博 士（情報科学） 河 東 孝

学 位 論 文 題 名

Efficient Algorithms for Extracting Frequent Episodes form Event Sequences

（イベント例からの頻出エピソード抽出に対する効率良いアルゴリズム）

学位論文内容の要旨

エピソード抽出 (episode extraction) またはエピソード発見 (episode mining) は，時系列データに対するデータマイニング手法の一つであり，1997 年に Mannila らによって導入された．エピソードとは，一群のイベントの出現に関する半順序制約を表すパターンであり，イベントを頂点とし，イベントの出現順序に関する時間制約を有向辺とするラベル付き非巡回有向グラフとして定義される．各頂点はイベント型と呼ばれる記号をラベルとして持つ．エピソード抽出の目的は，入力イベント列からすべての頻出エピソードを抽出することである．エピソードはイベント間の半順序関係を記述できるため，従来から時系列マイニングで扱われてきた全順序しか表現できない系列パターンよりも豊かな表現力を持つ．

一般に，頻出パターン発見問題では，入力サイズに対して指数的に大きな数の解が存在するため，パターン発見アルゴリズムの性能を通常の計算量で測ることは難しい．そのため，列挙アルゴリズムの枠組みを用いて，出力 1 個あたりのアルゴリズムの計算時間で測る．Mannila らは一般的な設定で頻出エピソード発見方法を考察しているが，一般のエピソードの族に対しては効率良い頻出エピソード発見アルゴリズムの構成は難しいと考えられる．また，明示的な性能保障をもつ効率良いアルゴリズムについても未解決の問題である．そこで，本研究では，エピソードの部分族に対して，頻出エピソードを抽出する効率良いアルゴリズムを考察し，その性能の理論的解析を行う．とくに，従来のアルゴリズムがの多くは候補パターンの空間の幅優先探索を用いているのに対して，本研究では，深さ優先探索を用いる高速かつ記憶効率の良いアルゴリズムの開発を追求する．

エピソード抽出問題に関して，現在までに提案されている複数の効率的アルゴリズムの多くは，幅優先探索 (breadth-first search, BFS) を用いた階層探索アルゴリズムであり，一度発見した頻出エピソードを主記憶上に保持する必要がある．そのため，計算時間は出力多項式時間である一方で，同じ出力の重複を回避するために，中間解を記憶するための入力サイズの指数サイズの記憶領域を必要とするという問題点をもつ．そのため，大きなデータ量を持つ実世界の問題に対して，実際の計算機上で動かないことが多い．これに対して，本研究で提案する深さ優先探索 (depth-first search, DFS) を用いるエピソード抽出アルゴリズムは，サイズが最小のエピソードから開始して，候補エピソードを順次拡大しながら，候補エピソードを探索する．これには，候補エピソードの空間上に家系木 (family tree) と呼ばれる木構造をした探索空間をあらかじめ理論的に定めておき，その上で現在の計算状況を記憶するためのスタックを用いたバックトラック探索を行って，族に含まれるすべての頻出エピソードを重複なく列挙する．このため，中間解を記憶するための大きな記憶領域が必要

ない.

この効率良いエピソード抽出アルゴリズムの鍵は, 対象とする部分族に対する家系木と呼ばれる探索空間の設計にある. 本研究では, エピソードのさまざまな部分族に対して, 探索空間の設計方法を与え, 深さ優先探索を用いる効率良いエピソード抽出アルゴリズムを提案する. ここでの目標は, 出力解 1 個当たり, 入力サイズの多項式時間と多項式領域で計算する多項式遅延・多項式領域アルゴリズムの開発である. さらに, 応用例として, 医学分野における細菌検査の時系列データからのエピソード抽出を考察し, 提案したアルゴリズムを適用し, その有効性を検証する. 以下では, Σ および, N, n で, それぞれ, 文字のアルファベットおよび, 入力イベント系列の総イベント数と入力長を表す. また, 3 章以降の各章の末尾では, 提案アルゴリズムに対する人工データおよび実世界データを用いた評価実験を行う.

初めに, 第 2 章で基本的な概念と定義を準備した後で, 第 3 章では, 入力イベント列から, 頻出菱形エピソードを効率よく抽出する問題を扱う. 菱形エピソード (diamond episode) とは $a \to E \to b$ という形のエピソードであり, これはイベント a の出現後に, イベント集合 E 中の全イベントが出現し, それにイベント b の出現が続くことを意味する. これに対して, イベント列からすべての頻出菱形エピソードを, 出力一つ当たり $O(|\Sigma|^2 n)$ 時間と $O(|\Sigma| + n)$ 領域で重複なく抽出する多項式遅延・多項式領域アルゴリズム PolyFreqDmd を与える.

第 4 章では, イベント集合 A と B に対して, $A \to B$ のような形をした二部エピソード (bi-partite episode) の族を提案する. このエピソードは, A 中の全イベントの出現が, B 中の全イベントの出現に先行することを意味する. これに対して, イベント列からすべての頻出二部エピソードを, 出力一つ当たり $O(|\Sigma|^2 N)$ 時間と $O(|\Sigma|^2 n)$ 領域で重複なく抽出するアルゴリズム Bipar を導入する.

第 5 章では, はじめに, イベント集合 A_i $(1 \le i \le k)$ に対して, $\langle A_1, ..., A_k \rangle$ という形の多部エピソード (multi-partite episodes) を導入する. これは, 任意の $1 \le i < k$ に対して, 集合 A_i 中のすべてのイベントが集合 A_{i+1} 中のすべてのイベントの前に出現することを意味する. これに対して, イベント列からすべての頻出多部エピソードを, 出力一つ当たり $O(|\Sigma|^2 wn)$ 時間と $O(|\Sigma|^2 k)$ 領域で重複なく抽出するアルゴリズム Kpar を導入する.

第 6 章では, エピソードの高速抽出に有用なエピソードの直列構成可能性と呼ばれる性質について考察する. エピソードの直列構成可能性とは, エピソードのデータ中の出現位置が, そのエピソードが含む系列エピソードと呼ばれる単純な一次元の計上をしたエピソードのすべての出現位置全体から, 一意に再構成されることを意味する性質である. 一方で, 与えられたエピソードが非並列性 (parallel-freeness) を持つとは, ラベルが等しい頂点の間には必ず有向辺があることをいう. 本章の主結果として, 任意のエピソードの族において, エピソードの非並列性が, その直列構成可能性の必要十分条件であることを証明する.

第 7 章では, エピソード抽出アルゴリズムを細菌検査データに適用することで, 院内感染の原因と考えられている菌交代現象と薬剤耐性変化に関する時系列規則を表す頻出エピソードを抽出する. 具体的には, 菱形エピソードの特殊形である扇状エピソードと系列エピソード, 制約付きの二部エピソードに関するマイニングアルゴリズムを, 大阪府立急性期総合医療センターから提供された細菌検査データに適用することで, 菌交代現象と薬剤耐性変化に関する時系列規則を抽出する.

最後に, 第 8 章で本論文の結論と今後の課題を述べる.

学位論文審査の要旨

主　査　教　授　有　村　博　紀
副　査　教　授　原　口　　誠
副　査　教　授　湊　　真　一
副　査　教　授　今　井　英　幸
副　査　准教授　喜　田　拓　也

学位論文題名

Efficient Algorithms for Extracting Frequent Episodes form Event Sequences

（イベント例からの頻出エピソード抽出に対する効率良いアルゴリズム）

データマイニング (data mining) は，大量の蓄積されたデータから，人間にとって有用な規則性やパターンを発見するための効率的な計算手法の研究である．データマイニング分野では，従来，表形式のデータが主に研究されてきた．これに対して，2000年前後からグラフマイニングの名称で，時系列やグラフ等の離散構造データに対するデータマイニングの研究が盛んになっている．本研究では，グラフマイニングの一分野である時系列からの離散的パターンマイニングについて考察する．

エピソード抽出 (episode extraction) は，1997年にデータマイニング分野の Mannila らによって導入された時系列からの離散構造データマイニング手法である．エピソードは，複数のイベント間の依存関係を表現する有向グラフであり，与えられたエピソードのクラスに対して，エピソード抽出アルゴリズムは，イベントの時系列である入力イベント系列中に指定回数以上出現するエピソードを漏れなく発見する．エピソード抽出は，イベント間の複雑な依存関係の発見に適している．その一方で，任意のグラフ構造をパターンに許すために，パターン発見は大きな計算量をもつ．そのため，エピソード抽出の研究においては，応用の観点から十分な表現力をもち，同様に効率良い計算を許すようなエピソードの部分クラスの同定と，これらのクラスに対する効率良いアルゴリズムの設計が重要な研究課題となっている．

本研究では，エピソードのさまざまな部分族を新たに提案し，これらに対して深さ優先探索を用いた領域効率の良いエピソード抽出アルゴリズムを提案し，その計算量の理論的解析を与えている．さらに，応用として 医学分野における細菌検査の時系列データに提案アルゴリズムを適用し，その有効性を検証している．特に本研究の特色は，深さ優先探索に基づく設計原理を用いて，エピソード抽出アルゴリズムの系統的な設計と開発を行っている点にある．この点について，従来の研究は，主に探索空間の幅優先探索を用いており，重複解の回避のために入力の指数サイズの作業用記憶領域を必要としていた．これに対して，本研究で提案する深さ優先探索を用いたエピソード抽出アルゴリズムは，木構造の探索路の上でバックトラック探索を行い，すべての解を効率よく重複なしに列挙する

ことが可能である. このため, 中間解を記憶するための大きな記憶領域を必要とせず, 実際のデータの多くにおいて, 記憶領域効率が良いという長所をもつ.

第一の研究では, 入力イベント列から, 頻出菱形エピソード (diamond episodes) の抽出問題を提案し, その多項式遅延・多項式領域抽出アルゴリズムを与えている. 第二の研究では, 二部エピソード (bi-partite episode) のクラスを提案し, そのエピソード抽出問題の多項式遅延・多項式領域抽出アルゴリズムを与えている. 第三の研究では, 多部エピソード (multi-partite episodes) のクラスを提案し, そのエピソード抽出問題の多項式遅延・多項式領域抽出アルゴリズムを与えている. 第四の研究では, 第一から第三の研究で効率化の鍵となっているエピソードクラスの直列構成可能性と呼ばれる性質について考察し, 任意のエピソードのクラスにおいて, エピソードの非並列性が直列構成可能性の必要十分条件であることを証明している. 第五の研究では, 提案のエピソード抽出アルゴリズムを, 実際の細菌検査データに適用することで, 院内感染の原因と考えられている菌交代現象と薬剤耐性変化に関する時系列規則を抽出する実験を行っている.

本論文の成果は次のようにまとめられる

1. データマイニングにおけるイベント系列からのエピソード抽出問題について, 十分な表現力をもつと同時に, 効率良い計算を許すエピソードの種々の部分クラスを提案し, これらのクラスに基づくイベント系列マイニングの有用性を示した.

2. 特に, 種々のエピソード抽出問題に対して, 深さ優先探索に基づくアルゴリズムの系統的な設計に取り組み, 効率良いエピソード抽出アルゴリズムを開発した. これにより, データマイニングにおけるエピソード抽出の時間領域計算量に関する新しい結果を示した.

3. 実際の大規模イベント系列解析問題に提案手法を適用し, その有効性を示した.

これを要するに, 著者は, データマイニングにおけるエピソード抽出アルゴリズムの設計について, 新しい方法論を提案し, 種々の問題に対して効率良い解法が構成可能であるという新知見を得たものであり, データ工学と知能情報学において貢献するところ大なるものがある. よって著者は, 北海道大学博士 (情報科学) の学位を授与される資格あるものと認める.