

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

安井 壱陽[†] 新谷 隆彦[‡] 大森 匡[‡] 藤田 秀之[‡]
電気通信大学情報理工学域[†] 電気通信大学情報理工学研究科[‡]

1 はじめに

近年、スマートフォンなどの小型端末の普及により、日常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。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に行われた利用行動を抽出するモバイルマイニング [1] や、加速度センサで取得した腕の動きに関するライフログからエピソードマイニングを用いて日常行動に相当するエピソードを抽出し、生活特性を検出する研究 [2] が行われている。従来のエピソードマイニングでは行った回数が多い行動、行った時間の総和が長い行動に相当する長時間エピソードを抽出している。しかし、人の生活においては行動した時間帯も重要であると考えられる。特定の時間帯に偏って行われた行動は、人の生活特性を示すことが期待できる。

本研究では、時間帯を考慮したエピソードを検討する。長時間エピソードにおけるエピソードが発生した時間帯の偏りの評価を検討する。報告者が収集したライフログから時間帯の偏りを考慮したエピソードを抽出し、報告者が実際に行った生活行動との関連を考察する。

2 長時間エピソードマイニング

従来の長時間エピソード [2] における定義を示す。処理対象であるシーケンスデータ $S = \langle e_1, e_2, \dots, e_n \rangle$ は、イベントを開始日時[†]の順に並べたリストである。ここで、イベント $e = (\varepsilon, t_s, t_e)$ はあるイベントタイプ $\varepsilon \in E$ (E はイベントタイプの種類の集合) が、開始日時 t_s から終了日時 t_e まで行われたことを表す。

エピソード $\alpha = a_1 - a_2 - \dots - a_k$ はイベントタイプのリストである。ここで、 $a_i \in E$ ($1 \leq i \leq k$) を満たす。エピソード 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時、エピソードがシーケンスデータに発生したと表現する。長さ k のエピソード $\alpha = a_1 - a_2 - \dots - a_k$ は、シーケンスデータ $S = \langle (\varepsilon_1, t_{s_1}, t_{e_1}), (\varepsilon_2, t_{s_2}, t_{e_2}), \dots, (\varepsilon_n, t_{s_n}, t_{e_n}) \rangle$ に対して、 $a_i = \varepsilon_{j_i}$ ($1 \leq j_1 < j_2 < \dots < j_k \leq n, 1 \leq i \leq k$) を満たすとき、エピソード α がシーケンスデータ S に発生したとする。オカレンスはシーケンス S における α の先頭のイベントタイプ a_1 を含むイベントから末尾のイベントタイプ a_k を含むイベントまでのイベントのリストであり、 $[t_{s_{j_1}}, t_{e_{j_k}}]$

と表す。オカレンスの継続時間は $t_d = t_{e_{j_k}} - t_{s_{j_1}}$ である。オカレンスには制約条件として、最大継続時間と最大ギャップを用いる。最大継続時間により継続時間が長すぎるオカレンスを除外する。最大ギャップによりイベント間の時間間隔が長い場合エピソードと認めない。

エピソード α は頻度と総継続時間を評価値として持つ。 α の極小非重複オカレンス [3] の数が頻度であり、継続時間の和が総継続時間である。極小オカレンスとはオカレンスの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のことである。オカレンスが非重複であるということは、2つのオカレンスが互いに重複していないことを表す。エピソード α の極小かつ非重複なオカレンスの集合を極小非重複オカレンス $MANO(\alpha)$ と呼ぶ。

長時間エピソードマイニング問題は、総継続時間が最小総継続時間 $mindur$ 以上のエピソードをすべて抽出することである。

3 時間帯を考慮したエピソード

人は生活の中で様々な時刻に行動を行う。特定の時間帯に多く行われた行動は、生活特性を示すと考えられる。そこで、エピソードのオカレンスが発生した時間帯を考慮し、所定の時間帯においてどの程度発生したかを評価する。エピソードが総継続時間に対して特定の時間帯で長く発生している時、そのエピソードは時間帯に偏って発生している。エピソードが特定の時間帯に長く発生したとは、エピソードがその時間帯に偏って発生したことを意味する。オカレンスはエピソードが発生した区間であるため、所定の時間帯におけるオカレンスの継続時間を考えることでエピソードの発生した時間帯の偏りを評価できると考えた。

本研究では、ユーザが指定した時間帯に偏って発生したエピソードを考える。時間帯 TR は1分単位の開始時刻 T_f 、終了時刻 T_t までとし、 $TR = [T_f, T_t)$ と定義する。時間帯の幅は $T_t - T_f$ 分である。ここで、時間帯 TR の幅が1440分(24時間)以上の場合には考慮しない。エピソード α のオカレンス $occ(\alpha) = [t_s, t_e)$ の開始日時を開始時刻、終了日時を終了時刻に変換したものを $occ'(\alpha) = [t'_s, t'_e)$ と表す。ここで、オカレンスの継続時間が1440分以上の場合には考慮しない。エピソード α においてオカレンス $occ'(\alpha) = [t'_s, t'_e)$ が $t'_s < T_t$ かつ $T_f < t'_e$ を満たす時、時間帯 $[T_f, T_t)$ でオカレンスが発生したとし、オカレンスが時間帯と重なった時間を時間帯継続時間 tr_d とする。 tr_d は $t'_e - t'_s$ と計算する。ここで、 $t'_s < T_f$ の時 $t'_s = T_f$ 、 $T_t < t'_e$ の時 $t'_e = T_t$ とする。 $t'_s > T_t$ または $T_f > t'_e$ の時 $tr_d = 0$ とし、 $t_d > 1440 - (T_t - T_f)$ かつ $t'_s < T_t$ かつ $T_f < t'_e$ の時 $tr_d = t_d - \{1440 - (T_t - T_f)\}$ とする。

Consideration of the data skew of activity time range in long-duration episode mining

Kazuhi Yasui[†], Takahiko Shintani[‡], Tadashi Ohmori[‡], Hideyuki Fujita[‡]

Faculty of Infomatics and Engineering, The University of Electro-Communications[†]

Graduate of Information Systems, The University of Electro-Communications[‡]

エピソードが所定の時間帯に偏っているかを評価するために、総時間帯継続時間を用いることはできない。総継続時間が長いエピソードは総時間帯継続時間の総和も長くなる。しかし、時間帯に重ならなかった継続時間も長い場合、その時間帯に偏って発生したとはならない。そこで、エピソードを総時間帯継続時間の総継続時間に対する割合によって評価する。この評価値を時間帯偏り度 TB と呼び、時間帯偏り度を以下の式で定義した。

$$TB = \frac{\sum_{i \in MANO(\alpha)} tr_{d_i}}{\sum_{j \in MANO(\alpha)} t_{d_j}}$$

時間帯偏り度は値が高い程その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。

時間帯を考慮した長時間エピソード抽出問題は、ユーザが指定した時間帯 RT 、最小時間帯偏り度 $mintb$ 、補正した最小総継続時間 $mindur'$ を満たすエピソードをすべて抽出することである。オカレンスの制約条件として最大オカレンス継続時間 $maxspan$ 、最大ギャップ $maxgap$ もユーザが指定しておく。ここで、補正した最小総継続時間とは、時間帯の幅に相応する最小総継続時間である。最小総継続時間は継続時間が短いエピソードを除外する条件であるため、24 時間に対する最小総継続時間は所定の時間帯を考慮したエピソードにはふさわしくない。そこで、時間帯を考慮したエピソードにおいては、最小総継続時間を 24 時間に対する時間帯の幅に従うこととした。時間帯 $TR = [T_f, T_t)$ に対して補正した最小総継続時間 $mindur'$ は次の式で計算する。

$$mindur' = mindur \times \frac{T_t - T_f}{1440}$$

ここで、 $mindur$ は最小総継続時間である。

4 評価実験

報告者が収集したライフログを用いて時間帯偏り度の高い長時間エピソードを抽出し、抽出されたエピソードが報告者の生活において特定の時間帯に偏った行動に相当するかを考察する。

本実験では、リストバンド型センサを用いて報告者が収集した 209 日分の運動状態データを用いた。運動状態データは、いつからいつまでどの運動状態を継続したかを示すデータである。運動状態は、静止、安静、デスクワーク、軽作業、作業、運動、歩行、ジョギング、非装着に分類されている。さらに、頻度の高い運動状態である静止、デスクワーク、軽作業は、継続時間によってそれぞれ長い場合 (L) と短い場合 (S) に分けられている。運動状態データは長時間エピソードマイニングにおけるイベント、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の順に並べたリストはシーケンスデータとして扱うことができる。

時間帯を $[18:00, 24:00)$ 、時間帯偏り度の最小値 $mintb$ を 0.30、補正した総継続時間の最小値 $mindur'$ を 4179 分として時間帯を考慮したエピソードを抽出した結果を表 1 に示す。表 1 は時間帯偏り度の高い 3 つの極大エピソードである。ここで、最小総継続時間 $mindur$ を 16714 分 (週に 600 分)、最大オカレンス継

続時間 $maxspan$ を 360 分、最大ギャップ $maxgap$ を 60 分とした。本実験では、はじめに $mindur'$ を満たす長時間エピソードをすべて抽出し、それぞれのエピソードに対して極小非重複オカレンスから時間帯偏り度を計算し、最小時間帯偏り度 $mintb$ を満たしたエピソードを抽出した。 $mintb$ を満たす長時間エピソードの数は、はじめに抽出した $mindur'$ を満たす長時間エピソードの約 15% のみであった。

表 1 時間帯を考慮したエピソード

エピソード	TB
L 軽作業-L 軽作業-S 歩行	0.947
S 歩行-L 軽作業	0.715
S デスクワーク-L 軽作業	0.520

表 1 のエピソード”L 軽作業-L 軽作業-S 歩行”は、総継続時間の 94.7% が 18 時から 24 時の間に発生していたことを示している。このエピソードのオカレンスは、報告者がアルバイト中の時間であった。報告者が主にアルバイトを行っている時間は指定した時間帯である。アルバイトでは立ち仕事を中心であるが、1 つの作業を続けているわけではなく途中で異なる作業に移ることがある。また、常に立ち仕事を行っているわけではない。このエピソードの L 軽作業-L 軽作業の部分はアルバイトでの行動を反映している。また、オカレンスの終了時刻はアルバイトを終えて帰宅するときであるため、このエピソードの最後の S 歩行がこの行動に対応していると考えられる。このことから、時間帯偏り度の高いエピソードは報告者が特定の時間帯に偏って行った行動に相当していたことがわかった。長時間エピソードに対して発生した時間帯の偏りを時間帯偏り度で評価することによって、時間帯に偏りのある行動に相当するエピソードを抽出できた。

5 おわりに

本研究では長時間エピソードが発生した時間帯の偏りの評価について検討した。実データを用いた実験により、時間帯を考慮した長時間エピソードは実生活で特定の時間に偏って行った行動に相当するエピソードとなる見込みを得た。今後は効率よく抽出するための手法を検討する。

謝辞

本研究は、JST、CREST の支援を受けたものである。

参考文献

- [1] R. Rawassizadeh, E. Momeni, C. Dobbins, J. Gharibshah, M. Pazzani, *Scalable Daily Human Behavioral Pattern Mining from Multivariate Temporal Data*, IEEE TKDE, 2016.
- [2] T. Shintani, T. Ohmori, H. Fujita, *Method for Comparing Long-term Daily life using Long-duration episodes*, EDBT/ICDT Workshops 2019.
- [3] H. zhu, P. Wang, X. He, Y. Li, W. Wang, B. Shi, *Efficient Episode Mining with Minimal and Non-overlapping Occurrences*, IEEE ICDM, 2010.