飽和系列パターンの多項式時間列挙アルゴリズムの実装

谷　研究室　　栗原　純
Jun Kurihara

概要

有村　博紀、宇野　毅明によって提案された飽和系列パターンの多項式時間列挙アルゴリズムを [1]C 言語で実装する。

1　はじめに

データベースからの頻出アイテム集合発見 [2] は，データマイニングにおける代表的な研究トピックである．この頻出アイテム集合発見において，一般に膨大な数の互いに類似した頻出系列パターンが出力されるという問題がある．この問題に対して，頻出飽和系列パターンと呼ばれる特別なクラスの頻出系列パターンだけを発見するアプローチが提案されている．ここで，飽和系列パターンとは，データベース上で同じ出現位置を持つような互いに同値な系列パターンの中で，系列パターンの包摂関係（部分集合関係）に関して極大な系列パターンをいう．実際データにおいて，対応する頻出飽和アイテム集合の数は頻出集合の数よりも著しく少ないことが多く，頻出飽和系列パターンを用いることで，系列パターン発見を効率化できると考えられる．

2　飽和系列パターン

本研究では，この飽和アイテム集合の概念を系列パターン発見に拡張して飽和系列パターンのクラスを導入し，与えられたデータベースにおける頻出飽和系列パターンの効率よい列挙について考察する．

系列パターンは任意の記号例であり，系列データベースは系列の集合である．系列パターンは，ある系列の (必ずしも連続しない) 部分系列 (subsequence) となるときに，その系列に出現するという．飽和系列パターンが頻出であるとは，それが与えられた最小頻度パラメータ $\alpha \geqq 0$ 以上の数の系列に出現することをいう．指定された出現のクラスに対する飽和系列パターンとは，与えられたデータベース中で同じ出現集合をもつような系列パターン中で，部分系列関係に関して極大な系列パターンである．

現在までに出現集合として文書出現集合 $DO(\cdot)$ に基づく頻出飽和系列パターンを列挙するさまざまな経験的アルゴリズムが提案されている．しかし，これらのアルゴリズムの出力多項式時間性は示されていない．

これに対し，ここではより詳細な情報を与える出現集合として，パターンの右端の出現位置である位置出現集合 $PO(\cdot)$ を採用し，これに基づく位置飽和系列パターンを導入する．このクラスに対して，系列データベースからの頻出位置飽和系列パターンの列挙問題を考察する．

主結果として，系列データベース上のすべての位置頻出飽和パターンを，系列パターン一つあた $O(|\sum|Nl)$ 時間で重腹なしに列挙するアルゴリズム $EnumClosedSeq$ を与える．ここに，$N = ||S||$ は入力データベースサイズであり，l は S 中の最長の系列の長さである．

提案のアルゴリズムの鍵は，次のように定義される系列パターンの逆探索性である:長さ $n \geqq 0$ の任意の位置飽和系列パターンに対して，その右端の記号を取り除いて得られる長さ $k-1 \geqq 0$ の系列パターンはやはり位置飽和系列パターンである．これより，すべての位置飽和パターンの親として，サイズが一つ小さな位置飽和パターンが一意に定まる．

この過程を逆にして，最小の飽和パターン ϵ から出発して，与えられた親からその子供を生成するというやり方で，頻出飽和系列パターン全体の全域木を構成できる．この全域木を，根からはじめて深さ優先探索を行うことにより、$EnumCloseSeq$ はすべての飽和パターンを一つあたり入力の多項式時間で重複なく列挙する．

3　準備

集合 A に対して，A の要素数を $|A|$ で表し，自然数全体の集合を $N = \{0, 1, 2, \ldots\}$ で表す．任意の有限アルファベット $\sum$ に対して，空列を ϵ と書き，$\sum$ 上の有限系列全体を $\sum^*$ と表す．$\sum$ 上の長さ $n \geqq 0$ の系列 $s = a_1 \cdots a_n \in \sum^*$ に対して，s の長さを $|s| = n$ で表し，各 $i = 1, \ldots, n$ に対して $s[i] = a_i$ と表す．

3.1　準備

$\sum$ を有限アルファベットとする．長さ $n \geqq 0$ の系列パターンは任意の系列 $s = a_1 \cdots a_n$ である．$P = \sum^*$ で $\sum$ 上の系列パターン全体の集合を表す．$\sum$ 上の系列

データベース（データベース）は，系列の集合

$$S = \{s_1, \ldots, s_m\} \subseteqq 2^{\Sigma^*}$$

である．本研究をとおして，S は少なくとも二つの異なる記号を含むと仮定する．ここに，$|S| = m$ で S の要素数を表し，$||S|| = \sum_{s \in S} |s|$ で S の総サイズを表す．二つの系列 $x = a_1 \cdots a_m$ と $y = b_1 \cdots b_n$ （$m, n \geqq 0$）に対して，$x = a_1 \cdots a_m = b_{\varphi(1)} \cdots b_{\varphi(m)}$ を満たすような y の添え字の列 $1 \leqq \varphi(1) < \cdots \varphi(m) \leqq n$ が存在するならば，$x \sqsubseteq y$ と書き，x は y の部分系列である，または x は y に含まれるという．さらに，$x \sqsubseteq y$ かつ $y \sqsubseteq x$ ならば $x \sqsubset y$ と書き，系列 x は系列 y の真部分系列である．または，x は y に真に含まれるという．

関数 $\varphi : dom(x) \to dom(y)$ を x の添え字に対する y の添え字への写像関数という．パターン x に対する系列 y への写像関数全体のなす集合を $M(x, y)$ と書く．このとき、位置 $ro(\varphi) = \varphi(|x|) \in dom(y)$ を x の y における（右端の）位置出現という．

パターン x の S 上での位置出現集合 $PO^S(x)$ と文書出現集合 $DO^S(x)$ を次のように定義する：

$$PO^S(x) = \{\langle s, ro(\varphi)\rangle \mid s \in S, \exists \varphi \in M(x, s)\}$$

$$DO^S(x) = \{s \in S \mid \exists \varphi \in M(x, s)\}$$

非負整数である最小頻度パラメータ $0 \leqq a \leqq |S|$ に対して、系列 $x \in P$ が S 上で頻出であるとは $|DO^S(x)| \geqq a$ が成立することをいう．

3.2 位置出現に関する飽和系列パターン

S 上の位置飽和パターンとは，任意のパターン x で，x を部分系列として真に含み ($xsqsubsety$)，$PO^S(y) = PO^S(x)$ を満たすようなパターン y をもたないものをいう．

S 上の文書飽和パターンとは，任意のパターン x で，x を部分系列として真に含み ($xsqsubsety$)，$DO^S(y) = DO^S(x)$ を満たすようなパターン y をもたないものをいう．

4 多項式時間列挙アルゴリズム

4.1 拡張と局所飽和性

任意の系列パターン $x, y \in P$ に対して，$|y| = |x| + 1$ かつ $x \sqsubseteq y$ ならば，$x \to y$ と書き，y は x の拡張であるという．これは，x の任意の場所への適当な文字一つの挿入によって，x から y が得られることを意味する．

パターンの拡張において，記号を挿入する場所をパターンの右端に制限して，次の定義を得る．

[定義 1] パターンの最右拡張

任意のパターン $x \in P$ と文字 $a \in \Sigma$ に対して，$x \to_R xa$ と書き，パターン $xa \in P$ を x の最右拡張という．また，パターン y が x の拡張であるが，最右拡張でないならば，$x \to_{NR} y$ と書き，y を x の非最右拡張という．長さ 1 の任意の文字列 $a \in \Sigma$ は，空語 ϵ の最右拡張である．任意のパターン x の最右拡張と非最右拡張は，x の拡張でもある．

4.1.1 [定理 1] 飽和パターンの拡張による特徴づけ

任意の系列パターン x に対して，次の (1) と (2) は等価である:

(1)x が S 上で位置飽和である．

(2)$PO^S(y) = PO^S(x)$ を満たすような x の拡張 $y \in p(x)$ が存在しない．

4.2 頻出飽和パターンに対する逆探索戦略

任意のパターン x, yy に対して，もし x の最右拡張が y ならば，すなわち，$y = xa(\exists a \in \Sigma)$ ならば，x を最右拡張に関する y の親といい，y を最右拡張に関する x の子という．

4.2.1 [定理 2] 系列パターンの逆探索性

任意のパターン $x \in P$ と文字 $a \in \Sigma$ に対して，もしパターン xa が S 上で位置出現に関して飽和ならば，その親 x も位置出現に関して飽和である．

4.3 多項式時間列挙アルゴリズム

空パターンからはじめて，親パターンから子パターンをすべて試しながら、探索木を深さ優先探索することで，S 上のすべての飽和パターンを重複なく列挙する．各パターンにおいて，それが飽和性と頻出性を満たさなくなった場合は，ただちにその子孫の探索を打ち切って，バックトラックし次の候補を探索する．

5 今後の課題

・より多くの種類に分かれたデータを使用し、位置出現に関する飽和系列パターンのアルゴリズムを実装すること．

・文書出現に関する飽和系列パターンの多項式時間列挙アルゴリズムを実装すること．

参考文献

[1] 有村　博紀，宇野　毅明，飽和系列パターンの多項式時間列挙アルゴリズム，2004.

[2] 有村　博紀，データベースからの頻出アイテム集合発見，2005