LF-012

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
A Frequency Measure of Sequential Patterns on a Single Very-Large Data Sequence and its Anti-Monotonicity

高野 洋 † 岩沼 宏治 ‡ 鍋島 英知 ‡
Yo Takano Koji Iwanuma Hidetomo Nabeshima

1. はじめに

本論文では，単一系列データにおける頻出パターンの高速な自動抽出を目的として，逆単調性を満たす出現尺度を与え，その理論的な性質を考察する．また大規模テキスト時系列データに対して予備的な評価実験を行なった結果，興味深い系列が得られたので，結果の一部を示す．

時系列に代表される系列データを対象とするデータマイニングは，数多くの応用を持ち，近年盛んに研究されている [1, 2, 4, 7, 12, 13]．POSデータや金融情報，ゲノム，計算機ログの解析などの従来からの研究分野に加えて，近年では，時系列WEBデータの時間追跡 [5, 6, 9, 11] や情報検索の分野における続報記事抽出 [3] などの時系列テキストの研究も盛んに行なわれている．

系列データ上の頻出パターンの自動抽出に関しては，Agrawal らの系列パターンマイニング [1, 2, 13] に関する研究が代表的である．Agrawal らの手法では，複数のデータ系列の集合をデータベース D として捉え，複数の系列に出現する系列パターンを抽出する．この枠組みでは，目標とする系列パターン α を含むデータベース D 中の系列の数を調べる．D の各系列における α の複数回の出現は完全に無視され，数え上げられることは全く無い．これに対して，本論文では，単一の長大な系列データからの頻出パターンの高速な自動抽出法を考察する．高速な自動抽出のためには，逆単調性を満たすパターンの出現尺度が重要であるが，単一系列上で逆単調を満たす合理的な出現尺度を構成することは意外と困難である．本論文では，逆単調性を満たす**系列全体頻度**なる出現尺度を提案し，理論的な考察を行なう．また併せて予備的な評価実験の結果も示す．

文献 [4, 7, 12] でも，単一の系列データ上の頻出パターンの高速自動抽出が研究されている．Yang ら [12] では，周期的に出現する系列パターンの高速自動抽出を試みているが，逆単調性を満たす出現尺度は全く考察されていない．Mannila ら [7] では，スライド窓 (sliding window) を使った系列パターンの高速自動抽出を試みている．スライド窓を利用した出現尺度は逆単調性を満たす [7, 10] が，同一のデータを重複して数え上げることが知られており [10]，短い系列パターンほど重複して数え上げる欠点を持つ．そのため長い系列パターンは必要以上に不利な扱いを受け，自動抽出が著しく困難となる．本論文で提案する系列全体頻度はスライド窓機構を利用しておらず，重複数え上げが無い尺度となって いる．

†山梨大学大学院工学研究科、現在 (株) 日本システムディベロップメント勤務
‡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

2. 準備

表記法は文献 [8] に準拠する．全ての**アイテム**の集合を $I = \{i_1, i_2, \cdots, i_n\}$ とする．**系列**とはアイテムの順序付きリストである．系列 α を $\langle s_1 s_2 \cdots s_l \rangle$ で表すとき，各 s_j は α の**要素**と呼ばれる．系列 $\alpha = \langle s_1 s_2 \cdots s_n \rangle$ と $\beta = \langle t_1 t_2 \cdots t_m \rangle$ について，α が β の**部分系列**であるとは，$s_1 = t_{j_1}, s_2 = t_{j_2}, \cdots, s_n = t_{j_n}$ を満たす整数 $1 \leq j_1 < j_2 < \cdots < j_n \leq m$ が存在する場合をいう．これを $\alpha \sqsubseteq \beta$ と表す．

Agrawal の系列パターンマイニングにおいては，系列データベース D は系列の集合であり，頻出系列パターンとは，複数の D 中の系列に出現する部分系列のことである．この枠組みでは，目標とした α を含む D 中の系列の数が重要であり，D 中の各系列における α の出現回数は完全に無視されている．本論文では，単一の長大なデータ系列からなるデータベース S を考え，S 中に繰り返し出現する部分系列パターンの高速抽出法を提案する．

3. 系列先頭頻度と系列全体頻度

部分系列の頻出の尺度に対する**逆単調性 (anti-monotone)** は，頻出パターンの効率的なデータマイニングにとって，非常に重要な役割を担っている．M を系列から非負実数への写像とする．任意の 2 つの系列 α, β について，$\alpha \sqsubseteq \beta$ ならば $M(\alpha) \geq M(\beta)$ となるとき，M が**逆単調**であるという．本稿では，単一の長大な系列 S 中に複数回出現するパターンのマイニングのために有用な頻度の尺度の構築を試みる．ただし素朴な頻度の尺度のほとんどは，残念ながら逆単調ではない．

例 1 単一系列のデータベース $S = \langle aabbb \rangle$ を考える．S において部分系列 $\langle a \rangle$ が 2 回出現し，$\langle b \rangle$ が 3 回出現することは明らかである．一方，$\langle ab \rangle$ は 6 回出現する．従って，単純な頻度の尺度（すなわち S 中での部分系列の出現数）は，逆単調性を満たさない．さらに，部分系列の出現の割合に注目しても，それもまた逆単調ではない．例えば，S 中の系列 α に対する次のような割合 $R(S, \alpha)$ を考えてみる：

$$\frac{S \text{ 中での } \alpha \text{ の実際の出現数}}{S \text{ 中での } \alpha \text{ の可能な出現の総数}}$$

この例では，$R(S, \langle a \rangle) = 0.4 \ (= \frac{2}{{}_5C_1})$ であり，$R(S, \langle ab \rangle) = 0.6 \ (= \frac{6}{{}_5C_2})$ である．従って，$R(S, \alpha)$ は一般に逆単調ではない．

例 2 単一系列のデータベース $S = \langle abcde \rangle$ を考える．長さ 3 の窓（スライド窓 sliding window [7]）を通して S

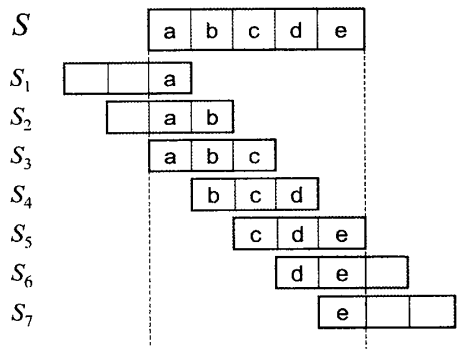

図 1: スライド窓の利用

を眺めることで，図1のように S を部分系列 $S_1, \cdots, S_7$ へと分割する．各 S_i を **窓**と呼び，部分系列 α の頻度 $W(S,\alpha)$ を，α を含む窓の数で与える．$W(S,\langle cd\rangle) = 2$ であり $W(S,\langle cde\rangle) = 1$ である．頻度 $W(S,\alpha)$ が逆単調性を満たすことは明らかである．しかし各窓が同一のデータを重複して保持するため，要素の間隔が短い系列ほど頻度が高くなる傾向がある．例えば，$W(S,\langle de\rangle) = 2$ であるが，より長い間隔の系列 $\langle ce\rangle$ においては $W(S,\langle ce\rangle) = 1$ となる．また長い系列パターンも必要以上に不利な扱いを受ける．この例では各要素はそれぞれ1回しか出現しないにもかかわらず，$W(S,\langle cde\rangle) = 1, W(S,\langle c\rangle) = 3$ である．従って $W(S,\alpha)$ は適切な頻度の尺度とはいえない [10].

以後，我々は逆単調な頻度の尺度を段階的に構築する．系列 $\alpha = \langle t_1 t_2 \cdots t_n\rangle$ が与えられたとき，$\boldsymbol{suf}(\alpha, i)$ により，α の接尾辞 $\langle t_i t_{i+1} \cdots t_n\rangle$ を表す．

定義 1 （系列先頭頻度）
S を単一系列のデータベース $\langle s_1 s_2 \cdots s_m\rangle$ 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** $\boldsymbol{H\text{-}freq}(S,\alpha)$ 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 を以下の関数と定める．

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

系列先頭頻度は，文字通り，系列の先頭要素の出現頻度を計るものである．

例 3 例1と同じ系列 $S = \langle aabbb\rangle$ について，

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

S の 2 つの接尾辞，すなわち $suf(S,1) = S$ と $suf(S,2) = \langle abbb\rangle$ の 2 つの接尾辞だけが部分系列 $\langle ab\rangle$ を含む点に注意して頂きたい．従って，$\langle a\rangle$ と $\langle ab\rangle$ のペアに対しては，$H\text{-}freq(S,\langle a\rangle) \geq H\text{-}freq(S,\langle ab\rangle)$ となり，先頭系列頻度は逆単調性を満たす．しかし残念ながら，一般には系列先頭頻度は逆単調性を満さない．例えば S' を $\langle aaaab\rangle$ とした場合，$H\text{-}freq(S',\langle b\rangle) = 1$ かつ $H\text{-}freq(S',\langle ab\rangle) = 4$ であり，

$$H\text{-}freq(S',\langle b\rangle) \not\geq H\text{-}freq(S',\langle ab\rangle)$$

となる．明らかに逆単調性が成り立たない．

系列先頭頻度は逆単調性は満たさないが，より限定された性質である**右逆単調性**は満足する．α を系列 $\langle s_1 s_2 \cdots s_m\rangle$ とし，$t \in I$ をアイテム，i を $1 \leq i \leq m$ を満たす整数とする．このとき，α **の位置** i **における** t **を用いた右拡張**を系列 $\langle s_1 \cdots s_i\ t\ s_{i+1} \cdots s_m\rangle$ と定める．また系列 $\langle t\ s_1 \cdots s_m\rangle$ を，α **の** t **を用いた先頭拡張**と呼ぶ．右拡張と先頭拡張を併せて**拡張**と呼ぶ．

定義 2 （右逆単調性）
M を系列から非負実数への写像とする．M **が右逆単調であるとは**，任意の系列 α, β について，β があるアイテムを使った α の右拡張であるときに，$M(\alpha) \geq M(\beta)$ である場合をいう．

補題 1 系列先頭頻度 $H\text{-}freq$ は右逆単調である．すなわち，任意の単一系列から成るデータベース S と，任意の 2 つの系列 α と β について，β が α の右拡張ならば，以下が成り立つ．

$$H\text{-}freq(S,\alpha) \geq H\text{-}freq(S,\beta)$$

証明 定義から容易に証明できる．紙面の都合上，省略する． ∎

$H\text{-}freq(S,\alpha)$ は，系列 α の先頭要素の出現頻度を計るが，その後続要素の出現は，その有無だけをチェックし，実際の出現回数は見ていない．例えば，系列データベース $S = \langle aabbbc\rangle$ に対して，

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

となり，S における要素 b と c の出現回数の違いは反映されない．直感的には，S における $\langle ab\rangle$ と $\langle ac\rangle$ の出現頻度には明らかな違いがあり，これは望ましくない．系列の後続要素の出現頻度を何らかの形で反映する尺度が必要である．そこで次に，系列の全ての部分系列の頻度を考慮した尺度を考える．

定義 3 （系列全体頻度）
S を単一の系列データベースとする．任意の系列 α に対し，γ を α の任意の部分系列とする．S **における** α **の系列全体頻度** $\boldsymbol{T\text{-}freq}(S,\alpha)$ を以下で定義する．

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

例 4 単一系列のデータベース $S = \langle aabbbc \rangle$ を考える．このとき

$$\begin{aligned} &T\text{-}freq(S, \langle ab \rangle) \\ &= \min(H\text{-}freq(S, \langle ab \rangle), H\text{-}freq(S, \langle a \rangle), H\text{-}freq(S, \langle b \rangle)) \\ &= \min(2, 2, 3) \\ &= 2 \end{aligned}$$

また同様にして，$T\text{-}freq(S, \langle ac \rangle) = 1$ である．

次の定理により，系列全体頻度が逆単調性を満たすことを示す．

定理 1 系列全体頻度は逆単調である．すなわち，任意の単一系列から成るデータベース S と，任意の 2 つの系列 α と β について，$\beta \sqsubseteq \alpha$ ならば

$$T\text{-}freq(S, \beta) \geq T\text{-}freq(S, \alpha) \qquad (1)$$

証明 α のすべての部分系列の集合を Γ_α とし，β の全ての部分系列の集合を Γ_β とする．$\beta \sqsubseteq \alpha$ であるので，$\Gamma_\beta \subseteq \Gamma_\alpha$ である．よって，系列全体頻度の定義より，明らかに式 (1) が成り立つ．∎

定義 3 より，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が，実際には，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．これを次の補題で示す．

補題 2 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

$$T\text{-}freq(S, \alpha) = \min_{i=1}^{n}(H\text{-}freq(S, suf(\alpha, i))) \qquad (2)$$

証明 α の任意の部分系列 $\beta = \langle u_1 u_2 \cdots u_m \rangle$ を考える．このとき，

$$u_1 = t_{j_1},\ u_2 = t_{j_2},\ \cdots,\ u_m = t_{j_m}$$

を満たす整数 $1 \leq j_1 \leq j_2 \leq \cdots \leq j_m \leq n$ が存在する．ここで α の接尾辞 $suf(\alpha, j_1)$ は，β の右拡張（を複数回繰り返したもの）となっている．よって補題 1 から，明らかに $H\text{-}freq(S, \beta) \geq H\text{-}freq(S, suf(\alpha, j_1))$ である．以上から α の部分系列 β には必ず $H\text{-}freq(S, \beta) \geq H\text{-}freq(S, suf(\alpha, k))$ を満たす接尾辞 $suf(\alpha, k)$ が存在する．よって，補題 2 は明らかである．∎

更に補題 2 から以下の系が成り立つ．この系は，α の系列全体頻度が，α より 1 つ短い接尾辞 $suf(\alpha, 2)$ の系列全体頻度と，α 自身の系列先頭頻度から求まることを示している．よって単一系列データベース中の頻出部分系列を，長さの短い系列から順次求めていくならば，動的プログラミング法により，系列全体頻度の計算は更に効率化できる．即ち，長さ $n-1$ の頻出系列の全体頻度を表に保存しておけば，長さ n の系列 α の系列全体頻度は α の先頭頻度（だけ）を求めて表中の値と比較することにより，容易に求めることができる．

系 1 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

$$\begin{aligned} &T\text{-}freq(S, \alpha) = \\ &\quad \min(H\text{-}freq(S, \alpha),\ T\text{-}freq(S, suf(\alpha, 2))) \end{aligned} \qquad (3)$$

証明 補題 2 より，

$$T\text{-}freq(S, suf(\alpha, 2)) = \min_{i=2}^{n}(T\text{-}freq(S, suf(\alpha, i)))$$

であるので，これを式 (3) に代入すると，式 (2) の右辺に等しくなる．∎

4. 大規模時系列テキストを用いた頻出パターン抽出実験

系列全体頻度 $T\text{-}freq(S, \alpha)$ は完全に逆単調性をみたすので，系列データのマイニングに既存手法 [1, 2, 8] を利用できる．我々は Agrawal [1] の AprioriAll 法を転用して，毎日新聞コーパスから重要イベント（単語）の頻出時系列パターンの抽出予備実験を試み，系列全頻度の有用性の評価を行なった．コーパス中の記事は日付属性を持っており，一日を単位とする時系列上に並べられる．記事中のイベントを表す重要語フレーズの選定法や抽出アルゴリズムなどの詳細は，紙面が限られているので，本稿では省略する．図 2 に，毎日新聞 2000 年上半期の社会面の記事 12,596 記事を対して行なった予備実験結果を示す．

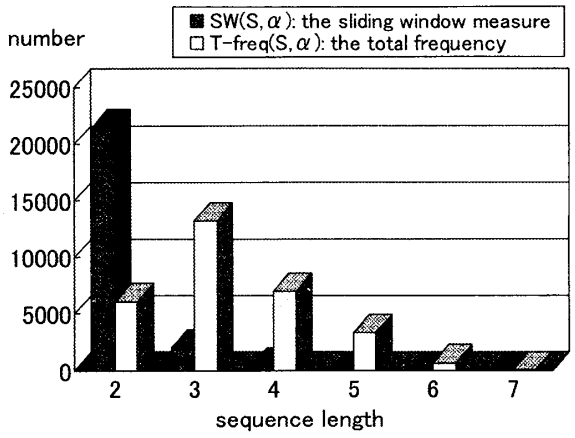

図 2: 毎日新聞コーパスから抽出された頻出系列の数

図 2 は頻出イベント系列の分布を示している．頻出尺度 $SW(S, \alpha)$ は，長さ 7 のスライド窓§を用いた尺度である．最小出現頻度はマイニング期間の日数の 15%（27 回）以上とした．ただし系列全体頻度においては，頻度 27 回以上の系列はほとんど得られなかったため，最小

§紙面の都合上省略したが，系列全体頻度でも**イベント期間長** と呼ばれるウインドウを使用して，連続する一連のイベントの発生期間を限定している．本実験でのイベント期間長は 7 である．

出現頻度を15回として実験を行った．このことは，スライド窓尺度が短い系列パターンを重複して数え上げていることを示している．図より明らかなように，系列全体頻度 $T\text{-}freq(S,\alpha)$ は，スライド窓尺度 $SW(S,\alpha)$ よりも，より小さな最小出現頻度において，より長い頻出系列の抽出に成功している．以下に，抽出に成功した幾つかのイベント系列の一部を示す．

- ⟨(放火)(判決)⟩, ⟨(判決)(弁護士)⟩, ⟨(地震)(転落)⟩, ⟨(子供)(文部省)⟩, ⟨(トラック)(爆発)⟩, ⟨(リコール)(車)⟩ ⟨(交際)(被害者)⟩ ...
- ⟨(放火)(起訴)(男児)⟩, ⟨(巡査部長)(放火)(自殺)⟩, ⟨(気象庁)(地震)(アパート)⟩...
- ⟨(巡査部長)(放火)(地震)(大阪)⟩ ...

5. まとめ

本論文では，単一の長大なデータ時系列上の頻出パターンを高速に抽出することを目的として，逆単調性を満たす出現頻度について議論した．また系列全体頻度を提案し，逆単調性を満たすこと，およびその理論的性質を示した．また新聞記事コーパスを用いた予備的な検証実験によって，重要イベント（単語）系列の抽出に有効に働くことを確認している．我々の知る限り，単一データ系列上の出現パターンの出現尺度の中で，重複数え上げを行なわずに逆単調性を満たすものは，本論文での系列全体頻度の他には無い．

本稿では紙面の都合から，これまでに我々が開発した頻出時系列パターンの高速抽出アルゴリズムや枝刈技術，および実装プログラムの詳細は省略した．今後あらためて報告を行ないたい．また系列先頭頻度の双対として，系列末尾頻度とも呼ぶべき頻度尺度が考えられる．系列末尾頻度をベース尺度とした系列全体頻度も定義でき，逆単調性を同様に満す．概念的には，系列先頭頻度がイベント系列の起点または原因の出現を数えるのに対して，系列末尾頻度はイベントの終点または帰結・結果を数えることに相当する，と考えられる．今後のこの２つの系列尺度の現実のデータ系列への適用性やその特徴や差異についても、考察を行なっていく予定である．

謝辞 日頃より多くの有益な議論を頂く福本文代助教授（山梨大学大学院）に深く感謝致します．また本研究は一部，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. In *Proceedings of ICDE-95*, pp.3–14, 1995.

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

[3] 福本文代, 鈴木良弥, 山田寛康. 話題の推移に基づく続報記事の自動抽出. 情報処理学会論文誌, 44(7) pp.1766–1777, 2003.

[4] K-Y. Huang and C-H. Chang. Asynchronous Periodic Patterns Mining in Temporal Databases In *Proceedings of IASTED Inter. Conf. on Databases and Applications*, pp.43-48, 2004.

[5] S. Hirokawa, E. Itoh and T. Miyahara. Semi-Automatic Construction of MetaData from A Series of Web Documents. In *Proceedings of 16th Australian Joint Conf. on Artificial Intelligence*, LNCS 2903, pp.942-953 (2003).

[6] M. Nakamura, K. Iwanuma and H. Nabeshima. Detecting Two Sorts of Correspondences between HTML Documents for Extracting Temporal Differences. In *Proceedings of the Third IASTED International Conference on Artficial Intelligence and Applications (AIA2003)*, pp.611-616 (2003).

[7] 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

[8] 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.

[9] 高間 康史. Web 情報ストリーム. 情報処理学会学会誌, 44(7) pp.720–725, 2003.

[10] 高野洋. 大規模な時系列テキストデータからのイベント時系列パターンの発見. 山梨大学大学院修士論文, 2004.

[11] 梅原雅之, 岩沼宏治, 鍋島英知. 事例に基づくシリーズ型 HTML 文書の意味論理構造の自動認識, 人工知能学会論文誌, 17(6E) pp.690-698. (2002).

[12] 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.

[13] M. J. Zaki: Efficient Enumeration of Frequent Sequences. In *Proc. the 7th Inter. Conf. on Information and Knowledge Management (CIKM'98), pp.68–75 (1998)*