. α が発生した区間は, 部分シーケンスの先頭のイベント $(\epsilon_{j_1}, t_{s_{j_1}}, t_{e_{j_1}})$ の開始日時から末尾のイベント $(\epsilon_{j_k}, t_{s_{j_k}}, t_{e_{j_k}})$ の終了日時までである. この区間をオカレンスと呼び, $[t_{s_{j_1}}, t_{e_{j_k}})$ と表す. 継続時間 d_t は $t_{e_{j_k}} - t_{s_{j_1}}$ である. 冗長なオカレンスを除外する制約として, オカレンスは最大オカレンス継続時間 $maxspan$ と最大ギャップ $maxgap$ を持つ. 最大オカレンス継続時間はオカレンスが取ることのできる継続時間の最大値を表し, $d_t > maxspan$ となる継続時間が極端に長いオカレンスが除外される. 最大ギャップは, エピソードの 2 つの連続するイベントタイプに対するイベントの時間間隔の最大値を表す. α における 2 つの連続するイベントタイプ a_i, a_{i+1} のイベント (a_i, t_s, t_e), (a_{i+1}, t'_s, t'_e) の時間間隔 $t'_s - t_e$ が $t'_s - t_e > maxgap$ となる場合, 区間 $[t_s, t'_e)$ を含むオカレンスは α のオカレンスから除外される.

エピソード α は頻度 $freq$ と総継続時間 $tdur$ を評価値として持つ. α の極小非重複オカレンス [7] の数が頻度であり, α の極小非重複オカレンスの継続時間の総和が総継続時間である. 極小オカレンスとは, 開始日時から終了日時の間に他のオカレンスを含まないオカレンスであり, エピソード α の極小オカレンスの集合を $MO(\alpha)$ と表す. つまり, α のオカレンス $occ = [t_s, t_e)$ に対して $t_s \leqq t'_s$ かつ $t_e \geqq t'_e$ を満たす α のオカレンス $[t'_s, t'_e)$ が存在しない時 occ は極小オカレンスとなる. 非重複オカレンスは, オカレンスが互いに重複していないことを表す. つまり, α のオカレンス $occ = [t_s, t_e)$ に対して $t_s < t'_e$ かつ $t_e > t'_s$ を満たす α のオカレンス $[t'_s, t'_e)$ が存在しない時 occ は非重複オカレンスとなる. 極小と非重複を同時に満たすオカレンスを極小非重複オカレンスと呼び, エピソード α の極小非重複オカレンスの集合を $MANO(\alpha)$ と表す. 例えば表 1 のシーケンスデータにおいて, エピソード $\alpha = \langle\epsilon_2\text{-}\epsilon_1\text{-}\epsilon_4\rangle$ の極小非重複オカレンスの集合は

$$MANO(\alpha) = \{[\text{2021-05-20 18:00}, \text{2021-05-20 19:25}), [\text{2021-05-20 19:25}, \text{2021-05-20 22:01})\}$$

である. また, 総継続時間は $85 + 156 = 241$ 分となる.

長時間エピソード抽出問題は, シーケンスデータから総継続時間が最小総継続時間 $mintdur$ を満たすエピソードをすべて抽出することである.

2.2 抽出アルゴリズム

長時間エピソードの抽出手順を説明する. エピソード α の末尾に 1 つのイベントタイプ e を接続して成長させたエピソード β を作成し, $MO(\beta)$ を $MO(\alpha)$ と $MO(e)$ から作成した後に, $MO(\beta)$ から $MANO(\beta)$ を作成し, β が最小総継続時間を満たす場合に長時間エピソードとして抽出する. これを再帰的に繰り返し, 長時間エピソードを深さ優先探索する. 長時間エピソード抽出アルゴリズムを Algorithm1 に示す.

Algorithm 1: ExtractLongDurationEpisodes

```
1  S をスキャンし, すべてのイベントタイプの MO を生成
2  E := lowfreq を満たすすべてのイベントタイプ
3  foreach e ∈ E do
4      tdur(e) = Σ_{[t_s,t_e)∈MO(e)} (t_e − t_s)
5      if tdur(e) ≧ mintdur then
6          output e
7      end
8      foreach e' ∈ E do
9          Extend(e, e')
10     end
11 end
```

長時間エピソードの抽出においてすべての探索候補を調べると, 探索候補数が膨大になり処理負荷が高くなる. 探索候補数を削減するために, 従来のエピソードマイニングでは Apriori の性質 [8] を用いて探索候補を枝刈りする. しかし, 長時間エピソードは頻度ではなく総継続時間で評価するが, 総継続時間は Apriori の性質を満たさないため, 総継続時間では従来の探索候補の枝刈りはできない. そこで長時間エピソードの抽出では, 最小総継続時間を満たすために最も小さい頻度によって探索候補を枝刈りする. エピソードのすべてのオカレンスの継続時間が $maxspan$ のときに最小総継続時間を満たす場合に頻度が最も小さくなる. この場合の頻度は $mintdur/maxspan$ であり, これを下限頻度 $lowfreq$ と呼ぶ. 長時間エピソードの抽出では, 下限頻度を閾値として探索候補を枝刈りする.

まず 1 行目でシーケンスデータをスキャンし, すべてのイベントタイプの極小オカレンス集合 MO を生成した後に, 2 行目で $lowfreq$ を満たすイベントタイプの集合 E を生成する. 4 行目から 10 行目までは, イベントタイプ e と e を成長させたエピソードを探索する処理であり, E のすべてイベントタイプに対して行う. 4 行目から 7 行目で e の $tdur$ を計算し, $mintdur$ を満たす場合に e を長時間エピソードとして出力する. そして,

E の各イベントタイプ e' で e を成長させたエピソードを探索するために, 9 行目で関数 Extend(e, e') を呼び出す.

関数 Extend は, エピソード α の末尾にイベントタイプ e を繋いで成長させたエピソード β が長時間エピソードであるかどうかを調べる関数である. 関数 Extend のアルゴリズムを Algorithm2 に示す.

Algorithm 2: Extend(α, e)

```
1  β := α を e で成長させたエピソード
2  MO(β) := MO(α) と MO(e) から作成した極小オカレンス
3  MANO(β) := MO(β) から作成した非重複オカレンス
4  tdur(β) = Σ_{[t_s,t_e)∈MANO(β)} (t_e − t_s)
5  if tdur(β) ≧ mintdur then
6      output β
7  end
8  if freq(β) ≧ lowfreq then
9      foreach e' ∈ E do
10         Extend(β, e')
11     end
12 end
```

まず, 1 行目で α の末尾に e を繋いで成長させたエピソード β を作成する. 次に, 2 行目で $MO(\alpha)$ と $MO(e)$ から β の極小オカレンス集合 $MO(\beta)$ を作成する. β の極小オカレンスは, $MO(\alpha)$ のオカレンスに $MO(e)$ のオカレンスをつないで作成したオカレンスのうち, $maxspan$ と $maxgap$ を共に満たす極小オカレンスである. 3 行目では, $MO(\beta)$ から β の極小非重複オカレンス集合 $MANO(\beta)$ を作成する. $MANO(\beta)$ は, $MO(\beta)$ に含まれる極小オカレンスのうち, 非重複となるオカレンスである. そして, 4 行目から 7 行目で β の $tdur$ を求め, $mintdur$ を満たす場合に β を長時間エピソードとして出力する. 最後に 8 行目から 12 行目で, β の $freq$ が $lowfreq$ を満たす場合, E のすべてのイベントタイプ e' で β を成長させたエピソードの探索を行うために Extend(β, e') を呼び出す.

3 行動時間帯に偏りのある長時間エピソード

3.1 定　　義

行動時間帯に偏りのある長時間エピソードの定義 [9] を示す. 時刻 r_s から r_e までを行動時間帯 T_r とし, $T_r = [r_s, r_e)$ と表す. ここで, r_s は時間帯開始時刻, r_e は時間帯終了時刻, 行動時間帯の幅は $r_e - r_s$ である. 本研究では, 行動時間帯におけるオカレンスの継続時間の計算を単純にするために, 行動時間帯の幅と最大オカレンス継続時間を共に 720 分未満とし, 1 つのオカレンスと行動時間帯の重なる区間が複数存在することを回避した.

エピソードのオカレンスが行動時間帯と重なるとき, 行動時間帯でエピソードが発生したとし, 行動時間帯と重なった区間の長さを時間帯継続時間 d_r とする. つまり, エピソード α のオカレンス $occ(\alpha) = [t_s, t_e)$ を時刻のみで表した区間 $[t'_s, t'_e)$ が行動時間帯 $T_r = [r_s, r_e)$ に対して $t'_s < r_e$ かつ $t'_e > r_s$ を満たすとき, T_r で α が発生したとする. 時間帯継続時間の計算は, オカレンスと行動時間帯の重なり方によって 5 通りに場合分けされる.

（1） $occ(\alpha)$ が T_r 内に発生, つまり $t'_s \geqq r_s$ かつ $t'_e \leqq r_e$ のとき, $occ(\alpha)$ の d_r は $t'_e - t'_s$ である.

（2） $occ(\alpha)$ が T_r 内で終了している, かつ, $occ(\alpha)$ が T_r より前に開始している, つまり $r_s < t'_e \leqq r_e$ かつ $t'_s < r_s$ のとき, $occ(\alpha)$ の d_r は $t'_e - r_s$ である.

（3） $occ(\alpha)$ が T_r 内で開始している, かつ, $occ(\alpha)$ が T_r より後に終了している, つまり $r_s \leqq t'_s < r_e$ かつ $t'_e > r_e$ のとき, $occ(\alpha)$ の d_r は $r_e - t'_s$ である.

（4） $occ(\alpha)$ が T_r をまたがる, つまり $t'_s < r_s$ かつ $t'_e > r_e$ のとき, $occ(\alpha)$ の d_r は $r_e - r_s$ である.

（5） $occ(\alpha)$ が T_r と重ならない, つまり $t'_s < r_e$ と $t'_e > r_s$ のいずれかのみを満たすとき, $occ(\alpha)$ の d_r は 0 である.

オカレンスの時間帯継続時間の例を示す. エピソード α のオカレンスが表 2 のとき, 各オカレンスの時間帯継続時間は図 1 の区間となる. ここで, 行動時間帯を $[08:00, 14:00)$ とする. 図 1 の横軸は時刻, 縦軸は日付を表している. 各日ともに上下に二つの区間があり, 下の色の薄い区間がオカレンス, 上の色の濃い区間が行動時間帯 $[08:00, 14:00)$ におけるオカレンスの時間帯継続時間となる区間を表しており, 色の濃い区間が存在していない場合はオカレンスの時間帯継続時間が 0 であることを表している. 例えば, 2021 年 1 月 4 日のオカレンスは行動時間帯内で開始している, かつ, 時間帯終了時刻より後に終了しているため, 場合 3 により時間帯継続時間となる区間は 12 時から 14 時, 時間帯継続時間は 120 分となる.

表 2: エピソード α のオカレンスの例

日付	オカレンス
2021-01-01	[2021-01-01 15:00, 2021-01-01 18:00)
2021-01-02	[2021-01-02 02:00, 2021-01-02 07:00)
2021-01-03	[2021-01-03 07:30, 2021-01-03 14:30)
2021-01-04	[2021-01-04 12:00, 2021-01-04 16:00)
2021-01-05	[2021-01-05 07:00, 2021-01-05 10:00)
2021-01-06	[2021-01-06 10:00, 2021-01-06 12:00)

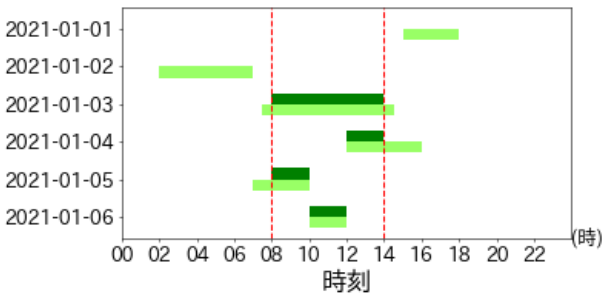

図 1: 時間帯継続時間の例

エピソード α は, 総時間帯継続時間 $trdur$ と時間帯偏り度 tb を評価値として持つ. α の極小非重複オカレンスの時間帯継続

時間の総和が総時間帯継続時間である. 時間帯偏り度は, α が所定の行動時間帯に偏っているかを表す評価値であり, 式 (1) で定義される.

$$tb(\alpha) = \frac{\sum_{occ \in MANO(\alpha)} d_r(occ)}{\sum_{occ' \in MANO(\alpha)} d_t(occ')} \quad (1)$$

$d_r(occ)$ は α の極小非重複オカレンス occ の時間帯継続時間, $d_t(occ')$ は α の極小非重複オカレンス occ' の継続時間を表す. 時間帯偏り度は $0 \leqq tb \leqq 1$ であり, 1 に近いほど行動時間帯に大きく偏ってエピソードが発生したことを示す. 例えば, 図 1 において α の総時間帯継続時間は 720 分, 総継続時間は 1440 分であるため, 時間帯偏り度は $720/1440 = 0.5$ となる.

行動時間帯に偏りのある長時間エピソード抽出問題は, シーケンスデータから, ユーザが指定した行動時間帯 $[r_s, r_e)$ における最小総時間帯継続時間 $mintrdur$ と最小時間帯偏り度 $mintb$ を満たす長時間エピソードをすべて抽出することである.

3.2 抽出手法

行動時間帯に偏りのある長時間エピソード抽出の基本手法を説明する. 基本手法は長時間エピソード抽出アルゴリズムをベースとし, 総継続時間に加え, 総時間帯継続時間と時間帯偏り度を求め, 最小総時間帯継続時間および最小時間帯偏り度を満たすエピソードを行動時間帯に偏りのある長時間エピソードとして抽出する. 行動時間帯に偏りのある長時間エピソード抽出アルゴリズムを Algorithm3 に示す.

Algorithm 3: ExtracttTBEpisodes

```
1  S をスキャンし, すべてのイベントタイプの MO を生成
2  E := lowfreq を満たすすべてのイベントタイプ
3  foreach e ∈ E do
4      trdur(e) := e の trdur
5      if trdur(e) ≧ mintrdur then
6          tdur(e) = Σ_{[t_s,t_e)∈MANO(e)} (t_e − t_s)
7          if trdur(e)/tdur(e) ≧ mintb then
8              output e
9          end
10     end
11     foreach e' ∈ E do
12         Extend'(e, e')
13     end
14 end
```

行動時間帯に偏りのある長時間エピソード抽出では, 下限頻度を $lowfreq = mintrdur/maxspan$ として探索候補を枝刈りする. まず 1 行目でシーケンスデータをスキャンし, すべてのイベントタイプの極小オカレンス集合 MO を生成した後に, 2 行目で $lowfreq$ を満たすイベントタイプの集合 E を生成する. 4 行目から 13 行目は, イベントタイプ e と e を成長させたエピソードを調べる処理であり, E のすべてのイベントタイプ e に対して行う. 4 行目で e の $trdur$ を計算し, $mintrdur$ を満たす場合に 6 行目で e の $tdur$, 7 行目で $tb(e)$ を計算し, $tb(e)$ が $mintb$ を満たす場合に 8 行目で e を行動時間帯に偏りのある長時間エピソードとして出力する. そして, E のすべてイベントタイプ e' で e を成長させたエピソードの探索を行うために, 13 行目で Extend$'(e, e')$ を呼び出す.

Extend$'$ は, エピソード α をイベントタイプ e で成長させたエピソード β が行動時間帯に偏りのある長時間エピソードであるかを調べる関数である. 関数 Extend$'$ のアルゴリズムを Algorithm4 に示す.

Algorithm 4: Extend$'(\alpha, e)$

```
1  β := α を e で成長させたエピソード
2  MO(β) := MO(α) と MO(e) から作成した極小オカレンス
3  MANO(β) := MO(β) から作成した極小非重複オカレンス
4  trdur(β) := e の trdur
5  if trdur(β) ≧ mintrdur then
6      tdur(β) = Σ_{[t_s,t_e)∈MANO(β)} (t_e − t_s)
7      if trdur(β)/tdur(β) ≧ mintb then
8          output β
9      end
10 end
11 if freq(β) ≧ lowfreq then
12     foreach e ∈ E do
13         Extend'(β, e)
14     end
15 end
```

1 行目で α の末尾に e を繋いで成長させたエピソード β の作成, 2 行目で $MO(\alpha)$ と $MO(e)$ から β の極小オカレンス集合 $MO(\beta)$ の作成, 3 行目で $MO(\beta)$ から非重複なオカレンスを取り出し β の極小非重複オカレンス集合 $MANO(\beta)$ の作成を行う. 4 行目で β の $trdur$ を計算し, $mintrdur$ を満たす場合に 6 行目で β の $tdur$, 7 行目で $tb(\beta)$ を計算し, $tb(\beta)$ が $mintb$ を満たす場合にを 8 行目で β を行動時間帯に偏りのある長時間エピソードとして出力する. 最後に 11 行目から 15 行目で, β の $freq$ が $lowfreq$ を満たす場合, E のすべてのイベントタイプ e' で β を成長させたエピソードの探索を行うために Extend$'(\beta, e')$ を呼び出す.

4 提案手法

4.1 とり得る総時間帯継続時間による枝刈り

行動時間帯に偏りのある長時間エピソードは行動時間帯における継続時間で評価するため, 最小総時間帯継続時間は行動時間帯の幅に応じて小さくなり, 下限頻度も小さくなる. よって, 基本手法では探索空間が広くなり, 探索候補数が増大する. そこで本研究では, 効率よく行動時間帯に偏りのある長時間エピソードを抽出するために, 探索候補のとり得る総時間帯継続時間に着目した探索候補の枝刈りと, イベントタイプがエピソードを構成する際にとり得る時間帯継続時間に着目したイベントタイプの枝刈りを提案する.

4.1.1 探索候補の枝刈り

オカレンスの継続時間の最大値は最大オカレンス継続時間であるが, 時間帯継続時間は行動時間帯と重なる区間を考慮するため, オカレンスがとり得る時間帯継続時間の最大値は最大オカレンス継続時間以下となる. 探索候補のオカレンスがとり得る時間帯継続時間を探索候補のとり得る時間帯継続時間と呼ぶ. 探索候補のとり得る総時間帯継続時間の最大値は, 探索候補のとり得る時間帯継続時間の最大値の総和であり, この値が最小総時間帯継続時間を満たさない場合, 探索候補の総時間帯継続時間が最小総時間帯継続時間を満たすことはないため, 探索候補を枝刈りできる [9]. そこで提案手法は, エピソードを成長させる際に下限頻度ではなく, とり得る総時間帯継続時間によって探索候補を枝刈りする.

また, 探索候補のとり得る総時間帯継続時間は Apriori の性質が成り立つ. エピソード α を成長させたエピソード β について, β のオカレンスは α のオカレンスから作成されるため α と β のとり得る時間帯継続時間の最大値は等しくなる. さらに, β のオカレンス数は α より大きくならないため, β のとり得る総時間帯継続時間の最大値は α のとり得る総時間帯継続時間の最大値より大きくならない. したがって, とり得る総時間帯継続時間が最小総時間帯継続時間を満たさない探索候補とその探索候補を成長させた探索候補の探索を省略できる.

エピソード α のオカレンスが $occ(\alpha) = [t_s, t_e)$ のとき, α を成長させた探索候補のとり得る時間帯継続時間の最大値をオカレンスの開始から $maxspan$ 後までの区間 $[t_s, t_s + maxspan)$ と行動時間帯 $T_r = [r_s, r_e)$ が重なる区間から求める. これは, 行動時間帯との重なり方によって 3 通りに場合分けされる. ここで, α のオカレンスの開始時刻 t'_s から $maxspan$ 後までの区間を $[t'_s, t'_s + maxspan)$ とする.

（1） $occ(\alpha)$ が T_r 内で開始している, つまり $r_s \leqq t'_s < r_e$ のとき, $occ(\alpha)$ のとり得る時間帯継続時間の最大値は $min\{r_e - t'_s, maxspan\}$ である.

（2） $occ(\alpha)$ が T_r 外で開始している, かつ, $occ(\alpha)$ の開始から $maxspan$ 後までの区間が T_r と重なる, つまり $t'_s < r_s$ かつ $t'_s + maxspan > r_s$ のとき, $occ(\alpha)$ のとり得る時間帯継続時間の最大値は $min\{(t'_s + maxspan) - r_s, r_e - r_s\}$ である.

（3） $occ(\alpha)$ の開始から $maxspan$ 後までの区間が T_r と重ならない, つまり $t'_s < r_e$ または $t'_s + maxspan > r_s$ のいずれかのみを満たす場合, $occ(\alpha)$ のとり得る時間帯継続時間の最大値は 0 である.

表 2 のオカレンスにおいて, エピソード α を成長させた探索候補 β がとり得る最大の区間と β のとり得る時間帯継続時間の最大値を図 2 に示す. ここで, 行動時間帯を $[08:00, 14:00)$, 最大オカレンス継続時間を 420 分とする. 図 2 の横軸は時刻, 縦軸は日付を表している. 各日ともに上下に二つの区間があり, 下の区間は α のオカレンス, 上の区間は β がとり得る最大の区間である. 上の区間において色が濃くなっている区間は, β のとり得る時間帯継続時間の最大の区間を表している. 例えば, 2021 年 1 月 4 日のオカレンスは行動時間帯内で開始しているため, 場合 1 によりとり得る時間帯継続時間が最大となる区間は 12 時から 14 時まで, とり得る時間帯継続時間の最大値は 120 分となる. また, β のとり得る総時間帯継続時間の最大値は図 2 の上の濃い色の区間の総和 1140 分であり, これが最小総時間帯継続時間を満たさない場合 β の探索を枝刈りする.

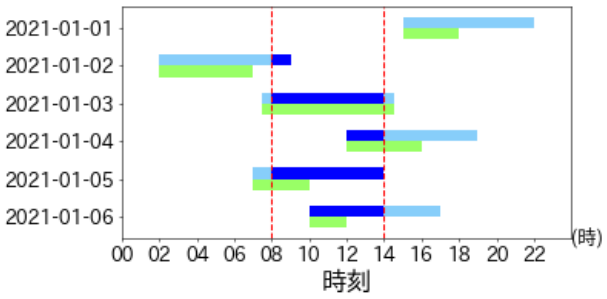

図 2: 探索候補のとり得る時間帯継続時間の最大値の例

4.1.2 イベントタイプの枝刈り

イベントタイプ e を含むエピソードがとり得る総時間帯継続時間の最大値が最小総時間帯継続時間を満たさないとき, e を含む探索候補を枝刈りできる. e を含むエピソードが発生し得る区間は, e がエピソードの先頭となり継続時間が $maxspan$ の区間から e がエピソードの末尾となり継続時間が $maxspan$ の区間までであるため, この区間と行動時間帯が重なる区間からイベントタイプのオカレンスがとり得る時間帯継続時間を求める. これをイベントタイプの時間帯継続時間と呼ぶ. イベントタイプ e のとり得る総時間帯継続時間の最大値は, イベントタイプのとり得る時間帯継続時間の最大値の総和であり, この値が最小総時間帯継続時間を満たさない場合に e を含むエピソードの総時間帯継続時間が最小総時間帯継続時間を満たすことはないため, e を枝刈りできる. そこで提案手法では, 下限頻度を満たすイベントタイプの集合ではなく, イベントタイプのとり得る総時間帯継続時間の最大値が最小総時間帯継続時間を満たすイベントタイプの集合を生成する.

イベントタイプ e を含むエピソードは, オカレンスが終了する $maxspan$ だけ前の日時から, オカレンスの開始から $maxspan$ 後までの区間に発生し得る. そこで, e を含むエピソードが発生し得る区間から e のとり得る時間帯継続時間の最大値を求める. オカレンスと行動時間帯が重なる区間は 1 つのみであるため, e のオカレンスが $occ(e) = [t_s, t_e)$ のとき, オカレンスが終了する $maxspan$ だけ前の日時からオカレンスの終了までの区間 $[t_e - maxspan, t_e)$ と, オカレンスの開始から $maxspan$ 後までの区間 $[t_s, t_s + maxspan)$ のうち, 行動時間帯 $T_r = [r_s, r_e)$ と長く重なる区間を e のとり得る時間帯継続時間の最大値とする. イベントタイプのとり得る時間帯継続時間の最大値は, 行動時間帯との重なり方によって 3 通りに場合分けされる. ここで, e のオカレンスを時刻で表した区間を $[t'_s, t'_e)$, e を含むエピソードが発生し得る区間を時刻で表した区間を $[t'_e - maxspan, t'_s + maxspan)$ とする.

（1） $occ(e)$ と T_r が重なる, つまり $t'_s < r_e$ かつ $t'_e > r_s$ のとき, $occ(e)$ のとり得る時間帯継続時間の最大値は $min\{r_e - r_s, maxspan, (t'_s + maxspan) - r_s, r_e - (t'_e - maxspan)\}$ で

ある．

（2） $occ(e)$ と T_r が重ならず，e を含むエピソードの発生し得る区間が T_r に重なる，つまり $t'_s < r_e$ と $t'_e > r_s$ のいずれかのみを満たす，かつ $t'_s + maxspan > r_s$ かつ $t'_e - maxspan < r_e$ のとき，$occ(e)$ のとり得る時間帯継続時間の最大値は $min\{r_e - r_s, max\{(t'_s + maxspan) - r_s, r'_e - (t'_e - maxspan)\}\}$ である．

（3） e を含むエピソードが発生し得る区間と T_r が重ならない，つまり $t'_s + maxspan > r_s$ または $t'_e - maxspan < r_e$ のいずれかのみを満たす場合，$occ(e)$ のとり得る時間帯継続時間の最大値は 0 である．

表 2 のオカレンスをイベントタイプ e のオカレンスとしたとき，e がとり得る時間帯継続時間の最大値を図 3 に示す．ここで，行動時間帯を $[08:00, 14:00)$，最大オカレンス継続時間を 420 分とする．図 3 の横軸は時刻，縦軸は日付を表している．各日ともに上下に二つの区間があり，下の区間は e のオカレンス，上の区間は e を含むエピソードが発生し得る区間を表している．上の区間において色の濃くなっている区間が，e のとり得る時間帯継続時間の最大の区間である．例えば，2021 年 1 月 4 日のオカレンスは行動時間帯と重なっているため，場合 1 によりとり得る時間帯継続時間の最大の区間は 9 時から 14 時まで，とり得る時間帯継続時間の最大値は 300 分となる．また，e のとり得る総時間帯継続時間は，図 3 の上の濃い色の区間の総和 1620 分である．これが最小総時間帯継続時間を満たさない場合，e を枝刈りする．

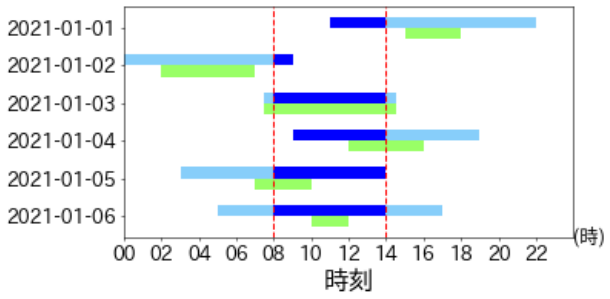

図 3: イベントタイプのとり得る時間帯継続時間の最大値の例

4.2 抽出アルゴリズム

提案する探索候補の枝刈りとイベントタイプの枝刈りを導入した行動時間帯に偏りのある長時間エピソード抽出手順を説明する．基本手法である Algorithm3 をベースとして，イベントタイプの集合の生成とエピソードを成長させる際の探索候補の枝刈りを提案手法に変更する．提案した枝刈りを導入した行動時間帯に偏りのある長時間エピソード抽出のアルゴリズムを Algorithm5 に示す．

基本手法から変更した点を説明する．Algorithm3 は 2 行目で $lowfreq$ を満たすイベントタイプの集合を生成していたが，Algorithm5 では 2 行目でイベントタイプのとり得る総時間帯継続時間の最大値が $mintrdur$ を満たすイベントタイプの集合

Algorithm 5: ExtractTBEpisodes'

```
1  S をスキャンし，すべてのイベントタイプの MO を生成
2  E := イベントタイプのとり得る総時間帯継続時間が mintrdur
     を満たすイベントタイプ
3  foreach e ∈ E do
4      if e を成長させた探索候補のとり得る総時間帯継続時間
           ≧ mintrdur then
5          trdur(e) := e の trdur
6          if trdur(e) ≧ mintrdur then
7              tdur(e) = Σ_{[ts,te)∈MO(e)} (te − ts)
8              if trdur(e)/tdur(e) ≧ mintb then
9                  output e
10             end
11         end
12         foreach e' ∈ E do
13             Extend''(e, e')
14         end
15     end
16 end
```

Algorithm 6: Extend''(α, e)

```
1  β := α を e で成長させたエピソード
2  MO(β) := MO(α) と MO(e) から作成した極小オカレンス
3  if βを成長させた探索候補のとり得る総時間帯継続時間
       ≧ mintrdur then
4      MANO(β) := MO(β) から作成した極小非重複オカレンス
5      trdur(β) := β の trdur
6      if trdur(β) ≧ mintrdur then
7          tdur(β) = Σ_{[ts,te)∈MANO(β)} (te − ts)
8          if trdur(β)/tdur(β) ≧ mintb then
9              output β
10         end
11     end
12     foreach e' ∈ E do
13         Extend''(β, e')
14     end
15 end
```

を生成する．また，Algorithm5 では，Algorithm3 の 4 行目から 13 行目までの各イベントタイプと各イベントタイプを成長させた探索候補を調べる処理を行う前に，4 行目で探索候補を枝刈りするためにイベントタイプを成長させた探索候補のとり得る総時間帯継続時間の最大値が $mintrdur$ を満たすか調べる．さらに，13 行目で探索候補の枝刈りを導入した関数 Extend'' を呼び出し，イベントタイプを成長させたエピソードの探索を行う．

探索候補の枝刈りを導入した関数 Extend'' を Algorithm6 に示す．基本手法の下限頻度による枝刈りは極小非重複オカレンスから $freq$ を調べる必要があるが，探索候補のとり得る総時間帯継続時間の最大値は極小オカレンスから求めることができる．そのため，Extend'' では 3 行目で極小オカレンスを作成した直後に，β から作成される探索候補のとり得る総時間帯継続時間

の最大値が最小総時間帯継続時間を満たすか調べ, 満たさない場合に探索候補を枝刈りする. また, 13 行目で関数 Extend″ を呼び出し, β を成長させた探索候補の探索を行う.

5 評価実験

提案手法が基本手法より処理負荷を低減できていることを評価するために実験を行った. 提案手法と基本手法の探索候補数と処理時間を調べ, 提案手法が基本手法に対してどのくらい減少できたかを評価した. また, 本実験において最小時間帯偏り度を 0.30, 最大オカレンス継続時間を 360 分, 最大ギャップを 60 分とした. 行動時間帯は, 幅が 6 時間となる $[08:00, 14:00), [16:00, 22:00)$ と, 幅が 4 時間となる $[08:00, 12:00), [16:00, 20:00)$ の 4 通りを用いた. 最小総時間帯継続時間は, 行動時間帯の幅が 6 時間の場合に 1 週間当たり 120 分, 4 時間の場合に 1 週間当たり 80 分とした.

本実験では処理対象データとして, 報告者が収集した 2019 年 5 月 1 日から 2021 年 10 月 1 日までの 885 日間の運動状態データを用いた. 運動状態データとは, リストバンド型センサで取得した, いつからいつまでどの程度の運動状態が継続したかを表すデータである. 運動状態には, ”静止”, ”安静”, ”デスクワーク”, ”軽作業”, ”作業”, ”歩行”, ”ジョギング”, ”運動”がある. さらに, 頻度の多い運動状態である”静止”, ”デスクワーク”, ”軽作業”, ”歩行”は, 運動状態データの継続時間の長さで短い運動状態と長い運動状態に分けられている. 運動状態データは長時間エピソードの抽出におけるイベント, 運動状態はイベントタイプに相当するため, 運動状態データを開始時刻順に並べたリストはシーケンスデータとなる.

まず, 探索候補数を比較した. データサイズを変化させた場合の提案手法と基本手法の探索候補数を図 4 に示す. 横軸はデータサイズである 2019 年 5 月 1 日からの月数, 縦軸は探索候補数を表している. さらに, 提案手法の探索候補数の基本手法に対する減少率を図 5 に示す. 横軸はデータサイズである 2019 年 5 月 1 日からの月数, 縦軸は減少率を表している. 減少率は式 (2) で求めた.

$$\text{減少率} = \frac{\text{基本手法の探索候補数} - \text{提案手法の探索候補数}}{\text{基本手法の探索候補数}} \tag{2}$$

提案手法の探索候補数が基本手法より少ないとき, 減少率は正

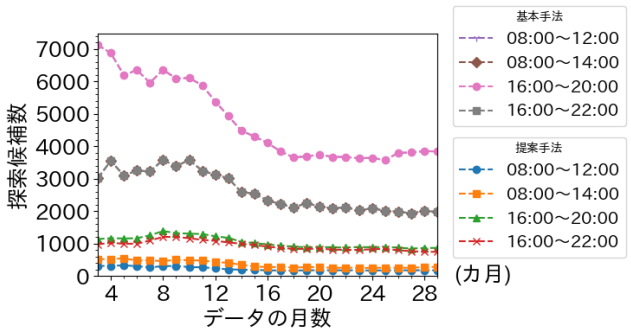

図 4: 提案手法と基本手法の探索候補数

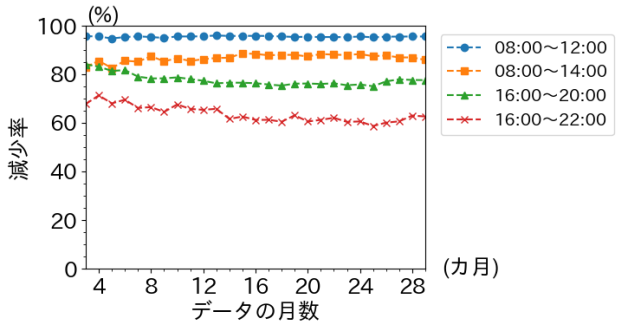

図 5: 提案手法の探索候補数の基本手法に対する減少率

の値をとる. 値が大きいほど大きく減少したことを表す.

図 4 と図 5 より, 提案手法は基本手法よりも大幅に探索候補数を削減できたことが確認できる. 基本手法は行動時間帯の幅が等しいとき探索候補数に差がないが, 提案手法は行動時間帯によって探索候補数に差が見られる. 基本手法は最小総時間帯継続時間と最大オカレンス継続時間から計算した下限頻度を用いて枝刈りするため, 行動時間帯の幅が等しいとき探索候補数が等しくなる. しかし, 提案手法はとり得る総時間帯継続時間の最大値を利用して枝刈りするため, 異なる行動時間帯では探索候補数が等しくならず, 探索候補数に差が生じると考えられる.

次に, 提案手法と基本手法の処理時間を調べた. データサイズを変化させた場合の提案手法と基本手法の処理時間を図 6 に示す. 横軸はデータサイズである 2019 年 5 月 1 日からの月数, 縦軸は処理時間を表している. 処理時間は 100 回計測した平均値とした. さらに, 提案手法の処理時間の基本手法に対する減少率を図 7 に示す. 横軸はデータサイズである 2019 年 5 月 1 日からの月数, 縦軸は減少率を表している. 減少率は式 (2) と同様に求めた.

図 6 と図 7 より, 提案手法は基本手法より処理時間を短くできたことが確認できる. しかし, 探索候補数の減少幅に比べ処理時間の減少幅は小さくなっている. 原因として, 行動時間帯に偏りのある長時間エピソードの探索の処理時間のみが減少したことが考えられる. 探索処理は全体に対して 45%から 76%の処理時間を占めており, 提案手法では探索処理に要する時間のみ短くできるため, 全体の処理時間の減少率は小さくなった.

最後に, 探索候補を調べる処理として極小オカレンスを作成

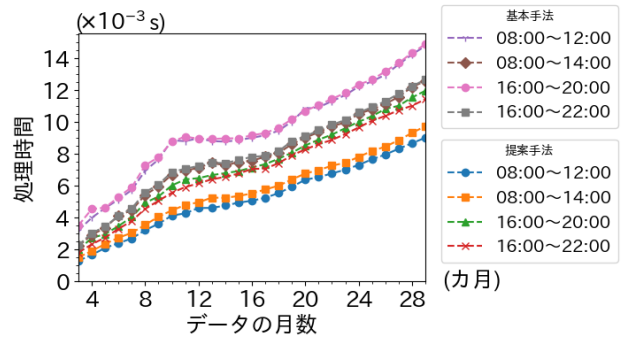

図 6: 提案手法と基本手法の処理時間

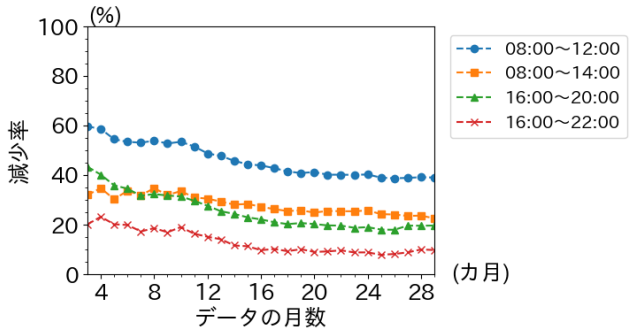

図 7: 提案手法の処理時間の基本手法に対する減少率

する処理の負荷についても調べた．データサイズを変化させた場合の提案手法と基本手法の処理全体のオカレンス比較回数を図 8 に示す．横軸はデータサイズである 2019 年 5 月 1 日からの月数, 縦軸はオカレンス比較回数を表している．オカレンス比較回数は, エピソード α の極小オカレンスとイベントタイプ e の極小オカレンスを接続してエピソード β のオカレンスを作成した回数である．この回数が多いほど処理負荷が高いことを表す．さらに, 提案手法のオカレンス比較回数の基本手法に対する減少率を図 9 に示す．横軸はデータサイズである 2019 年 5 月 1 日からの月数, 縦軸は減少率を表している．減少率は式 (2) と同様に求めた．

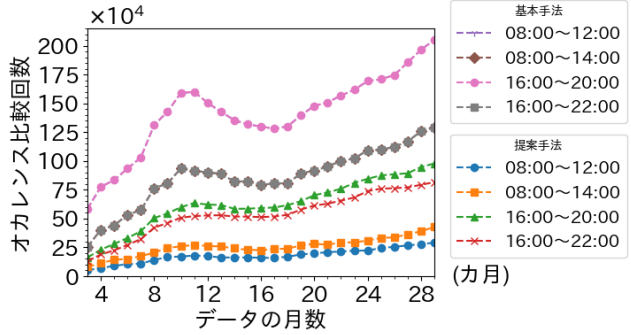

図 8: 提案手法と基本手法のオカレンス比較回数

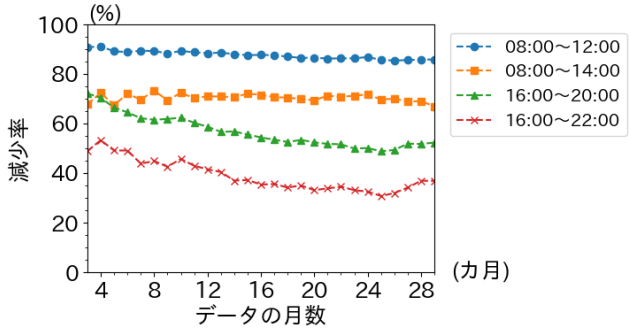

図 9: 提案手法のオカレンス比較回数の基本手法に対する減少率

探索候補数の減少率は図 5 より 58%から 97%であり, いずれのデータの月数, 行動時間帯においてもオカレンス比較回数の減少率は探索候補数の減少率よりも小さくなっていることがわかる．さらに, 提案手法で枝刈りできた探索候補と枝刈りできなかった探索候補のオカレンス比較回数を比較すると, 枝刈りできなかった探索候補のほうが少ない場合が多かった．例えばデータの月数が 24 カ月のときを比較すると, 提案手法で枝刈りできた探索候補 1 つあたりのオカレンス比較回数の平均は 901 回から 1377 回であったが, 枝刈りできなかった探索候補は 283 回から 424 回であった．行動時間帯に偏りのある長時間エピソードとならないエピソードはオカレンス数が少ない場合が多い．オカレンス数が少ないときオカレンス比較回数も少なくなるため, 提案手法は調べる必要のない探索候補を枝刈りできたと考えられる．

6 終わりに

本研究では, 行動時間帯に偏りのある長時間エピソードの抽出における処理負荷が高くなる問題点の解決を行った．成長させたエピソードとエピソードを構成するイベントタイプのオカレンスのとり得る時間帯継続時間に着目した探索候補とイベントタイプの枝刈り手法を提案した．報告者のデータを用いた評価実験によって, 提案手法の有効性を示した．今後は, 複数の行動時間帯に偏りのある長時間エピソードなどの検討を行う．

謝　辞

本研究は JST, CREST の支援を受けたものである．

文　献

[1] A.Ksibi, A.S.D.Alluhaidan, A.Salhi, S.A.El-Rahman, *Overview of Lifelogging: Current Challenges and Advances*, IEEE Access, vol.9, pp. 62630–62641, 2021.

[2] R. Rawassizadeh, E. Momeni, C. Dobbins, J. Gharibshah, M. Pazzani, *Scalable Daily Human Behavioral Pattern Mining from Multivariate Temporal Data*, IEEE Transactions on Knowledge and Data Engineering , vol.28, pp. 3098-3012, 2016.

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

[4] C.W.Wu, Y.F.Lin, P.S.Yu, and V.S.Tseng, *Mining high utility episodes in complex event sequences*, International Conference on Knowledge Discovery in Data, 2013.

[5] T. Shintani, T. Ohmori, H. Fujita, *Method for Comparing Long-term Daily life using Long-duration episodes*, EDBT/ICDT Workshops, International Workshop on Data Analytics solutions for Real-LIfe APplications, vol.2322-darliap8, 2019.

[6] 安井壱陽, 新谷隆彦, 大森匡, 藤田秀之, ”継続時間を考慮したエピソードマイニングにおける行動時間帯の偏りに関する一考察”, 情報処理学会第 82 回全国大会, 4N-04, 2020

[7] H. Zhu, P. Wang, X. He, Y. Li, W. Wang, B. Shi, *Efficient Episode Mining with Minimal and Non-overlapping Occurrences*, IEEE International Conference on Data Mining, pp. 1211-1216, 2010.

[8] R. Agrawel, R. Srikant, *Fast Algorithms for Mining Association Rules*, International Conference on Very Large Data Bases,1994.

[9] 安井壱陽, 新谷隆彦, 大森匡, 藤田秀之, ”行動時間帯に偏りのある長時間エピソード抽出における発生区間の範囲による探索候補の枝刈りの提案”, 情報科学技術フォーラム, D-005, 2021