LF-002

大規模データ系列中に頻出する部分系列のオンライン抽出アルゴリズム
On-line Extraction Algorithm of Frequent Subsequences from a Single Very-Long Data Sequence

石原 龍一 † 岩沼 宏治 ‡ 鍋島 英知 ‡
Ryuichi ISHIHARA Koji IWANUMA Hidetomo NABESHIMA

1. はじめに

本論文では，大規模な単一データ系列中に出現する頻出部分系列のオンライン型抽出アルゴリズムを提案し，その有用性を実験的に示す．

従来のボトムアップタイプの Apriori 型アルゴリズム [1] や，その改良型であるトップダウン型の PrefixSpan [5] は，どれも系列データベースを何度も読み直す必要があり，そのために実行時間が膨大になるという問題を抱えている．そこで，近年ではデータベース (ストリームデータ) を一度しか読まないオンライン型アルゴリズムの研究が盛んに行われ始めている [2]．

系列データを対象とするデータマイニングは，顧客の購入履歴，遺伝子情報，時系列テキスト，計算機ログの解析など数多くの応用を持っている．これまでにも多くの研究が行なわれている [1, 5, 7] が，これまでは，頻出する単一要素の抽出に関する研究が多く，頻出する部分系列を抽出する研究は殆どない [2]．

本論文では，先行研究 [6, 3] で提案した系列全体頻度を基礎として，長大な系列データ上の頻出する部分系列を高速に自動抽出するオンライン型アルゴリズムの提案を行い，予備的な実験的な評価を行なう．筆者の知る限りにおいて，これまで頻出部分系列をオンラインで抽出する試みはなく，初めての試みと考えられる．

2. 準備

本論文で扱う表記法や，用語の定義を以下に示す．全てのアイテムの集合を $\mathcal{I} = \{i_1 i_2 \cdots i_k\}$ とする．

定義 1 アイテム集合系列 (以下，系列と呼ぶ) とはアイテムの集合 $s_1, s_2, \ldots, s_m \subset \mathcal{I}$ の列であり，$\langle s_1 s_2 \cdots s_m \rangle$ と表記する．系列 α を $\langle s_1 s_2 \cdots s_m \rangle$ とするとき，$suf(\alpha, i)$ で**接尾辞** (suffix) 系列 $\langle s_i \cdots s_m \rangle$ を表す．また**空列**を λ で表す．系列 $\alpha = \langle s_1 s_2 \cdots s_m \rangle$ が系列 $\beta = \langle t_1 t_2 \cdots t_n \rangle$ の**部分系列**であるとは，$s_1 \subset t_{j_1}, s_2 \subset t_{j_2}, \cdots, s_m \subset t_{j_m}$ を満たす整数 $1 \leq j_1 < j_2 < \cdots < j_m \leq n$ が存在する場合をいい，$\alpha \sqsubseteq \beta$ と表す．

以下では簡単化のために，アイテム $a_1, a_2, \ldots, a_l \in \mathcal{I}$ の集合を $(a_1 a_2 \cdots a_l)$ と表記する．また系列中では単一要素集合の丸括弧を省略し，$\langle (ac)(a)(b)(c)(bc)(b) \rangle$ を $\langle (ac)abc(bc)b \rangle$ のように略記する．このとき，例えば $\langle abc \rangle$ は $\langle a(bd)d(dc) \rangle$ の部分系列となる．

定義 2 系列データベースと呼ばれる系列の集合 S を考える．また S 中での系列 α の出現頻度を返す**出現頻度関数**を，S と α の組を非負実数へ写像する関数 $M(S, \alpha)$ とする．このとき，M を用いた S 中の**頻出系列** α とは，与えられた閾値 MS (以下，**最小サポート**と呼ぶ) に対して，$M(S, \alpha) \geq MS$ を満たすものを言う．S 中の頻出系列 α が**極大**であるとは，$\alpha \sqsubseteq \beta$ かつ $\alpha \neq \beta$ なる頻出系列 β が S 中に存在しない場合をいう．

Agrawal に始まる系列パターンマイニングの研究 [1, 5] では，系列データベース S は複数の系列の集合と仮定している．本研究では，先行研究 [6, 3, 4, 7] と同様に，長大なデータ系列一本からなる系列データベース S を考える．S 中に繰り返し出現する部分系列，即ち極大な頻出部分系列の高速なオンライン抽出法を考察する．

本論文でのオンラインアルゴリズムとは，単一系列データベースの末尾に新たなデータが追加された時，これまでの結果を利用して，漸近的に極大頻出パターンを更新して出力するアルゴリズムのことである．

3. 系列全体頻度

M を出現頻度関数とするとき，系列データベース S と任意の 2 つの系列 α, β に対して，$\alpha \sqsubseteq \beta$ のとき必ず $M(S, \alpha) \geq M(S, \beta)$ となるならば，M は**逆単調** (anti-monotonic) であると言う．出現頻度の逆単調性は，頻出系列マイニングの効率的化に非常に重要であるが，単一系列データベース上の素朴な頻度尺度の殆んどは逆単調ではない [6, 3]．Mannila ら [4] は，スライド窓 (sliding window) という逆単調性を満たす出現尺度を提案しているが，同一の部分系列の重複数え上げが生じるなど，尺度の合理性に問題がある [6]．先行研究 [6, 3] で提案された**系列全体頻度**は，逆単調性を満たし，重複数え上げが生じない合理的な頻度尺度である

定義 3 (Iwanuma et.al.[3]) 単一系列データベース S を $S = \langle s_1 s_2 \cdots s_m \rangle$ とし，α を系列 $\langle t_1 t_2 \cdots t_n \rangle$ とする．S における α の**系列先頭頻度** $H\text{-}freq(S, \alpha)$ を以下で定義する．

$$H\text{-}freq(S, \alpha) = \sum_{i=1}^{m} \delta(suf(S, i), \alpha)$$

ここで δ は以下の関数と定める．

$$\delta(\langle s_i \cdots s_m \rangle, \langle t_1 \cdots t_n \rangle) = \begin{cases} 1 & \text{if } t_1 \subset s_i \text{ and } \langle t_2 \cdots t_n \rangle \sqsubseteq \langle s_{i+1} \cdots s_m \rangle \\ 0 & \text{otherwise} \end{cases}$$

例として，系列 $S = \langle a(ab)bbc \rangle$ を考えると，

$$H\text{-}freq(S, \langle a \rangle) = H\text{-}freq(S, \langle ac \rangle) = 2$$

†山梨大学大学院医学工学総合教育部修士課程 コンピュータ・メディア工学専攻
‡山梨大学大学院 医学工学総合研究部 コンピュータ・メディア工学専攻担当，{iwanuma,nabesima}@iw.media.yamanashi.ac.jp

系列先頭頻度は，文字通り系列の先頭要素の出現頻度を数えるので重複数え上げは生じない．しかし後続要素の出現頻度を考慮しないため，逆単調性を満たさない．そこで全ての部分系列の頻度を考慮した尺度を考える．

定義 4 (Iwanuma et.al. [3]) S を単一系列データベース，α を任意の系列とする．S における α の**系列全体頻度** $T\text{-}freq(S,\alpha)$ を以下で定義する．

$$T\text{-}freq(S,\alpha) = \min_{\gamma \sqsubseteq \alpha}(H\text{-}freq(S,\gamma))$$

例えば $S = \langle a(ab)bbc\rangle$ について

$$\begin{aligned} &T\text{-}freq(S,\langle ac\rangle) \\ &= min(H\text{-}freq(S,\langle ac\rangle), H\text{-}freq(S,\langle a\rangle), H\text{-}freq(S,\langle c\rangle)) \\ &= min(2,2,1) = 1 \end{aligned}$$

となる．また最小サポート値を 2 としたとき，$S' = \langle (ab)c(abc)(abc)\rangle$ 中の極大頻出系列は，$\langle (ab)(abc)\rangle$ と $\langle c(abc)\rangle$ の 2 つとなる．

系列全体頻度は逆単調性を満たす [3]．定義 4 より，n の系列 α の系列全体頻度を求めるためには，α の全ての部分系列（α の長さを n とするとき，α の 部分系列は $2^n - 1$ 個存在する）の系列先頭頻度を求める必要がある．しかし実際には，次の補題から α の全ての接尾辞（n 個存在）の系列先頭頻度の最小値を求めるだけでよい．

補題 1 (Iwanuma et.al. [3]) S を単一の系列データベースとし，α を系列 $\langle t_1 t_2 \cdots t_n\rangle$ とする．このとき，以下が成り立つ．

$$T\text{-}freq(S,\alpha) = \min_{i=1}^{n}(H\text{-}freq(S, suf(\alpha,i)))$$

次節で分かるように，補題 1 はオンライン型計算のために極めて重要である．系列全体頻度のその他の詳細は，先行研究 [6, 3] を参照していただきたい．本論文では，系列全体頻度を前提としたときの，極大頻出部分系列のオンライン型高速抽出アルゴリズムについて考察する．

4. オンライン抽出型アルゴリズム

本章では単一系列データベース中の頻出系列の高速抽出計算法を導出する．オンライン化が可能なのは，以下の性質が成り立つからである．

系列 $\langle s_1 \cdots s_m\rangle$ の末尾にアイテム集合 s' を連結した系列 $\langle s_1 \cdots s_m\ s'\rangle$ を，S の s' による**末尾拡大**と呼び，$S \circ s'$ と略記する．また単一系列データベース S 中の**アイテム集合 s の出現頻度**を $T\text{-}freq(S,\langle s\rangle)$ と定め，$Fr(S,s)$ と略記する．$Fr(S,s)$ は S 中の s の単純な出現回数 (部分集合として出現する場合も含める) と一致する．最小サポート値を MS とするとき，アイテム集合 s が単一系列データベース S で**頻出**であるとは，$Fr(S,s) \geq MS$ である場合を言う．

定理 1 (系列全体頻度を用いた極大頻出系列の漸近性) MS を最小サポート値，S を単一系列データベースとし，α を S 上の任意の極大頻出系列とする．また s' を任意のアイテム集合，t' を以下の 2 つの条件を満たす s' の部分集合と仮定する．

1. t' は末尾拡大 $S \circ s'$ 上の頻出アイテム集合である．
2. t' は s' の部分集合の中で極大である．即ち t' を真に含む s' の部分集合で，$S \circ s'$ 上の頻出アイテム集合となるものは存在しない．

このとき α の t' による末尾拡大 $\alpha \circ t'$ は，末尾拡大 $S \circ s'$ 上における極大頻出系列である．

証明: $S = \langle s_1 s_2 \cdots s_m\rangle$，$\alpha = \langle t_1 t_2 \cdots t_n\rangle$ と仮定すると，2 つの拡大 $S \circ s'$ と $\alpha \circ t'$ はそれぞれ $S \circ s' = \langle s_1 s_2 \cdots s_m\ s'\rangle$，$\alpha \circ t' = \langle t_1 t_2 \cdots t_n\ t'\rangle$ なる形になることに注意する．以下の 2 つに分けて証明する．

1. 頻出性，即ち $T\text{-}freq(S \circ s', \alpha \circ t') \geq MS$
2. 系列 $\alpha \circ t'$ の $S \circ s'$ 上での極大性

[頻出性の証明] まず定理の仮定より $T\text{-}freq(S,\alpha) \geq MS$ が成り立つ．よって補題 1 より，α の全ての接尾辞 $suf(\alpha,i) = \langle t_i t_{i+1} \cdots t_n\rangle$ $(i = 1,\ldots,n)$ に対して以下が成り立つ．

$$H\text{-}freq(S,\langle t_i t_{i+1} \cdots t_n\rangle) \geq MS$$

ここで拡大系列 $\alpha \circ t'$ の任意の接尾辞 $suf(\alpha \circ t', i) = \langle t_i t_{i+1} \cdots t_n\ t'\rangle$ $(i = 1,\ldots,n)$ に対して，系列先頭頻度 $H\text{-}freq(S \circ s', \langle t_i t_{i+1} \cdots t_n\ t'\rangle)$ を考える． 定義 3 より

$$\begin{aligned} &H\text{-}freq\ (S \circ s', \langle t_i t_{i+1} \cdots t_n\ t'\rangle) = \\ &\qquad \delta(\langle s_1 s_2 \cdots s_m\ s'\rangle, \langle t_i t_{i+1} \cdots t_n\ t'\rangle) + \\ &\qquad \delta(\langle s_2 \cdots s_m\ s'\rangle, \langle t_i t_{i+1} \cdots t_n\ t'\rangle) + \\ &\qquad \vdots \\ &\qquad \delta(\langle s_m\ s'\rangle, \langle t_i t_{i+1} \cdots t_n\ t'\rangle) + \\ &\qquad \delta(\langle s'\rangle, \langle t_i t_{i+1} \cdots t_n\ t'\rangle) \end{aligned}$$

定理の仮定より $t' \subseteq s'$ なので，δ の定義 (定義 3 参照のこと) より任意の $k = 1,\ldots,m$ に対して以下が成り立つ．

$$\begin{aligned} &\delta(\langle s_k \cdots s_m\ s'\rangle,\ \langle t_i t_{i+1} \cdots t_n\ t'\rangle) = \\ &\qquad \delta(\langle s_k \cdots s_m\rangle,\ \langle t_i t_{i+1} \cdots t_n\rangle) \end{aligned}$$

よって系列先頭頻度の定義から，各 $i = 1,\ldots,n$ に対して以下が明らかに成り立つ．

$$\begin{aligned} &H\text{-}freq(S \circ s', \langle t_i t_{i+1} \cdots t_n\ t'\rangle) \\ &= H\text{-}freq(S, \langle t_i t_{i+1} \cdots t_n\rangle) + \delta(\langle s'\rangle, \langle t_i t_{i+1} \cdots t_n\ t'\rangle) \\ &\geq H\text{-}freq(S, \langle t_i t_{i+1} \cdots t_n\rangle) \geq MS \end{aligned}$$

更に定理の仮定より，t' は拡大 $S \circ s'$ 上の頻出アイテム集合であるから，$H\text{-}freq(S \circ s', \langle t'\rangle) \geq MS$ は明らかである．よって補題 1 から $T\text{-}freq(S \circ s', \alpha \circ t') \geq MS$ が成り立つ．

[極大性の証明] 拡大系列 $\alpha \circ t'$ の極大性は背理法で証明する．$\alpha \circ t'$ が $S \circ s'$ 上で極大でないと仮定すると，$S \circ s'$ 上の極大頻出系列 β で $\alpha \circ t' \subset \beta$ かつ $\alpha \circ t' \neq \beta$ なる $\beta = \langle u_1 u_2 \cdots u_{q-1} u_q\rangle$ が存在する．よって

$$t_1 \subset u_{j_1}, t_2 \subset u_{j_2}, \cdots, t_n \subset u_{j_n}, t' \subset u_{j_{n+1}}$$

を満たす整数 $1 \leq j_1 < j_2 < \cdots < j_n < j_{n+1} \leq q$ が存在する．β の末尾要素 u_q は s' の部分集合で，かつ $S \circ s'$ 上で頻出な s' の部分集合の中で極大でなくてならない．さもなくば明らかに β の極大性に矛盾する．よって δ の定義から，任意の $k = 1, \ldots, m$ と $i = 1, \ldots, q-1$ に対して 以下が成り立つ．

$$\delta(\langle s_k \cdots s_m\ s' \rangle, \langle u_i u_{i+1} \cdots u_{q-1} u_q \rangle) = \delta(\langle s_k \cdots s_m \rangle,\ \langle u_i u_{i+1} \cdots u_{q-1} \rangle)$$

β の頻出性から $T\text{-}freq(S \circ s', \langle u_1 \cdots u_q \rangle) \geq MS$ が成り立つので，上記の [頻出性の証明] での議論と同様にして，$T\text{-}freq(S, \langle u_1 \cdots u_{q-1} \rangle) \geq MS$ が成り立つ．即ち系列 $\langle u_1 \cdots u_{q-1} \rangle$ は S 上の頻出系列となる．

このとき，$t' = u_q$ であれば，$\langle u_1 \cdots u_{q-1} u_q \rangle$ が $\langle t_1 \cdots t_n\ t' \rangle$ を真に含むことから，$\langle u_1 \cdots u_{q-1} \rangle$ が $\langle t_1 \cdots t_n \rangle$ を部分系列として真に含まなくてならない．これは $\alpha = \langle t_1 \cdots t_n \rangle$ が S 上の極大頻出系列であることに矛盾する．また $t' \neq u_q$ の場合，$\langle u_1 \cdots u_{q-1} \rangle$ が $\langle t_1 \cdots t_n\ t' \rangle$ を含むので，結局 $\langle u_1 \cdots u_{q-1} \rangle$ が $\langle t_1 \cdots t_n \rangle$ を部分系列として真に含む．これも α の極大性に矛盾する．以上より $\alpha \circ t'$ の極大性の証明が完了した． ∎

以下，定理1で示されている極大頻出系列の生成手順を逐次的に繰り返すアルゴリズム，即ちオンライン型アルゴリズムを示す．以下では，S の末尾に新たなアイテム集合を順に追加し，直前の極大頻出集合を利用し，漸近的に極大頻出系列の集合を更新していく．

定義 5 各 $F_i (1 \leq i \leq m)$ をアイテム集合の集合とする時，その直積 $\mathcal{F} = F_1 \times F_2 \times \cdots \times F_m$ により，長さ m のアイテム集合の系列の集合を表すものとする．即ち

$$\mathcal{F} = \{\langle s_1 s_2 \cdots s_m \rangle \mid s_1 \in F_1, s_2 \in F_2, \cdots, s_m \in F_m\}$$

[オンライン型極大頻出系列抽出アルゴリズム]

初期入力として最小サポート値 MS を与える．逐次入力として，現在までの単一系列データベース $S_n = \langle s_1 s_2 \cdots s_n \rangle$ と，S_n の末尾を拡大するアイテム集合 s'，ならびに S_n に出現する各アイテム集合 t の $Fr(S_n, t)$ を記録した作業表 FR_n と S_n 上の極大頻出系列の集合 $\mathcal{F}_n = F_1 \times \cdots \times F_{h_n}$ を与える．

逐次出力は，末尾拡大されたデータベース $S_{n+1} = \langle s_1 s_2 \cdots s_n s' \rangle$ と，S_{n+1} 中の極大頻出系列の集合 $\mathcal{F}_{n+1}$ と，更新された作業表 FR_{n+1} である．

データベースの初期値 S_0 は空列 λ，作業表の初期値 FR_0 は，任意のア イテム集合 t に対して $Fr(S_0, t) = 0$ なる $Fr(S_0, t)$ を集めた表，$\mathcal{F}_0$ は空集合とする．このとき

1. S_n を末尾拡大し，$S_{n+1} = \langle s_1 s_2 \cdots s_n s' \rangle$ とする．
2. 各アイテム集合 $t \subset s'$ に対して，以下のように $Fr(S_{n+1}, t)$ を定め 表 FR_{n+1} を作成する．
$$Fr(S_{n+1}, t) = \begin{cases} Fr(S_n, t) + 1 & \text{if } t \subset s' \\ Fr(S_n, t) & \text{otherwise} \end{cases}$$
3. 新しく出現した頻出アイテム集合の集合 C を求める．
$$C = \left\{ t \subset s' \;\middle|\; \begin{array}{l} Fr(S_{n+1}, t) > Fr(S_n, t) \text{ かつ} \\ Fr(S_{n+1}, t) \geq MS \end{array} \right\}$$
4. $C \neq \emptyset$ と $C = \emptyset$ の2つの場合を考える．
   - $C \neq \emptyset$ の場合：C 中の極大な元の集合 C' を求める．即ち $C' = \{ t \in C \mid \neg \exists u \in C\ (t \subset u \text{ かつ } t \neq u)\}$ とする．S_n 上の極大頻出系列の集合 $\mathcal{F}_n = F_1 \times \cdots \times F_{h_n}$ を拡大して，$\mathcal{F}_{n+1} = F_1 \times \cdots \times F_{h_n} \times C'$ を作成する．
   - $C = \emptyset$ の場合：$\mathcal{F}_n$ をそのまま $\mathcal{F}_{n+1}$ とする．

以上のアルゴリズムについて以下が成り立つ．

定理 2 単一系列データベースを S とし，S 上の極大頻出系列の集合を $\mathcal{F}_n$ とする．この時，S の末尾を任意のアイテムの集合 s' で拡大した系列を $S \circ s'$ とすると，アルゴリズムの出力結果である $\mathcal{F}_{n+1}$ は，$S \circ s'$ から得られる極大頻出系列全ての集合である．

証明： 定理1からほぼ自明である． ∎

例 1 アルゴリズムの動作の流れを例を用いて以下に示す．最終的に構成される単一系列データベースは，$S = \langle (ab)c(abc)(abc) \rangle$ とし，S の要素が順次届き，漸近的に構成されていく状況を考える．最小サポート値は $MS = 2$ とする．以下では各ステップでの主な値を列記する．

Step 1: 入力：$S_0 = \lambda$, $s' = (ab)$, $FR_0 =$ 空表, $\mathcal{F}_0 = \emptyset$

1. $S_1 = \langle (ab) \rangle$
2. $(a), (b), (ab)$ の出現頻度値を1増やし，FR_1 を作成．

FR_1:

(a)	(b)	(ab)
1	1	1

3. $C = \emptyset$
4. $\mathcal{F}_1 = \mathcal{F}_0$

出力：$S_1 = \langle (ab) \rangle$, FR_1, $\mathcal{F}_1 = \emptyset$

Step 2: 入力：$S_1 = \langle (ab) \rangle$, $s' = (c)$, FR_1, $\mathcal{F}_1 = \emptyset$

1. $S_2 = \langle (ab)c \rangle$
2. (c) の出現頻度値を1増やし，FR_2 を作成．

FR_2:

(a)	(b)	(c)	(ab)
1	1	1	1

3. $C = \emptyset$
4. $\mathcal{F}_2 = \mathcal{F}_1$

出力：$S_2 = \langle (ab)c \rangle$, FR_2, $\mathcal{F}_2 = \emptyset$.

Step 3: 入力：$S_2 = \langle (ab)c \rangle$, $s' = (abc)$, FR_2, $\mathcal{F}_2 = \emptyset$

1. $S_3 = \langle (ab)c(abc) \rangle$
2. $(a), (b), (c), (ab), (ac), (bc), (abc)$ の出現頻度値を1増やし，FR_3 を作成．

FR_3:

(a)	(b)	(c)	(ab)	(ac)	(bc)	(abc)
2	2	2	2	1	1	1

3. $C = \{(a), (b), (c), (ab)\}$

4. $\mathcal{F}_3 = \left\{ \begin{array}{c} (c) \\ (ab) \end{array} \right\}$

出力：$S_3 = \langle (ab)c(abc) \rangle$, FR_3, $\mathcal{F}_3$

Step 4: 入力：$S_3 = \langle (ab)c(abc) \rangle$, $s' = (abc)$, FR_3, $\mathcal{F}_3$

1. $S_4 = \langle (ab)c(abc)(abc) \rangle$
2. (a), (b), (c), (ab), (ac), (bc), (abc) の出現頻度値を 1 増やし，FR_4 を作成.

FR_4:

(a)	(b)	(c)	(ab)	(ac)	(bc)	(abc)
3	3	3	3	2	2	2

3. $C = \{(a), (b), (c), (ab), (ac), (bc), (abc)\}$
4. $\mathcal{F}_4 = \left\{ \begin{array}{c} (c) \\ (ab) \end{array} \right\} \times \left\{ (abc) \right\}$

出力：$S_4 = \langle (ab)c(abc)(abc) \rangle$, FR_4, $\mathcal{F}_4$

5. 評価実験と考察

オンライン型極大頻出系列抽出アルゴリズムをC言語で実装し，単一系列データベース S の系列長 n とアイテム集合の大きさ m を変化させ，データベースの拡張に伴う実行時間の変化を調べた．データベース系列は $S = \langle (i_{11}i_{12}\cdots i_{1m})\ (i_{21}i_{22}\cdots i_{2m})\ \cdots (i_{n1}i_{n2}\cdots i_{nm}) \rangle$ という形式であり，各アイテム i には 1〜100 の範囲の数値を乱数で生成し割り当てた．実験環境は OS: Turbo Linux 8 Workstation，CPU: Pentium4 2.53GHz，メモリ：2GB である．

- **系列長 n のスケーラビリティ**
  実験結果のグラフを図 1 に示す．実験条件はアイテム集合の大きさ m を 10，最小サポート値 MS は，系列長 n の 25% とした．
- **アイテム集合の大きさ m のスケーラビリティ**
  実験結果のグラフを図 2 に示す．実験条件は系列長 n を 100，最小サポート値 MS は 25 とした．

図 1 より，n の増加に対して，実行時間が一次関数的に増加していることが分かる．また図 2 より，m の増加に対しては，実行時間が指数関数的に増加している．これらのことから本手法は，単一系列データベースの系列長の拡張性に優れているが，系列中のアイテム集合の大きさに対して弱点をもっていることがわかる．

6. まとめ

本論文では，先行研究 [6, 3] で提案された逆単調性を満たす全体頻度なる出現尺度を用いて，大規模時系列データを対象とした頻出パターンのオンライン型高速抽出アルゴリズムの提案を行った．またアルゴリズムに関する正当性の証明を行い，評価実験の結果から本アルゴリズムの有用性を示した．

今後の課題としては，まずアイテム集合の大きさの拡張性に対する改善が挙げられる．また系列データベース中にデータ有効期間を設定し，有効期間の移動に伴う極大頻出系列の変化をオンラインで高速抽出する問題は実用上重要であるが，これに関しては別の機会に発表したい．また先行研究 [3] でのウィンドウ機構を用い，頻出系列の最初と最後の要素の間の出現期間を一定時間以下に制限する問題も応用上重要である．本論文の手法をこの問題に緩和手法として応用することで，解の高速抽出を助ける可能性があると考えられる．

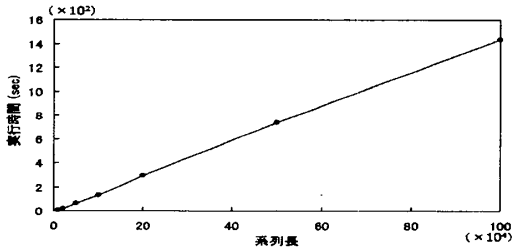

図 1: 系列長 n のスケーラビリティ

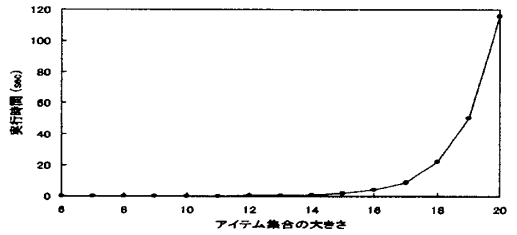

図 2: アイテム集合の大きさ m のスケーラビリティ

謝辞 本研究は一部，文科省科学研究費補助金 (No.16500078) ならびに中部電力基礎技術研究所研究助成の援助を受けている．

参考文献

[1] R. Agrawal and R. Srikant. Mining sequential patterns: Generalizations and performance improvements. In *Proc. of ICDE-96*, pp.3-17, 1996.

[2] 有村博紀，喜田拓也: データストリームのためのマイニング技術. 情報処理，Vol46, No.1, pp.4-11, 2005

[3] Koji Iwanuma, Yo Takano, Hidetomo Nabeshima. On Anti-Monotone Frequency Measures for Extracting Sequential Patterns from a Single Very-Long Data Sequence. *Proc. of IEEE CIS2004*, No. WP6.5 (2004)

[4] H. Mannila and H. Toivonen. *Knowledge Discovery in Databases: The Search for Frequent Patterns*, 1998, URL://www.cs.helsinki.fi/u/htoivone/teaching/timuS02/b.ps

[5] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M.-C. Hsu. PrefixSpan: Mining sequential patterns efficiently by prefix projected pattern growth. In *Proceedings of ICDE-01*, pp.215-226, 2001.

[6] 高野 洋, 岩沼 宏治, 鍋島 英知. 単一の長大なデータ系列上の系列パターンの出現尺度とその逆単調性. FIT2004（第 3 回情報科学技術レターズ），LE-012, pp.115-118, 2004.

[7] J. Yang, W. Wang and P. S. Yu. Mining Asynchronous Periodic Patterns in Time Series Data. In *Proceedings of 6th ACM SIGKDD*, pp.275-279, 2000.