特徴的な時系列パターンを発見するための新指標の提案

Proposal of a New Indicator to Discover Characteristic Sequential Patterns

櫻井 茂明♦　北原 洋一♠
折原 良平♠

Shigeaki SAKURAI　Youichi KITAHARA
Ryohei ORIHARA

本論文では，特徴的な時系列パターンを発見する新指標として系列興味度を提案する．提案する系列興味度は，時系列パターンの頻度と特定の部分時系列パターンとの出現のしやすさを評価することができる．本論文では，系列興味度がアプリオリ性を満たすことを示すとともに，従来の指標である支持度及び信頼度との関係を理論的に明らかにする．また，系列興味度に基づいて特徴的な時系列パターンを効率的に発見する方法を提案する．系列興味度の効果を SFA システムによって収集された営業日報から得られた時系列データに適用し，その効果を検証する．

This paper proposes a new indicator, called sequential interestingness, in order to discover interesting sequential patterns. The indicator can evaluate both frequency of sequential patterns, and relationships between the patterns and their specific sub-patterns. This paper shows the indicator satisfies with apriori property and shows relationships between the indicator and previous indicators: support and confidence. This paper also proposes a method that efficiently discovers all interesting sequential patterns based on the sequential interestingness. In addition, this paper verifies its effectiveness by numerical experiments based on sequential data of daily business reports collected by our SFA system.

1. はじめに

コンピュータ環境及びネットワーク環境の進展に伴って，時間情報が付随したデータを簡便に収集できるようになった．従来の研究の多くは，数値的な時系列データを対象としていたものの，近年時系列的なテキストデータを扱う方法が研究されるようになっている[3][4][10]．

これに対して，新たな観点での時系列テキストデータの分析法として，時間情報を持ったテキストデータから複数のテキストにまたがる時系列パターンを発見する方法が提案されている[8][9]．提案法では，分析者の背景知識を利用することにより，特徴的な時系列パターンを発見することができる．しかしながら，十分な背景知識が存在しない場合や分析者の思いもかけないような時系列パターンを発見したい場合には，提案法を適用することはできなかった．

一方，時系列データの中から頻出する時系列パターンを効率的に発見する方法も提案されている[1][5][7]．提案法では，頻出する時系列パターンを特徴的な時系列パターンとして発見するものの，発見された時系列パターンが必ずしも特徴的な時系列パターンになっているとはいえなかった．これに対して，制約を与えることにより，特徴的なパターンを発見する方法も提案されている[2][6]ものの，[8][9]と同様に背景知識に依存する問題を抱えていた．

そこで，本論文では，背景知識を利用することなく，特徴的な時系列パターンを発見するために，特徴的な時系列パターンを発見する新指標を提案し，提案する指標に基づいた効率的な時系列パターンの発見法を提案する．また，その効果を営業日報から得られた時系列データに適用して検証する．

2. 系列興味度

2.1 従来の指標

複数のイベントから構成される要素が順序構造を持って並べられた要素の列を時系列パターンとする．この時系列パターンを特徴付ける従来指標として，式(1)及び式(2)によって定義される，支持度及び信頼度が知られている．

$$supp(s) = f_s(s) / N \quad (1)$$
$$conf(s) = f_s(s) / f_s(s_p) \quad (2)$$

ただし，$f_s()$ を時系列パターン s が含まれる時系列データの頻度，N を時系列データの総数，s_p を s に含まれる時系列パターンとする．また，s が s_p を含むとは，s_p のすべての要素が s のいずれかの要素に，順序関係を保存したままで含まれることを意味している．

はじめに，支持度の性質について考えてみれば，任意の時系列パターンの支持度には，そのすべての部分時系列パターンの支持度以下になるといった性質(アプリオリ性)が成立している．このため，小さな時系列パターンを順次大きな時系列パターンへと成長させていくことにより，頻出する時系列パターンを効率的に発見することができる．しかしながら，頻出する時系列パターンはありふれた時系列パターンであることも多く，必ずしも分析者が求める特徴的な時系列パターンを発見することはできなかった．

これに対して，高い信頼度を持つ時系列パターンは，その部分時系列パターンが得られた段階で，次の状況をもっともらしく予測するルールとして利用することができる．このため，信頼度の高い時系列パターンはある種の特徴的な時系列パターンとみなすことができる．しかしながら，信頼度においてはアプリオリ性が成立していないため，信頼度の高い時系列パターンだけを直接発見することはできなかった．

そこで，特徴的な時系列パターンを効率よく発見するために，アプリオリ性を満たす新たな指標を次の節で検討する．

2.2 系列興味度の定義

特定の時系列パターンの中に，相対的な頻度がそれ程高くない部分時系列パターンが含まれている場合を考えてみる．このような時系列パターンは，相対的な頻度がそれ程高くない部分時系列パターンが与えられた段階で，時系列パター

♦ 正会員　(株)東芝 研究開発センター システム技術ラボラトリー shigeaki.sakurai@toshiba.co.jp
♠非会員　(株)東芝 研究開発センター システム技術ラボラトリー (youichi.kitahara, ryohei.orihara)@toshiba.co.jp

に含まれる残りのイベントを精度よく予測することができる．このため，ある種の特徴的な時系列パターンとみなすことができる．そこで，相対的な頻度がそれ程高くないことを時系列パターンに含まれる部分時系列パターンの頻度の逆数の最小値によって評価することにより，このような時系列パターンを発見する指標として，系列興味度を式(3)のように定義する．

$$inst(s) = \min_{s_p \subseteq s}\{(1/f_s(s_p))^{\alpha}\} \times ((f_s(s))^{(1+\alpha)}/N) \quad (3)$$

ただし，α≧0 を系列興味度パラメーターとする．本式は，α=0 の場合に，通常の支持度の定義を表しており，時系列パターンに含まれるイベントの数が 1 の場合には，支持度と一致する．次に，本式がアプリオリ性を満たすことを証明する．

［証明］ s_1，s_2 を条件 $s_1 \subseteq s_2$ を満たす時系列パターンとする．このとき，以下に示す関係が成立する．

$$\begin{aligned} inst(s_2) &= \min_{s_p \subseteq s_2}\{(1/f_s(s_p))^{\alpha}\} \times ((f_s(s_2))^{(1+\alpha)}/N) \\ &\le \min_{s_p \subseteq s_2}\{(1/f_s(s_p))^{\alpha}\} \times ((f_s(s_1))^{(1+\alpha)}/N) \\ &= \min[\min_{s_p \subseteq s_1}\{(1/f_s(s_p))^{\alpha}\}, \\ &\quad \min_{s_p \subseteq ((s_p \subseteq s_2) \cap (s_p \not\subseteq s_1))}\{(1/f_s(s_p))^{\alpha}\}] \times ((f_s(s_1))^{(1+\alpha)}/N) \\ &\le \min_{s_p \subseteq s_1}\{(1/f_s(s_p))^{\alpha}\} \times ((f_s(s_1))^{(1+\alpha)}/N) = inst(s_1) \end{aligned}$$

従って，系列興味度においてはアプリオリ性が成立する□

一方，式(3)を変形することにより，式(4)を得ることができ，支持度を時系列パターンに含まれる部分時系列パターンの最小信頼度によって補正した値とみなすこともできる．

$$inst(s) = \min_{s_p \subseteq s}\{(conf(s \mid s_p))^{\alpha}\} \times \mathrm{supp}(s) \quad (4)$$

2.3 系列興味度に基づいた時系列パターンの発見

アプリオリ性を利用した AprioriAll[1]ライクなアルゴリズムを構成することにより，指定した最小系列興味度以上となるすべての時系列パターンを効率的に発見することを試みる．構成する時系列パターンの発見法は，イベントの発見，イベント集合の発見，時系列パターンの発見といった 3 つのプロセスから構成されており，各プロセスを順に説明していくことにする．

第 1 のプロセスであるイベントの発見では，系列データの中からイベントをひとつ取り出して，取り出したイベントが出現する系列データの個数を計算する．ここで，イベントの場合における系列興味度が支持度と一致することに注意すれば，計算した頻度を系列データの総数で割ることにより，系列興味度を計算することができる．このようにして計算した系列興味度が，指定した最小系列興味度以上になるかどうかを判定し，最小系列興味度以上となる場合に，当該イベントを特徴的なイベントとして抽出する．このイベントの発見プロセスを系列データに含まれるすべてのイベントに対して順次実施し，特徴的なすべてのイベントを発見する．

第 2 のプロセスであるイベント集合の発見では，イベントの発見プロセスで発見されたふたつのイベントを組み合わせることにより，イベントの個数が 2 となる候補イベント集合を生成する．この候補イベント集合を系列データに適用し，当該候補イベント集合に対応する系列興味度を計算する．この系列興味度が最小系列興味度以上となる場合に，当該候補イベント集合をイベントの個数が 2 となる特徴的なイベント集合として抽出する．一般には，(i-2)個のイベントが一致する，ふたつのイベントの個数が(i-1)となる特徴的なイベントの集合 $evs_1 = \{ev_1,\cdots,ev_{i-2},ev_{i-1}\}$ 及び $evs_2 = \{ev_1,\cdots,ev_{i-2},ev_i\}$ から，イベントの個数 i がとなる候補イベント集合 $evs = \{ev_1,\cdots,ev_{i-2},ev_{i-1},ev_i\}$ を生成する．ただし，重複なく候補イベント集合を生成するために，イベント間には特定の順序関係(例えば，辞書順)が指定されているとする．このとき，この候補イベント集合の系列興味度がしきい値以上になるかどうかを判定するため，その頻度 $f_s(evs)$ 及び候補イベント集合に含まれるイベントの最大頻度 $\max_{ev_k \in evs}\{f_s(ev_k)\}$ を計算する．ただし，その和集合がイベント集合に一致するようなふたつのイベント部分集合においては，各イベント部分集合に含まれるイベントの最大頻度の最大値がイベント集合におけるイベントの最大頻度と一致している．このため，イベントの個数が(i-1)となる特徴的なイベント集合の最大頻度を格納しておき，ふたつの最大頻度の最大値を計算することにより，当該候補イベント集合におけるイベントの最大頻度を容易に計算でき，最小系列興味度以上になるかどうかを容易に判定することができる．このような特徴的なイベント集合の発見を，含まれるイベントの個数を増やしながら、特徴的なイベント集合が発見されなくなるまで順次繰り返すことにより，すべての特徴的なイベント集合を発見する．以上により，第 1 及び第 2 のプロセスで発見された特徴的なイベント及び特徴的なイベント集合を 1 次時系列パターンとする．

第 3 のプロセスである時系列パターンの発見では，発見されたふたつの 1 次時系列パターンを組み合わせることにより，2 次候補時系列パターンを生成する．この 2 次候補時系列パターンを系列データ集合に適用することにより，当該候補時系列パターンの系列興味度を計算する．この系列興味度が最小系列興味度以上であるとすれば，当該候補時系列パターンを 2 次時系列パターンとして抽出する．一般には，(k-1)次時系列パターンから前方の(k-2)個の要素が一致するふたつの(k-1)次時系列パターン $esq_1 = (s_p, el_1)$ 及び $esq_2 = (s_p, el_2)$ を抽出して組み合わせることにより，k 次候補時系列パターン $esq = (s_p, el_1, el_2)$ を生成する．ただし，s_p を(k-2)次部分時系列パターン，el_1 及び el_2 をそれぞれ時系列パターンの要素とする．このとき，当該の k 次候補時系列パターンの系列興味度がしきい値以上になるかどうかを判定するため，その頻度 $f_s(esq)$ を計算する．また，候補時系列パターンに含まれるイベントの最大頻度 $\max_{ev \in esq}\{f_s(ev)\}$ を計算する．ここで，組み合わせた時系列パターンが時系列パターンに一致するようなふたつの部分時系列パターンにおいては，各部分時系列パターンに含まれるイベントの最大頻度の最大値が時系列パターンにおけるイベントの最大頻度と一致している．このため，候補時系列パターンの元になる特徴的な時系列パターンに対応する最大頻度を格納しておくことにより，当該最大頻度も容易に計算することができ，最小系列興味度以上になるかどうかを容易に判定することができる．このような系列の延伸を時系列パターンが生成できなくなるまで繰り返すことにより，すべての時系列パターンを発見する．

// イベント発見
$L_{11} = \phi$;
For each event $ev \in el, el \in es, es \in SeqDB$
  $freq = calc\_freq(ev, SeqDB, 1)$;
  $inst = freq / |SeqDB|$;
  If $inst \geq MinInst$;
  Then store $freq$ to $sf[ev]$; add ev to L_{11};
// イベント集合発見
For $(i = 2; L_{1i-1} != \phi; i++)$
  $L_{1i} = \phi$; $N_{1i-1} = \phi$;
  For each event set $evs_1 \in L_{1i-1}$
    add evs_1 to N_{1i-1};
    For each event set $evs_2 \in (L_{1i-1} - N_{1i-1})$
      If $subset(evs_1, i-2) == subset(evs_2, i-2)$;
      Then $evs = evs_1 \cup evs_2$;
        $freq = calc\_frea(evs, SeqDB, i)$;
        $tinst = \max(sf[evs_1], sf[evs_2])$;
        $inst = (freq/tinst)^{\alpha} \times (freq/|SeqDB|)$;
        If $inst \geq MinInst$;
        Then store $tinst$ to $sf[evs]$; add evs to L_{1i};
$L_1 = \cup_i L_{1i}$
//時系列パターン発見
For$(k = 2; L_{k-1} != \phi; k++)$
  $L_k = \phi$;
  For each sequence $esq_1 \in L_{k-1}$
    For each sequence $esq_2 \in L_{k-1}$
      If $subseq(esq_1, k-2) == subseq(esq_2, k-2)$
      Then $esq = esq_1 \Diamond_{seq} esq_2$;
        $freq = calc\_freq(esq, SeqDB, k)$;
        $tinst = \max(sf[esq_1], sf[esq_2])$;
        $inst = (freq/tinst)^{\alpha} \times (freq/|SeqDB|)$;
        If $inst \geq MinInst$;
        Then store $tisnt$ to $sf[esq]$; add esq to L_k;

図 1　系列興味度に基づいた時系列パターン発見法

Fig.1　Discovery Method of Sequential Patterns based on Sequential Interestingness

以上に説明した擬似コードを図 1 に示す．この擬似コードに従うことより，系列興味度がしきい値以上となる時系列パターンを効率的に発見することができる．図 1 においては，*SeqDB* を時系列データ集合，*MinInst* を最小系列興味度，L_{i1} をイベントの個数が *i* となる特徴的なイベント集合の集合，L_k を *k* 次時系列パターンの集合，*calc_freq*()を時系列パターンの頻度を計算する関数，*sf*[]を時系列パターンに含まれるイベントの最大頻度を格納する領域，*subset*()を指定した個数のイベントを先頭から辞書順にイベント集合から取り出す関数，*subseq*()を指定したサイズの部分時系列パターンを時系列パターンの前方から取り出す関数，$\Diamond_{seq}$ を最後尾の要素を除いた部分時系列パターンが一致するふたつの時系列パターンからサイズが 1 大きい候補時系列パターンを生成する演算とする．演算 $\Diamond_{seq}$ によって，ふたつの (*k*-1)次時系列パターンからひとつの *k* 次候補時系列パターンを生成することができる．

3. 数値実験

3.1　実験データ

社内の 5 つの営業部門に導入されていた SFA システムから入手した 27,731 件の営業日報を実験データとして利用する．各データから営業日報の分析にとって重要なイベントを営業日報の本文からシソーラスに基づいて抽出し，各データを顧客名，案件名でグループ化し，活動日の順に並べることにより，6,434 件の時系列データを生成する．

3.2　実験方法

第 1 の実験として，最小系列興味度及び最小支持度として，1.0%及び 2.0%の 2 種類を利用し，系列興味度パラメーターを 0.25，0.5，1.0，2.0，5.0，10.0 と変化させた実験を行う．また，第 2 の実験として，最小支持度が 3.0%の場合に発見される 1 次時系列パターンの数と系列興味度によって発見される 1 次時系列パターンの数が一致するように，各系列興味度パラメーターの最小系列興味度を調整した実験を行う．このとき，本パラメーターは，それぞれ 1.85E+0，1.19E+0，5.07E-1，9.10E-2，3.84E-4，4.34E-8 と与えられる．

3.3　実験結果

第 1 の実験結果として，2%の場合において発見される時系列パターンの数が変化する様子を図 2 に示す．また，第 2 の実験結果として，系列サイズ 1 の数が同数の場合における実験結果を図 3 に示す．各図においては，*x* 軸が系列サイズを示し，*y* 軸が時系列パターンの数を示している．

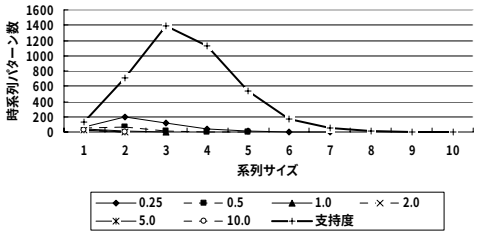

図 2　2%の場合における時系列パターン数

Fig.2　Number of Sequential Patterns in Case of 2%

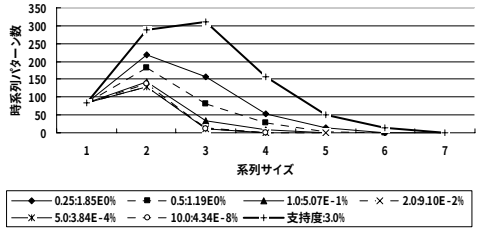

図 3　1 次時系列パターン数が一致する時系列パターン数

Fig.3　Number of Sequential Patterns in Case of the Same Number of 1-size Sequential Patterns

3.4　考察

系列興味度の特徴:式(4)にから分かるように，系列興味度の値は支持度の値よりも小さくなる一方，系列興味度パラメーターの値が大きくなるにつれて，その補正値も単調に減少

している．このため，系列興味度パラメーターの値も単調に減少する．従って，同じ値をしきい値として利用した場合には，系列興味度によって発見される時系列パターンの数は系列興味度パラメーターの値が増大するにつれて少なくなっている．図 2 はその性質を示しており，発見される時系列パターンの数が単調に減少している．

このため，支持度によって発見された時系列パターンと同程度の数の時系列パターンを発見したい場合には，系列興味度パラメーターの値に応じて，最小系列興味度を小さくする必要がある．系列興味度パラメーターの値と最小系列興味度の次数に着目してみれば，その値は近い値を示しており，従来の支持度と同じ感覚で系列興味度を利用するには，最小支持度の値を 1/α乗した値を参考にして，最小系列興味度を設定すればよいと考えられる．これは，支持度を最小信頼度のα乗によって補正したことに起因した現象と考えられる．

発見されるパターンの特徴:図 3 に示すように，系列サイズが 1 の時系列パターンの数が同数であったとしても，長い時系列サイズにおいて発見される時系列パターンの数は少なくなっている．この傾向は，系列興味度パラメーターの値が大きくなるにつれて顕著に現れており，支持度と最小信頼度のふたつの減少要因が存在することがその原因と考えられる．系列興味度は，支持度によって絞り込んだ後に信頼度によって絞り込むことと類似の効果を発揮できると考えられ，従来の支持度よりも特徴的な時系列パターンを発見しやすくなると考えられる．

また，系列興味度の場合には，一部のイベントの組み合わせを変更した類似の時系列パターンの数が減少する一方，それ程頻出していないイベントを含んだ時系列パターンを発見している．実験結果としては示していないものの，この傾向は系列興味度パラメーターの値が増大するにつれて顕著になっており，対応する最小信頼度の影響が大きくなることが原因と考えられる．従って，系列興味度の場合には，バリエーションに富んだ，頻度が少なく見逃されていたイベントを含んだ時系列パターンを発見することが期待できる．

計算時間:系列興味度においては，対応する最小頻度を計算する必要があるものの，発見された時系列パターンに対応する最大イベント頻度を記憶しておくことにより，効率的に系列興味度を計算することができる．一方，時系列パターンの発見法では，各パターンが系列データの集合に含まれる頻度を計算するのに最も多くの時間が必要であり，その他の計算部分は比較的小さな計算時間になっている．このため，計算時間としては同程度になると考えられ，実験システムにおいてもその差はそれ程大きなものとはならなかった．

以上の議論に基づいて，系列興味度は支持度よりも特徴的な時系列パターンを効率的に発見できると考えられる．

4. まとめと今後の課題

本論文では，支持度，信頼度に代わる新たな指標である系列興味度を定義し，その性質を理論的に明らかにした．また，系列興味度に基づいた特徴的な時系列パターンの発見法を提案し，その効果を SFA システムに基づく営業日報から得られた時系列データに適用して検証した．

今後の課題としては，発見された時系列パターンが分析者にとって真に興味ある時系列パターンになっているか，より詳細に検証する予定である．また，他の応用として分析している健診データ等への適用に向けて，テキストデータと数値データをシームレスに扱う方法を検討する予定である．

[文献]

[1] Agrawal, R. and Srikant, R.: “Mining Sequential Patterns", Proceedings of the 11th Int. Conf. Data Engineering, pp.3-14 (1995).

[2] Garofalakis, M. N., Rastogi, R. and Shim, K.: “SPIRIT: Sequential Pattern Mining with Regular Expression Constraints", Proceedings of the Very Large Data Bases Conf. 1999, pp.223-234 (1999).

[3] Lavrenko, V., Schmill, M., Lawrie, D., Ogilvie, P., Jensen, D., and Allan, J.: “Mining of Concurrent Text and Time-Series", Proceedings of the KDD-2000 Workshop on Text Mining, pp.37-44 (2000).

[4] Lent, B., Agrawal, R., and Srikant, R.: “Discovering Trends in Text Databases", Proceedings of the 3rd Int. Conf. on Knowledge Discovery and Data Mining, pp.227-230 (1997).

[5] Pei, J., Han, J., Mortazavi-Asl, B., Pinto, H., Chen, Q., Dayal, U., and Hsu, M.: “PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth", Proceedings of the 2001 Int. Conf. Data Engineering, pp.215-224 (2001).

[6] Pei, J., Han, J., and Wang, W.: “Mining Sequential Patterns with Constraints in Large Databases", Proceedings of the 11th ACM Int. Conf. on Information and Knowledge Management, pp.4-9 (2002).

[7] Srikant, R. and Agrawal, R.: “Mining Sequential Patterns: Generalizations and Performance Improvements", Proceedings of the 5th Int. Conf. Extending Database Technology, pp.3-17 (1996).

[8] Sakurai, S. and Ueno, K.: “Analysis of Daily Business Reports Based on Sequential Text Mining Method", Proceedings of the 2004 IEEE Int. Conf. on Systems, Man and Cybernetics, pp.3279-3284 (2004).

[9] Sakurai, S., Ueno, K., and Orihara, R.: “Introduction of Time Constraints to a Sequential Mining Method", WWW/Internet2005, 2, pp.328-332 (2005).

[10] Swan, R. and Jensen, D.: “TimeMines: Constructing Timelines with Statistical Models of Word Usage", Proceedings of the KDD-2000 Workshop on Text Mining, pp.73-80 (2000).

櫻井 茂明 Shigeaki SAKURAI

(株)東芝 研究開発センター システム技術ラボラトリー研究主務. 1991 東京理科大学大学院修士(数学)課程修了. 博士(工学), 技術士(情報工学). 機械学習，データ・テキスト・時系列マイニングの研究開発に従事．電子情報通信学会，日本知能情報ファジィ学会，人工知能学会，日本データベース学会各会員.

北原 洋一 Youichi KITAHARA

(株)東芝 研究開発センター システム技術ラボラトリー研究主事．2003 九州大学総合理工学部修士課程修了．データベースシステム，データマイニングの研究開発に従事

折原 良平 Ryohei ORIHARA

(株)東芝 研究開発センター システム技術ラボラトリー主任研究員．東京工業大学連携助教授．1988 筑波大学大学院工学研究科電子・情報工学専攻博士前期課程修了. 博士(工学). 発想支援技術，類推，機械学習，データ・テキストマイニングの研究開発に従事．人工知能学会，情報処理学会，日本ソフトウェア科学会各会員.