様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　６　月　１８　日現在

機関番号：　１２６１２
研究種目：　基盤研究(C)（一般）
研究期間：　2014～2017
課題番号：　２６３３０１２９

研究課題名（和文）大規模データからの多種の時間概念が混在するシーケンシャルパターン高速抽出技術

研究課題名（英文）Mining algorithm for sequential patterns with multiple concepts of time

研究代表者
新谷　隆彦（Shintani, Takahiko）

電気通信大学・大学院情報理工学研究科・准教授

研究者番号：３０６０４６２３

交付決定額（研究期間全体）：（直接経費）　　3,500,000円

研究成果の概要（和文）：本研究では、継続時間を持つイベントからなるシーケンスデータから多種の時間概念を考慮したイベントの順序パターンを抽出する手法を検討した。実応用に基づいて時間概念とパターンを検討し、頻度ではなく、継続時間によってパターンを評価することが重要であることがわかったため、長い時間を占めたパターンである長時間パターンを定義した。さらに、イベントの継続時間と開始時刻を考慮したパターンに拡張することで多種の時間概念を含む長時間パターンの抽出を実現した。長時間パターン抽出は冗長なカウントを省略することで処理性能を向上させた。実データを用いた実験によって提案手法の有効性を確認できた。

研究成果の概要（英文）：This research work proposed a method for finding sequential patterns containing various time concepts from sequence data of data events with duration. We investigated the time concepts and patterns by real application and clarified that it is important to evaluate patterns by pattern duration instead of frequency. We then defined a long duration pattern that occupied a long time, extended to a long duration pattern containing various time concepts by considering the duration and start time of data events, and proposed an efficient method to find long duration patterns by omitting redundant counts. We confirmed the effectiveness of our method by experiments using real data.

研究分野：　データ工学

キーワード：　データマイニング　シーケンスデータ　時間概念　長時間パターン

１．研究開始当初の背景

データ収集技術の発展と普及により膨大なデータが収集されるようになっている。これら膨大なデータに基づいて価値ある情報を見つけ出し、より効率の良い高度な社会を実現するサービスやシステムに結び付ける研究開発が情報爆発、ビッグデータなどのキーワードで進められてきた。その１つに、膨大なデータに潜む規則性や傾向などの有用な情報を抽出するデータマイニングにおけるシーケンシャルパターン抽出技術がある。時刻を持つイベントからなるシーケンスデータからイベントの時間軸上での出現順序のパターンを抽出する技術である。従来のシーケンシャルパターン抽出技術では、例えば、携帯端末、電子マネーIC カード、電子乗車券などで取得された「誰がいつ何をしたか」などの履歴データを処理対象としていたため、イベントの順序のみを考慮した技術の研究は数多く行われてきており、効率の良いアルゴリズムが報告されていた。しかし、データマイニング技術の適用範囲の広がりによって、時刻だけでなく、継続時間を持つイベントも処理対象となってきた。例えば、身体装着型センサなどにより「誰がいつからいつまで何をしていたか」を示す生活における行動などが長期間に渡って連続して記録されるようになってきた。その後、時間概念として、あるイベントが起きてから次のイベントが起きるまでの時間差であるギャップ、一連のイベントの組とみなす時間窓、等しいパターンの繰り返し、少しの違いを許容するゆらぎ、イベントが起きていた時間間隔である継続時間、イベントの発生した時間帯が考慮されるようになってきたが、複数の時間概念を混在させたパターンを考慮することができなかった。また、これまでの研究は理論的な興味という観点での研究が中心であり、数多くの優れたアルゴリズムが報告されているものの、問題を簡略に整理した上でのみ優れた手法であり、実際に利用する場面で課題を考慮していないため、実用性の面での性能が不十分であった。

２．研究の目的

膨大なデータから有用な情報を抽出するデータマイニング技術としてイベントの時間軸上での出現順序のパターンを抽出するシーケンシャルパターン抽出技術の研究が進められてきたが、継続時間を持つイベントからなるデータを処理対象とすること、様々な時間概念を考慮することができておらず、また、実用性の面での性能が十分でなかった。本研究では、シーケンシャルパターン抽出に対する実用性の面から考慮した課題として、継続時間を持つイベントデータにおける複数の時間概念を考慮したパターンと、その抽出処理性能の向上を挙げ、これら課題を解決する技術の実現を目的とする。

具体的には、パターンの発生する頻度だけではなく、ギャップ、時間窓、繰り返し、ゆらぎ、継続時間、時間帯などの時間概念について、複数の時間概念の考慮したパターンを実応用に基づいて定義する。これら時間概念によるパターンの評価についても検討する。さらに、膨大なデータに対する実用的な時間での処理を実現する方式を確立する。これにより、実世界に通用する実用性の高い技術を実現することを目指す。

３．研究の方法

本研究を実現するために、実用的な性能向上という観点から取り組んだ。様々な時間概念を考慮したパターンを実応用に基づいて定義し、その抽出手法を設計した。そして、実データを用いて性能評価を行った。

実応用として、ライフログからの生活特性抽出を対象とした。申請者らは、ライフログとして、いつからいつまでどのような運動状態を続けていたかを示す運動状態データをリストバンド型センサで 24 時間 365 日連続して取得している。この運動状態データは継続時間を持ち、運動状態データのパターンは実際の行動を示す。これらに実問題において求められる情報は何かを検討した上で様々な時間概念を考慮したパターンと効率の良いパターン抽出手法を設計した。ここでは、パターンが発生した区間から様々な時間概念によりパターンを評価する基準についても検討した。さらに、抽出手法が意味のあるパターンが抽出できるか、および、実用的な時間で処理できるかを実データを用いた評価実験によって検証した。

４．研究成果

本研究の成果として、継続時間を持つイベントからなるシーケンスデータからのシーケンシャルパターン抽出について、長い時間を占めた長時間パターンを定義すること、多種の時間概念を考慮した長時間パターンに拡張すること、および、効率良い抽出方法を確立することができた。特に、頻度ではなく継続時間によって評価した長時間パターンは時間概念を考慮する際に重要であり、従来手法では見つけ出すことができない。

(1)長時間パターンの定義

継続時間を持つイベントからなるデータに対するシーケンシャルパターンでは、そのパターンが発生した頻度だけでは有用な情報が抽出できないことが分かった。そのため、継続時間によってパターンを評価する長時間パターンを定義した。

継続時間を持つイベントからなるシーケンスデータに従来のパターン抽出手法を適用した場合には有用な情報が抽出できない。従来手法では、そのパターンが発生した回数である頻度によってパターンの良さを評価した。高頻度で発生するパターンであるほど良いパターンであるとされ、ユーザが指定し

た頻度の最小値を満たすパターンである頻出パターンのみを抽出した。ライフログにおける生活特性抽出において頻出パターンは行われた回数が多い行動に相当する。回数が多い行動は生活特性を示す重要な行動である。しかし、それぞれの行動によって要する時間の長さが異なる。長い時間を要する行動の場合、行われた回数は多くないが要した時間の総和としては長い時間を占めた行動も生活特性として非常に重要な行動である。短時間で行うことができる行動は回数を多く行うことができるが、長い時間を要する行動は行うことができる回数に限度がある。例えば、1 回行うために 3 時間を要する行動は、1 日の中では最大でも 8 回しか行うことができない。そのため、継続時間を持つイベントにおいては、頻度は高くないが、長い時間を占めたパターンを抽出する必要がある。このパターンを長時間パターンと名付けた。

イベントはアイテム、開始日時、終了日時の組からなる。アイテムはイベントの内容を示し、開始日時と終了日時の時間差がイベントの継続時間となる。シーケンスデータは、開始日時の順に並べたイベントのリストである。パターンは、アイテムを発生した順に並べたリストであり、パターンが含むアイテムの数をパターンの長さとする。シーケンスデータにおいて、パターンの各アイテムが同じ順序で現れる区間をインスタンスと呼ぶ。インスタンスの先頭のイベントの開始日時と末尾のイベントの終了日時の時間差はインスタンスの継続時間である。したがって、長時間パターンは、そのパターンのすべてのインスタンスの継続時間の総和である総継続時間によって評価する。ユーザが指定した総継続時間の最小値を満たすパターンを長時間パターンとすることとした。

(2)多種の時間概念を含む長時間パターン抽出問題の定義

総継続時間によって長時間パターンを定義したが、インスタンスに制約が必要であることがわかり、インスタンス継続時間の最大値、ギャップの最大値によってインスタンスと認める条件を設定した。さらに、イベントの継続時間と行った時間帯の情報をアイテムに付加した。そして、長時間パターン抽出問題を「処理対象のシーケンスデータ、および、最小総継続時間、最大インスタンス継続時間、最大ギャップが与えられたとき、これら条件を満たす長時間パターンをすべて抽出すること」と定義した。これによって時間概念としてギャップ、ギャップ、時間窓、繰り返し、ゆらぎ、継続時間、時間帯を含むパターンの抽出を実現した。

長時間パターンを総継続時間が長いほど良い評価をした場合、シーケンスデータ全体を 1 つのインスタンスとするパターンが最も良いパターンとなる。このパターンは、ライフログにおける生活特性抽出ではユーザの全データとなり、生活特性を示すとは言えない。このような極端に頻度が少ないパターンが長時間パターンとして抽出されることを回避しなければならない。そこで、個々のインスタンスの継続時間の最大値を制限することとした。この制約を最大インスタンス継続時間と呼ぶ。最大インスタンス継続時間により、例えばライフログにおける生活特性抽出においては、最大インスタンス継続時間によって、1 回の行動とみなすには所要時間が長すぎる場合を除外することが可能となる。また、シーケンスデータにおいてパターンが現れた区間であるインスタンスはパターンの各アイテムが同一の順序で発生している必要がある。しかし、パターンの連続するアイテムに対応するインスタンスのイベントの間に任意のイベントが挟まれることを認めるため、イベント間の時間差の最大値を制限した。これを最大ギャップと呼ぶ。これによって、パターンが現れたことの判定に誤差を許容することができる。ライフログにおける生活特性抽出では、完全に一致した動作でなくとも同一の行動とみなすことが可能となる。特に人は同じ内容の行動でも完全に同一の動作を行うわけではない。最大ギャップはこの問題を回避できる。

長時間パターンを構成するアイテムは、イベントの内容を示している。例えば、本研究で用いたリストバンドセンサで取得した運動データはイベントの内容である運動状態を静止、安静、デスクワーク、軽作業、作業、スポーツ、歩行、ジョギングで表現している。このようにシーケンスデータのアイテムは時間概念を含まない。多種の時間概念を含む長時間パターンとするため、アイテムに継続時間と時間帯の情報を付加することとした。アイテム毎にイベントを継続時間と開始時刻でクラスタリングすることで、同様の時間帯に同様の時間続けられたイベントをクラスタにまとめることができる。そして、各クラスタの代表の継続時間と開始時刻を、そのクラスタに属するイベントのアイテムに付加したシーケンスデータに変換する。このシーケンスデータから長時間パターンを抽出することで多種の時間概念を考慮できる。

以上から、最大ギャップによってギャップとゆらぎ、最大インスタンス継続時間によって時間窓、シーケンスデータからインスタンスを求めることで繰り返し、継続時間を持つイベントからなるデータから総継続時間で評価した長時間パターンを抽出することによって継続時間、アイテムにイベントの継続時間と開始時刻の情報を付加することによって継続時間と時間帯の時間概念を考慮することができることがわかる。

(3)長時間パターン抽出手法の確立

長時間パターンは、探索するパターンのインスタンスのリストを作成し、総継続時間を求め、最小総継続時間を満たすパターンを選

出することで抽出される。長時間パターンにおいてはアプリオリの性質が成り立たないため、探索候補パターンの枝刈りができないことがわかった。そこで、探索候補パターンのインスタンス作成に着目し、最小総継続時間を満たさないことを判定できた時点より後で作成するインスタンスを省略する戦略をとることとした。これによって、処理効率を向上させ、実用的な時間での処理を実現した。

あるパターン X と Y(Y は X を含む)について、Y の頻度は常に X の頻度以下となるため、X が頻出とならないときには Y の探索を省略することができた。しかし、Y の総継続時間は X の総継続時間より長くなる場合がある。Y のインスタンス数は X のインスタンス数より多くなることはないが、Y の方が長いため、それぞれのインスタンスの継続時間は長くなる。そのため、総継続時間については単調性が成り立たない。そのため、X が最小総継続時間を満たさない場合でも Y を調べなければならない。そこで、長時間パターン抽出におけるインスタンスリストの作成に着目した。パターン P にアイテム q を追加したパターン Pq のインスタンスのリストの作成は、P のすべてのインスタンスについて、そのインスタンスの末尾のイベントの終了日時以降の開始日時を持つ q のインスタンスをつなげることで作成する。P が r 個のインスタンスを持つとき、j 番目のインスタンスまで Pq のインスタンスを作成した時点で最小総継続時間を満たせないことが判明した場合、j+1 番目以降のインスタンスについては Pq のインスタンスを作成する必要がない。この性質を利用し、インスタンス作成処理を省略することとした。具体的には、P の j 番目のインスタンスまで Pq のインスタンスを作成したとき、つなげた q のインスタンスの終了時刻からシーケンスデータの末尾の時刻までの残り時間と Pq の作成済みのインスタンスリストの総継続時間の和が最小総継続時間未満となった場合、j+1 番目以降については Pq のインスタンス作成を省略する処理を行う。

また、長時間パターン抽出においては、頻度が最小総継続時間を最大インスタンス継続時間で割った値以上のパターンを抽出することとした。最小総継続時間と最大インスタンス継続時間を共に満たす長時間パターンで頻度が最小となるのは、すべてのインスタンスが最大インスタンス継続時間と等しく、その和が最小総継続時間となる場合であることを利用した。

以上から、継続時間を持つシーケンスデータから長時間パターンを抽出する処理は、①シーケンスデータから各アイテムのインスタンスリストを作成し、②各アイテムについて、アイテムを１つ追加したパターンのインスタンスリストを作成する、さらに、これらパターンにアイテムを１つ追加したパターンのインスタンスリストを作成することを繰り返し、③インスタンスリストでインスタンスの継続時間の総和が最小総継続時間を満たすときに長時間パターンとして出力することを繰り返すこととした。ただし、前処理としシーケンスデータの各アイテムに継続時間と開始時刻の情報を付加しておく。ここで、②において、最大ギャップを満たさないインスタンスをつなげないこと、インスタンスが最大インスタンス継続時間を満たさない場合はリストについかしないこと、および、最小総継続時間を満たさないことが判明した時点より後のインスタンス作成を省略すること、また、③において、重複しないインスタンスの継続時間の和を求めることによって、処理量を削減する。

リストバンド型センサで取得した 562 日分の運動状態データの実データに対して、週に 20 時間の最小総継続時間として長時間パターンを抽出したときの処理時間は 3.2 秒であり、14.3%のインスタンスの作成を削減することができており、実用的な時間での処理を確認するとともに、多種の時間概念を考慮したパターンが抽出できることを確認できた。

５．主な発表論文等

〔学会発表〕（計 4 件）

① 楊デイ、新谷隆彦、大森匡、藤田秀之、リストバンド型センサで取得した腕の向きのパターンによる運動状態の分類の検討、情報処理学会第 77 回全国大会、2015.3.17、京都大学（京都府）

② 櫻田滋大、新谷隆彦、大森匡、藤田秀之、継続時間を閾値としたエピソードマイニングの提案、電子情報通信学会総合大会、2015.3.12、立命館大学（滋賀県）

③ 富金輝、新谷隆彦、大森匡、藤田秀之、長時間エピソードマイニングにおけるインスタンス数え上げ処理量低減の検討、電子情報通信学会総合大会、2016.3.16、九州大学（福岡県）

④ 新谷隆彦、中島彩花、大森匡、藤田秀之、リストバンド型センサで取得した運動量のレベル化による異なる期間の生活比較手法の検討、第 13 回日本感性工学会春季大会、2018.3.27、名古屋大学（愛知県）

６．研究組織

(1)研究代表者

新谷　隆彦（SHINTANI, Takahiko）
電気通信大学・大学院情報理工学研究科・准教授
研究者番号：３０６０４６２３