単一系列データ上の系列選択パターンと逆単調性

清水　一宏†　　三浦　孝夫†

† 法政大学 工学部 情報電気電子工学科　〒 184–8584 東京都小金井市梶野町 3-7-2
E-mail: †{c02d3051,miurat}@k.hosei.ac.jp

あらまし　本論文では，単一系列データ上での系列選択の頻出パターン発見手法を論じる．このため，高速な自動抽出方法を提案して逆単調性を満たす出現尺度を与え，APRIORI 流の計算によりその有用性を検証する．
キーワード　系列選択パターン，逆単調性, 単一系列データ

Disjunctive Sequential Patterns on Single Data Sequence and its Anti-Monotonicity

Kazuhiro SHIMIZU† and Takao MIURA†

† Dept.of Elect.& Elect. Engr., HOSEI University　3-7-2, KajinoCho, Koganei, Tokyo, 184–8584 Japan
E-mail: †{c02d3051,miurat}@k.hosei.ac.jp

Abstract　In this work, we proposes a novel method for mining frequent disjuctive patterns on single data sequence. For this purpose, we introduce a sophisticated measure that satisfies *anti-monotonicity*, by which we can discuss efficient mining algorithm based on APRIORI. We discuss some experimental results.
Key words　Disjunctive Sequence Pattern, Anti-Monotonicity, Single Data Sequence

1. 前　書　き

近年, 文書データ分析にデータ発見手法を使用することが多くなってきた．従来，文書データは定性的であり量的に判断することが無かった．このため，統計的推定やデータ発見手法など定量的な分析に基づく分析は難しい [10], [13]．しかし，テキストの特徴を数量化し解析しようとする研究（テキスト発見）が注目されるにつれ，従来提案されてきた手法についても適用可能であると考え出されている．

データ発見手法の特徴は，頻度 (重要な項目は何度も生じる) や共起性 (複数の項目が同時に生じる) にある [2], [3]．例えば，書店の購買情報を分析することで,「”蛇にピアス”を購入する客は同時に”蹴りたい背中”も購入する」といった相関性(注 1)の検出することができ，書籍の配置やキャンペーン方法を工夫するきっかけとなる．

この考え方は多方面に拡張できる．系列データ (sequence data) とは情報を時系列などのある順序で並べたものであり，購買パターンを動作として検出することができる．例えば，書店の購買情報を分析することで,「”蛇にピアス”を購入する客は 1 ヶ月以内に”蹴りたい背中”も購入する」といった相関性検出により，DM など販売促進の方法を示唆することができる．この他，Web アクセスパターンの解析，医療診断，DNA 系列の解析など幅広い応用が考えられる [4], [12], [14].

このうちテキスト発見は，近年注目を浴びている研究分野である．テキストは語の並びで構成された意味を表し，系列データの一種である．テキスト発見手法の特徴は，頻度 (重要な語は何度も生じる) や共起性 (類似した語は同時に近くに表れる) にある．頻度の高いパターン (語の並び，フレーズ) を抽出し，テキストの要約 (抽象化) や重要な事象のラベル付けを行うことができる．

長大なテキストから発生頻度の高い語句すべてを発見することは容易ではない．探索空間が巨大であり組合わせ的な探索問題となっている．例えば，次の例を考える．

(1) ”友人の兄の母に会った．今日，兄の友人の母に会った”
(2) ”赤くて大きい旗を見た．今日，大きくて赤い旗を見た”

明らかに，(1) で論じているのは異なる人物であるが，(2) で論じているのは同じ旗である．頻出パターンを発見するとき，後者の場合では複数回カウントすべきである．テキスト発見技法ではこれを ”[赤い，大きい] 旗” というパターンとしてを検出したい．すなわち，”赤い旗” でも”大きい旗” ではなく，”赤

(注 1)：”蛇にピアス”金原 (かねはら) ひとみ著と”蹴りたい背中” 綿矢 (わたや) りさ著は 2003 年第 130 回芥川賞を同時に受賞した文学作品であり，両著者が 20 才前後であったことから注目を浴びた．

い”と”大きい”の語順 (permutation) を無視したパターンとして捕らえたい．本研究では，これを**選択** *(disjunctive)* **パターン**という．Kleene 閉包を加え正規表現を考察できるが，計算処理量が増大し，本稿では議論しない．選択パターンを発見するとき，膨大な数の候補を生成すべきであり，そのひとつが ”蛇にピアス 蹴りたい背中”である．候補を絞って収束させるには何度もデータベースを走査する必要があり，ハードディスク上に格納されたデータに数多くのアクセスが生じる．しかし，サンプリング手法や特殊なデータ構造の利用ではデータの分布特性によって大きく性能に差が生じる．

これまで主要な研究は APRIORI に基づいている [10], [13]. このことで探索量を大幅に下げることができるが，系列パターンについては十分ではない．本来，APRIORI アプローチでは，パターン q がパターン p の”部分”であるとき，p に適合する系列データは必ず q にも適合する，という性質 (パターンの逆単調性) により探索量を削減する．ところがテキストでは逆単調性が成り立たず，もはや APRIORI アプローチを利用できない．

例えば テキスト `"aabbba"` においてパターン `"ab"` の適合回数は 6 回, `"a"`, `"b"` は共に 3 回, `"[ab]"` は 9 回である．

単一の長大な系列データ S に対して，パターン p でその出現尺度 $\mathcal{M}_S(p)$ がある整数値 m 以上のものをすべて検出することを論じる．従って，発生回数を意識したカウント方法 $\mathcal{M}$ が問題である．本稿では，既に提案されている手法 [1] を拡張し，選択パターンを APRIORI で取り扱う枠組みの提案を行う．2 章で問題を定式化し，3 章で逆単調性を満たす出現尺度を導入する．続く 4 章でこれを用いたテキスト発見アルゴリズムを述べ，5 章でいくつかの実験結果を示す．

2. 系列データからのパターン発見

本稿では，語 (word) を基本単位 (item) として扱う．アイテム集合 $I = \{i_1, .., i_L\}, L > 0$ の要素をアルファベット，**系列データ** (テキスト) $S = s_1...s_m, m > 0$ とはアイテムの順序リストである．S には同じアイテムが複数回生じてよく，S に含まれる (重複を含めた) 語の数 m に対し，S を m-系列データという．

選択パターン (あるいは単にパターン) p とは $t_1t_2...t_n$ で表され，各 t_i はアルファベット a または選択 $[a_1a_2...a_m], m > 0$ (各 a_j は相異なるアルファベット) である．パターン $p = t_1t_2...t_n$, $q = v_1v_2...v_m, m \leqq n$ があるとき，q が p の**部分パターン** である ($q \sqsubseteq p$ と記す) とは，$1 \leqq j_1 < ... < j_m \leqq n$ があり，各 v_k は t_{j_k} に対応して次を満たす (混乱の無い限り $v_k \sqsubseteq t_{j_k}$ と記す)：

> v_k がアルファベット a のとき， $t_{j_k} = a$ もしくは a を含む選択
> v_k が 選択 $[a_1a_2...a_m]$ のとき，$t_{j_k} = [b_1b_2...b_l]$ でかつ $\{a_1, .., a_m\} \subseteqq \{b_1, .., b_l\}$

［例 1］ ”ac” は ”abcd” の部分パターンである．同様に，”[ac]” は ”[abcd]” の，”bd” は ”[ab]b[cd]de” の，”b” は ”[ab]” の，そして”ac”は ”[ab][cd]” の部分パターンである．

しかし，”ab” は ”[ab]” の部分パターンではない．また”[ab]” は ”ab” の部分パターンではない．

□

系列データ $S = c_1c_2...c_m$ にパターン $p = t_1t_2...t_n$ が**適合する** (match) とは，t_1 がアルファベット a_1 のとき，$t_1 = a_1 = c_{i_1}, 1 \leqq i_1 \leqq m$ でありかつ 部分パターン $t_2...t_n$ が系列データ $c_{i_1+1}...c_m$ に適合するときを言う．t_1 が選択 $[a_1a_2...a_m]$ のとき $a_1, ..., a_m$ のある並びからなるパターン $a_{j_1}...a_{j_m}$ が系列データ $c_1...c_{i_1}$ に適合し，かつ 部分パターン $t_2...t_n$ が系列データ $c_{i_1+1}...c_m$ に適合するときを言う．

［例 2］ 系列データ S が `aabbba` であるとき，パターン `a` は S に 3 回適合する．また，パターン `ab` は 6 回適合する．一方 `[ab]` は 9 回適合する．

□

系列データ S に関してパターンから非負整数への関数 $\mathcal{M}$ が逆単調性 (Anti Monotonicity, AM) を満たすとは，パターン p, q に対して $q \sqsubseteq p$ のとき $\mathcal{M}_S(q) \geqq \mathcal{M}_S(p)$ が成り立つときを言う．以下で考慮する系列データは単一であり S を略す．

逆単調な $\mathcal{M}$ が与えられ，また最小頻度 m が与えられているとする．このとき，$\mathcal{M}(q) < m$ ならば $q \sqsubseteq p$ となる p は $\mathcal{M}(p) \geqq m$ ではない．この性質を用いれば，部分パターンの探索範囲を減少させることができる．これが APRIORI に基づく探索アルゴリズムであり [6]，以下の手順で頻出パターンを計算することができる．

> (1) 「最小のパターン」で頻出のものを探す
> (2) 大きいパターン p で，そのすべての部分パターンが頻出である ものを探す (候補パターン集合を得る). 候補がなくなれば終了する
> (3) S をスキャンして頻出なものだけを求める．(2) へ

単一の長大な系列データ S に対して，選択パターン p でその出現尺度 $\mathcal{M}(p)$ がある整数値 m 以上のものをすべて検出すること，が本論文の目的である．しかし，$\mathcal{M}$ で逆単調なものを見つけるのは簡単ではない．例えばパターン p に対して系列データ S に適合する回数を $\mathcal{M}(p)$ とする．このとき p の部分パターン q で $\mathcal{M}(q) \geqq \mathcal{M}(p)$ を満たさないものがあるのは例 2 に示した．

3. 逆単調出現尺度

単純な適合頻度では逆単調性が得られないため，系列の先頭要素の適合頻度を考察する．

系列データ $S = s_1s_2...s_r$，パターン p を $t_1t_2...t_n$ とする．このとき **系列先頭頻度** $H(S, p)$ を次のように定義する．

$$H(S, p) = \Sigma_{i=1}^{r} Val(S, i, p)$$

ただし $Val(S, i, p)$ は次を満たすとき 1, さもなければ 0 であるとする:

> $S(i)$ を S の i 番目からの接尾辞 $s_i...s_r$ とする．t_1 がアルファベット a のとき，$s_i = a$ かつ $t_2t_3...t_n$ が $S(i+1)$

に適合する．t_1 が選択 $[a_1a_2...a_m]$ のとき，$s_i = a_j$ となる j があり (例えば $j = 1$)，$[a_2a_3...a_m]t_2...t_n$ が $S(i+1)$ に適合する．

$H(S,p)$ は S またはその接尾辞において p が先頭から適合する回数を表している．

［例 3］ (1) S を bbba とするとき，$p = ba$ のとき $H(S,p) = 3$，実際の適合回数は 3 回．$p = a$ のとき $H(S,p) = 1$，実際の適合回数は 1 回である．定義から $a \sqsubseteq ba$ であるが $H(S,a) > H(S,ba)$ ではない．

(2) S を aabbba とする．$p = ab$ のとき $H(S,p) = 2$，実際の適合回数は 6 回．$p = ba$ のとき $H(S,p) = 3$，実際の適合回数は 3 回．$p = [ab]$ のとき $H(S,p) = 5$，実際の適合回数は 9 回 $p = a$ のとき $H(S,p) = 3$，実際の適合回数は 3 回である．定義から $a \sqsubseteq [ab]$ であるが $H(S,a) > H(S,[ab])$ ではない．

□

この例が示すように，系列先頭頻度 $H(S,p)$ は逆単調性を満たさない．p の先頭での適合回数は，後続系列での適合した回数を無視している (そうでなければ逆単調にならない)．そこで p のすべての部分パターン q を調べ，その系列先頭頻度 $H(S,q)$ のうちの最小の値を**全体出現頻度** $D(S,p)$ と呼ぶ．

$$D(S,p) = \mathrm{MIN}\{H(S,q) | q \sqsubseteq p\}$$

［定理 1］ $D(S,p)$ は逆単調性を満たす．

(証明) パターン p, q が $q \sqsubseteq p$ を満たすとき，定義より $D(S,p) = \mathrm{MIN}\{H(S,r) | r \sqsubseteq p\}$，$D(S,q) = \mathrm{MIN}\{H(S,r) | r \sqsubseteq q\}$ であるから，$q \sqsubseteq p$ より $D(S,q) \geqq D(S,p)$ となる．(証明終わり)

［例 4］ (1) S を bbba とするとき，$p = ba$ のとき $D(S,p) = 1$，$p = a$ のとき $D(S,p) = 1$ である．

(2) S を aabbba とする．$p = ab$ のとき $D(S,p) = 2$，$p = ba$ のとき $D(S,p) = 3$，$p = [ab]$ のとき $D(S,p) = 3$，$p = a$ のとき $D(S,p) = 3$ である．

(3) S を caabbbc, $p = ab$ とすれば $H(S,p) = 2, D(S,p) = 2$ である．$p = ac$ とすれば $H(S,p) = 2, D(S,p) = 2$ である．また $p = [ac]$ とすれば $H(S,p) = 3, D(S,p) = 2$ である (実際の適合頻度は 4 回)．ac や $[ac]$ の部分系列 a, c が分離して生じていても適合するとみなすため，系列先頭頻度や適合回数とは合わない．

□

定理 1 から，長さ n のパターン p の全体出現頻度を得るには，p のすべての部分パターンの系列先頭頻度を調べればよい．次の定理から，この計算は p の接尾辞 (n 個存在) のうち，長さの小さいものから順にその最小値を決定すればよいことがわかる．

［定理 2］ 系列データ S, パターンを p とすると，$D(S,p) = \mathrm{MIN}\{H(S,p), D(S,p(2))\}$

(証明) $S = s_1s_2...s_m$, $p = t_1t_2...t_n$, $p(i) = t_i...t_n$ とする．$D(S,p) = \mathrm{MIN}\{H(S,p(i)) | i = 1, ..., n\}$ を言えばよい．

p の任意の部分パターンを $q = u_1u_2...u_k$ とする．仮定より $1 \leqq j_1 < ... < j_k \leqq n$ で $u_i \sqsubseteq t_{j_i}, i = 1, ..., k$ (アルファベットなら同一，選択なら部分集合) となるものがある．

q の位置 i での**右拡張** q' を次のように定義する ($i > 0$):

q' は次のいずれかの形をしている
(i) $u_1u_2...u_ivu_{i+1}...u_k$, つまり u_i の直後にパターン v が挿入されている
(ii) $u_1u_2...u'_i...u_k$, つまり u_i が u'_i に変更され $u_i \sqsubseteq u'_i$ となっている

このとき $H(S,q) \leqq H(S,q')$ が成り立つ．すなわち，$Val(S,i,q') = 1$ ならば必ず $Val(S,i,q)$ をいう．実際，(i) による右拡張なら，適合検査時にこれを無視すれば $Val(S,i,q) = 1$ となる．(ii) による右拡張なら，適合検査時に新たに拡張されたパターンを無視すれば $H(Val,i,q) = 1$ となる．

t_{j_1} は q を何度か右拡張したものなので $H(S,q) \geqq H(S,t_{j_1})$ となる．これより，どの部分パターン q でも系列先頭頻度値でそれより小さい $t_{j'}$ が存在する．(証明終わり)

4. 実　　験

4.1 準　　備

実験データとして，NTCIR-3 特許検索タスクコレクション中の「日本国英語特許出願抄録データ PAJ」を使用する．これは 1995 年から 1999 年までの JAPIO 出願抄録を英語に翻訳したコーパスであり，コーパス中の抄録はそれぞれ特許出願日時順に並んでいる．今回は，1995 年の特許出願抄録データから 1000 件分の特許内容要約文を用い，各要約文について不要語 (stop word) の削除および単語のステミング [5] を行ったものを実験データとして使用する．以下に実験データの一部を示す．

...control adjust speed thresh depth regul grain culm state oper load combin shape initi puls power high load devic regul control thresh depth grain culm detect thresh depth sensor...

4.2 実験の手順と評価

本実験は [6] の APRIORI 法を転用し，実験データから頻出パターンの抽出を試みる．この結果と，[1] の系列全体頻度 $T\text{-}freq(S,p)$ の手法で得られる頻出パターンとを比較し，どちらの手法がより多くの頻出パターンを抽出できるかを調べる．生成される候補パターンはそれぞれ，系列全体頻度 (a,ab,ba,abc,bac...)，全体出現頻度 (a,[ab],[ab]c,[abc]d....) とする．なお，全体出現頻度において [ab]c[de] 等の 1 つのパターン中に複数の選択が出現するものについては対象としない．

4.3 実 験 結 果

最小出現頻度は両手法とも 2%(20 回) 以上とし，全体出現頻度，系列全体頻度それぞれの手法を用いて実験データから頻出パターンを抽出する実験を行う．実験結果を以下の表 1，図 1 に示す．表 1 では，全体出現頻度と系列全体頻度のそれぞれが抽出に成功した頻出パターンの数を示しており，図 1 ではその分布を示している．括弧内の数値は全体出現頻度でのみ抽出が成功したパターン数を示す．また，それぞれの手法において実際に抽出に成功した頻出パターンの一部を以下に示す．

[全体出現頻度]
・([medicin,patient]),([prepar,medicin]),([surfac, sheet])...
・([prepar,medicin]patient)...

[系列全体頻度]

・(medicin,prepar),(medicin,patient),(devic,capabl)...

・(provid,capabl,devic)...

n-頻出パターン	全体出現頻度	系列出現頻度
$n=1$	207(0)	207
$n=2$	121(76)	45
$n=3$	3(1)	2
$n=4$	0(0)	0

表 1　特許出願抄録データから抽出された頻出パターンの数

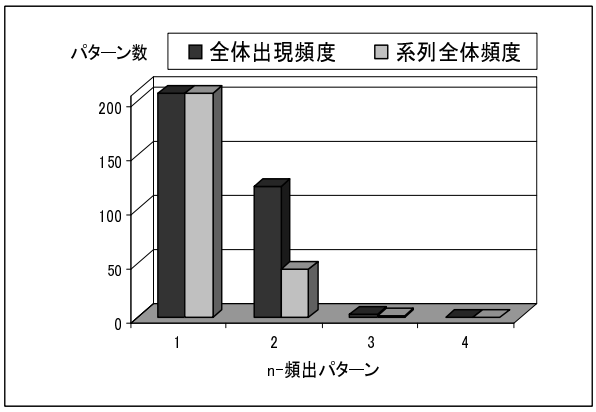

図 1　特許出願抄録データから抽出された頻出パターンの数

4.4 考　　察

表 1 からも明らかなように，$n=2$ および $n=3$ において全体出現頻度は系列全体頻度と比べ，より多くのパターンの抽出に成功している．また，全体出現頻度のみでしか抽出されない頻出パターンが $n=2$ の時 76 組，$n=3$ の時 1 組あり，これは系列全体頻度では抽出されない，最小出現頻度より少ない出現頻度を持つパターンが，全体出現頻度では選択パターンとして抽出されるからだと考えられる．例えば，具体例として上記にも示した ([surfac, sheet] 22 回) や ([prepar,medicin]patient 21 回) は全体出現頻度のみでしか抽出されていない．実験で与えた最小出現頻度は両手法共に 2%と同条件であったので，この点において全体出現頻度は系列全体頻度と比べ優れていると言える．

5. 関 連 研 究

系列パターンを発見する問題 [7], [16] は，これまで相関性を検出する APRIORI アルゴリズム [6] を利用して研究されている．系列データ集合があり，各系列は項目集合のリストから成るとする．このとき最小頻度個以上の系列に出現する系列パターン (これも項目集合のリストとして定義される) をすべて発見する，というのが扱うべき問題である．素直に考えれば組み合わせ的問題であるが，高速化アルゴリズムについては APRIORI に基づく FreeSpan, PrefixSpan の提案 [9], [15] や，束 (Lattice) を用いた形式化による SPADE が知られている [17]．正規表現を直接扱う SPIRIT アルゴリズム [8] や，これを制約充足問題とみなす提案もある [11]．しかし，最小頻度条件を満たすパターンで，しかも正規表現パターンに属するものを発見しようとしており，本研究のように選択系列パターンを発見するものではない．さらに，複数系列を対象としており，出現回数とは系列パターンを含む系列数と定義されているため，テキスト発見に直接利用できない．

(本稿で論じているような) 単一系列データ上で頻出パターンを発見する問題(注 2)は，事象列に関するアプローチ (エピソード) [12] が知られているが，テキスト発見に特化したものではない．本研究と直接関連を有する提案がある [1]．しかし選択系列を扱わず，本稿で扱う問題に十分ではない．

6. 結　　論

本論文では，単一系列データ上での系列選択の頻出パターン発見を目的として，高速な自動抽出方法を提案して逆単調性をみたす出現尺度である全体出現頻度を提案した．また系列全体頻度との頻出パターン抽出数の比較実験を行い，同条件の最小出現頻度で，より多くの頻出パターンを抽出することができ，本手法の有用性を確認した．今後は，抽出するパターンを正規表現に広げ考察していく予定である．

謝　　辞

本研究の一部は文部科学省科学研究費補助金 (課題番号 16500070) の支援をいただいた．本実験に対しては国立情報学研究所より，NTCIR-3 特許検索タスクテストコレクションの提供をいただきました．関係各位に深く感謝します．

文　　献

[1] 高野，岩沼，鍋島: 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, 第 3 回情報科学技術フォーラム (FIT), 2004, pp.115-118

[2] 長尾 真: 自然言語処理, 岩波書店, 1996

[3] 永田, 平田: "テキスト分類–学習理論の「見本市」–", 情報処理,vol.42(1),pp:32–37(2001)

[4] 本吉，三浦，塩谷: 回帰分析によるストリームデータのクラスタリング, 日本データベース学会 DBSJ Letters Vol.2,No.3, 2003, pp.45-48

[5] 北, 津田, 獅子堀: 情報検索アルゴリズム, 共立出版, 2002

[6] Agrawal, R. and Srikant, R.: Fast Algorithm for Mining Association Rules, proc. VLDB, 1994, pp.487-499

[7] Agrawal, R. and Srikant, R.: Mining Sequential Patterns, proc. ICDE, 1995, pp.3-14

[8] Garofalakis, M., Rastogi, R. and Shim, K.: SPIRIT : Sequential Pattern Mining with Regular Expression Constraints, proc. VLDB, 1999, pp.223-234

[9] Han, J., Pei, J. Mortazavi, B., Chen, Q, Dayal, U. and Hsu, M-C.: FreeSpan: Frequent Pattern-Projected Sequential Patterns Mining, proc. KDD, 2000, pp.355-359

[10] Han,J. and Kamber, M.: Data Mining : Concepts and Techniques, Morgan Kaufmann, 2000

[11] Albert-Lorincz, H. and Boulicaut, J-F.: Mining Frequent Sequential Patterns under Regular Expressions: A Highly Adaptative Strategy for Pushing Constraints, proc. SIAM DM, 2003, pp.316-320

[12] Mannila, H. and Toivonen, H. and Verkamo, I.: Discovery of Frequent Episodes in Event Sequences, *Data Mining and*

(注 2)：従って出現する回数も考慮している．

Knowledge Disvocery 1(3), 1997, pp.259-289

[13] Hand, D., Mannila, H. and Smyth, P.: Principles of Data Mining, MIT Press, 2001

[14] Motoyoshi, M., Miura, T., Watanabe, K. and Shioya, I.: Temporal Class Mining for Time Series Data, proc. CIKM, 2002

[15] Pei, J., Han, J., Mortazavi, B., Pinto H., Chen, Q, Dayal, U. and Hsu, M-C.: PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Patter Growth, proc.ICDE, 2001, pp.215-224

[16] Srikant, R and Agrawal, R.: Mining Sequential Patterns: Generalizations and Performance Improvements, proc.EDBT, 1996, pp.412-421

[17] Zaki, M.J.: Efficient Enumeration of Frequent Sequences, proc.CIKM, 1998, pp.68-75