特徴的な時系列パターンの効率的な発見法

櫻井　茂明†　　北原　洋一†　　折原　良平†

† 株式会社 東芝 研究開発センター　〒 212–8582 神奈川県川崎市幸区小向東芝町 1
E-mail: †{shigeaki.sakurai,youichi.kitahara,ryohei.orihara}@toshiba.co.jp

あらまし　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。
キーワード　時系列パターンマイニング、アプリオリ性、支持度、信頼度、営業日報

Efficient Discovery Method of Interesting Sequential Patterns

Shigeaki SAKURAI†, Youichi KITAHARA†, and Ryohei ORIHARA†

† Corporate Research & Development Center, Toshiba Corporation,　1, Komukai-Toshiba-cho, Saiwai-ku, Kawasaki, Kanagawa, 212–8582, Japan
E-mail: †{shigeaki.sakurai,youichi.kitahara,ryohei.orihara}@toshiba.co.jp

Abstract　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.
Key words　Sequential Pattern Mining, Apriori Property, Support, Confidence, Daily Business Reports

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、営業日報、Web ログ、生体情報等の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。旧来から行われている研究の多くは、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きている。論文 [5] では、テキストデータを単語の並びからなる系列データとみなして、頻出するフレーズを発見し、頻出するフレーズの期間ごとにおける頻度の変化からテキストデータに内在する傾向を発見する方法を提案している。また、論文 [4] では、時系列に与えられる数値データをいくつかの trend に分割し、その trend よりも前に出現するテキストデータと trend とを関連付けて分析する方法を提案している。さらには、論文 [11] では、テキストに含まれる名詞句及び固有名詞によりテキストを特徴付ける一方、指定した時間に含まれるテキストと含まれないテキストに分割し、特徴と時間との間の関係を χ^2 検定することによって、特徴のグループ化を行っている。

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 [9] [10]。提案法では、分析者が注目する部分時系列パターンを指定する、イベント間に時間制約を導入するといった、分析者の背景知識を利用することにより、頻出する時系列パターンの中から特徴的な時系列パターンを発見することができる。このような背景知識に基づいた発見法の場合、十分な背景知識が存在する場合には、特徴的な時系列パターンを適切に発見することができるものの、十分な背景知識が存在しない場合や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、発見すべき時系列パターンを見逃す危険性があった。

一方、テキストデータを必ずしも対象とはしていないものの、時系列データの中から頻出する時系列パターンを効率よく発見する方法も提案 [1] [6] [8] されている。このような発見法の場合、頻出する時系列パターンが特徴的な時系列パターンとして

発見される。しかしながら、頻出する時系列パターンは分析者にとっては既知の時系列パターンであることも多く、必ずしも特徴的な時系列パターンにはなっていないという問題があった。この問題に対して、論文 [2] では、正規表現を指定することにより、正規表現を受理する時系列パターンを抽出する方法を提案している。また、論文 [7] では、射影型の時系列パターン発見アルゴリズムにおいて、発見される時系列パターンに制約を与える方法を提案するとともに、効率的に制約を満たす時系列パターンを発見する方法 prefix-growth を提案している。これらの制約に基づいた時系列パターンの発見法は、先の注目する部分時系列パターンを指定する方法と同様に、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方法として、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とを評価可能な系列興味度を提案し、その性質及び従来の指標との関係を明らかにする。また、系列興味度の高い時系列パターンを特徴的な時系列パターンとして効率よく発見する方法を提案する。加えて、提案法の効果を Sales Force Automation (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。

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。

$$supp(s) = \frac{f_s(s)}{N} \tag{1}$$

$$conf(s|s_p) = \frac{f_s(s)}{f_s(s_p)} \tag{2}$$

ここで、$f_s()$ を時系列パターン s を含む時系列データの頻度、N を時系列データの総数、$s_p \subseteqq s$ とする。図 1 は時系列パターン s が時系列パターン s_p を含む様子を示している。

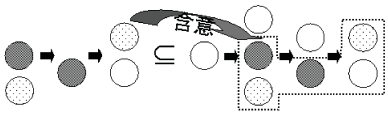


図 1　時系列パターンの含意

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) が成立する。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、分析者にとって必ずしも興味ある時系列パターンにはなっていなかった。このため、支持度に基づいて時系列パターンを発見したとしても、分析者が求める特徴的な時系列パターンを必ずしも発見することはできなかった。

これに対して、信頼度を利用することにより、時系列パターンを延ばしたとしてもそれ程頻度が変化しない時系列パターンを発見することができる。このような時系列パターンは、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、信頼度の定義から分かるように、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い。このため、支持度の高い時系列パターンを一旦発見し、その中から信頼度の高い時系列パターンを発見するといった方法が通常は行われている。このような 2 段階の発見法の場合、支持度の値を小さくし過ぎると、発見される時系列パターンが膨大になるため、特徴的な時系列パターンの発見に多くの時間が必要であった。一方、支持度の値を大きくし過ぎると、特徴的な時系列パターンが支持度を利用した判定の段階ではじかれてしまい、特徴的な時系列パターンを見落とす危険性があった。

この他の従来法として、参考文献 [12] では、時系列パターンに対して意外性を定義することにより、時系列パターンの中から特徴的な時系列パターンを発見する方法を提案している。しかしながら、提案法においては、意外性があるかどうかを判定する候補時系列パターンが発見されていることを前提としており、意外性のある時系列パターンを発見するには、GSP [8] などの時系列パターンの発見法を利用して、予め候補となる時系列パターンを発見しておかなければならなかった。また、候補となる時系列パターンの頻度の増減と意外性の値の増減との間には単調な関係が存在しないため、すべての意外性のある時系列パターンを発見するには、低頻度の時系列パターンをも候補として発見する必要があった。このため、先に説明した 2 段階の発見法と同様な問題を抱えていた。

このような背景の下、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時系列パターンの指標を新たに検討する。

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とにする。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、時系列パターン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。

$$inst(s) = \min_{s_p \subseteqq s} \{(\frac{1}{f_s(s_p)})^{\alpha}\} \times \frac{(f_s(s))^{(1+\alpha)}}{N} \tag{3}$$

ただし、$\alpha(\geqq 0)$ を系列興味度パラメーターとする。本式により、時系列パターン s としての頻度が比較的高く、相対的な頻度がそれ程高くない部分時系列パターン s_p とともに現れやすい時系列パターンを発見することができる。本式は、$\alpha = 0$ 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、$\min_{s_p \subseteqq s}(\frac{1}{f_s(s_p)}) = \frac{1}{f_s(s)}$ となるため、本式の値は支持度と一致する。以下においては、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 $s_1 \subseteqq s_2$ 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。

$$
\begin{aligned}
inst(s_2) &= \min_{s_p \subseteqq s_2}\{(\frac{1}{f_s(s_p)})^{\alpha}\} \times \frac{(f_s(s_2))^{(1+\alpha)}}{N} \\
&\leqq \min_{s_p \subseteqq s_2}\{(\frac{1}{f_s(s_p)})^{\alpha}\} \times \frac{(f_s(s_1))^{(1+\alpha)}}{N} \\
&= \min[\min_{s_p \subseteqq s_1}\{(\frac{1}{f_s(s_p)})^{\alpha}\}, \\
&\qquad \min_{s_p \subseteqq ((s_p \subseteqq s_2) \cap (s_p \not\subseteqq s_1))}\{(\frac{1}{f_s(s_p)})^{\alpha}\}] \times \frac{(f_s(s_1))^{(1+\alpha)}}{N} \\
&\leqq \min_{s_p \subseteqq s_1}\{(\frac{1}{f_s(s_p)})^{\alpha}\} \times \frac{(f_s(s_1))^{(1+\alpha)}}{N} = inst(s_1)
\end{aligned}
$$

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) を得ることができる。

$$
\begin{aligned}
inst(s) &= \min_{s_p \subseteqq s}\{(\frac{1}{f_s(s_p)})^{\alpha}\} \times \frac{(f_s(s))^{(1+\alpha)}}{N} \\
&= \min_{s_p \subseteqq s}\{(\frac{f_s(s)}{f_s(s_p)})^{\alpha}\} \times \frac{f_s(s)}{N} \\
&= \min_{s_p \subseteqq s}\{(conf(s|s_p))^{\alpha}\} \times supp(s)
\end{aligned}
\tag{4}
$$

式 (4) から分かるように、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して系列興味度を定義することができる。

また、式 (3) を変形することにより、式 (5) を得ることもできる。ただし、ev を時系列パターンを構成するイベントとする。

$$
\begin{aligned}
inst(s) &= \min_{s_p \subseteqq s}\{(\frac{1}{f_s(s_p)})^{\alpha}\} \times \frac{(f_s(s))^{(1+\alpha)}}{N} \\
&= \frac{1}{\max_{s_p \subseteqq s}\{(f_s(s_p))^{\alpha}\}} \times \frac{(f_s(s))^{(1+\alpha)}}{N} \\
&= \frac{1}{\max_{ev \in s}\{(f_s(ev))^{\alpha}\}} \times \frac{(f_s(s))^{(1+\alpha)}}{N} \\
&= \min_{ev \in s}\{(\frac{1}{f_s(ev)})^{\alpha}\} \times \frac{(f_s(s))^{(1+\alpha)}}{N} \\
&= \min_{ev \in s}\{(conf(s|ev))^{\alpha}\} \times supp(s)
\end{aligned}
\tag{5}
$$

式 (5) から分かるように、支持度を時系列パターンに含まれるイベントが与えられた場合における最小信頼度によって補正した値として系列興味度を定義することもできる。

加えて、式 (4)、式 (5) から分かるように、α の値を大きくすることにより、信頼度の影響を大きくすることができる。このため、信頼度を重視したい場合には、α の値を大きくする一方、支持度を重視したい場合には、α の値を小さくすることにより、タスクに応じて系列興味度を柔軟に指定することができる。

ここで、頻度が大きな時系列パターンよりも、頻度の小さな時系列パターンがどのくらいの確率で選ばれるのかを考えてみることにする。このような頻度の大小と系列興味度の大小における逆転現象が起きるには、以下の条件が成り立つ必要がある。

$$0 \leqq \alpha \tag{6}$$

$$0 \leqq n \leqq m \tag{7}$$

$$0 \leqq n \leqq y \tag{8}$$

$$0 \leqq m \leqq x \tag{9}$$

$$(\frac{n}{y})^{\alpha}\frac{n}{N} \geqq (\frac{m}{x})^{\alpha}\frac{m}{N} \tag{10}$$

ただし、n, m を時系列パターン s_a, s_b の頻度 $f_s(s_a), f_s(s_b)$、y, x を $\max_{ev \in s_a}\{f_s(ev)\}, \max_{ev \in s_b}\{f_s(ev)\}$ に対応する式の値とする。このとき、n と y、m と x の間には、時系列パターン s_a, s_b を介した従属関係が存在している。しかしながら、系列データベースを適切に設定することにより、条件 (8)、(9) が成り立つ範囲において、任意の n, m に対して、任意の y, x を設定することができる。このため、n と y、m と x は互いに独立しているとみなすことができる。従って、条件 (6)、(10) より、以下の条件を得ることができる。

$$(\frac{n}{m})^{\frac{\alpha+1}{\alpha}}x \geqq y \tag{11}$$

以上により、逆転現象が起こる領域 $S(x)$ は、図 2 の網掛けされた領域として与えられる。

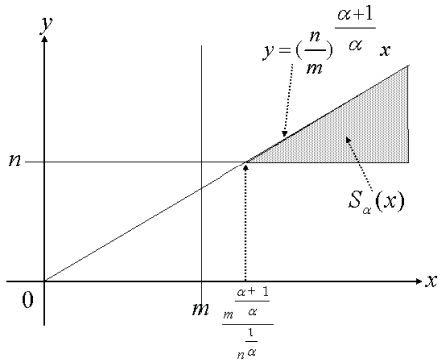


図 2　頻度と系列興味度の逆転

次に、系列興味度パラメーターを α として、点 (T,0), $T > \frac{m^{\frac{\alpha+1}{\alpha}}}{n^{\frac{1}{\alpha}}}$ までの領域の面積 $S_\alpha(x)$ を求めることにすれば、本面積は以下のように与えられる。

$$
\begin{aligned}
S_\alpha(x) &= \int_{\frac{m^{\frac{\alpha+1}{\alpha}}}{n^{\frac{1}{\alpha}}}}^{T}\{(\frac{n}{m})^{\frac{\alpha+1}{\alpha}}x - n\}dx \\
&= [\frac{1}{2}(\frac{n}{m})^{\frac{\alpha+1}{\alpha}}x^2 - nx]_{\frac{m^{\frac{\alpha+1}{\alpha}}}{n^{\frac{1}{\alpha}}}}^{T} \\
&= \{\frac{1}{2}(\frac{n}{m})^{\frac{\alpha+1}{\alpha}}T^2 - nT\} \\
&\qquad -\{\frac{1}{2}(\frac{m^{\frac{2(\alpha+1)}{\alpha}}}{n^{\frac{2}{\alpha}}})(\frac{n}{m})^{\frac{\alpha+1}{\alpha}} - n\frac{m^{\frac{\alpha+1}{\alpha}}}{n^{\frac{1}{\alpha}}}\} \\
&= \frac{T^2}{2}(\frac{n}{m})^{\frac{\alpha+1}{\alpha}} - nT + \frac{m^{\frac{\alpha+1}{\alpha}}}{2n^{\frac{1-\alpha}{\alpha}}}
\end{aligned}
\tag{12}
$$

また、ふたつの系列興味度パラメーター $\alpha, \beta(\alpha > \beta > 0)$ に対して、その面積の差を求めれば、その差は式 (13) のように与えられる。

$$\begin{aligned} S_\alpha - S_\beta &= \{\frac{T^2}{2}(\frac{n}{m})^{\frac{\alpha+1}{\alpha}} - nT + \frac{m^{\frac{\alpha+1}{\alpha}}}{2n^{\frac{1-\alpha}{\alpha}}}) \\ &\quad -\{\frac{T^2}{2}(\frac{n}{m})^{\frac{\beta+1}{\beta}} - nT + \frac{m^{\frac{\beta+1}{\beta}}}{2n^{\frac{1-\beta}{\beta}}}\} \\ &= \frac{T^2}{2}\{(\frac{n}{m})^{\frac{\alpha+1}{\alpha}} - (\frac{n}{m})^{\frac{\beta+1}{\beta}}\} \\ &\quad + \frac{n^2}{2}\{(\frac{m}{n})^{\frac{\alpha+1}{\alpha}} - (\frac{m}{n})^{\frac{\beta+1}{\beta}}\} \end{aligned} \quad (13)$$

このとき、$\alpha > \beta > 0$ より、$\frac{\beta+1}{\beta} > \frac{\alpha+1}{\alpha} > 0$ であり、$\frac{n}{m} < 1$ であるので、式 (13) の第 1 項は正の値となる。一方、$T \to \infty$ とするならば、第 2 項は T に対して定数項となるので、面積の差は常に正となる。従って、α の値を大きくするにつれて、頻度の小さなものの系列興味度が頻度の大きなものの系列興味度よりも大きくなる確率が高くなるといえる。すなわち、α の値を大きくするにつれて、支持度によって抽出される時系列パターンとは異なる時系列パターンを、系列興味度は発見しやすくなるといえる。

2.3 系列興味度に基づいた時系列パターンの発見法

系列興味度はアプリオリ性を満たしているため、アプリオリ性を利用することにより、効率的に系列興味度の高い時系列パターンを発見することができる。そこで、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いる。以下においては、各プロセスの概要を順に説明していく。

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することにより、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の頻度を計算する。この頻度と候補イベント集合の基になったふたつのイベントの頻度、系列データの総数、系列興味度パラメーターの値から、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。上述した処理を、イベントの発見プロセスで発見されたすべてのイベントの組み合わせに対して実施することにより、すべてのイベントの個数が 2 となる特徴的なイベント集合を発見する。ここで、候補の生成において、先のイベント発見プロセスで発見されたイベントの組み合わせだけから候補を生成している点に注意する必要がある。このような候補生成が可能なのは、最小系列興味度よりも小さなイベントを含むイベントの個数が 2 となるイベント集合の系列興味度は、含意されるイベントの系列興味度以下になるため、最小系列興味度よりも小さくなるからである。

次に、このイベントの個数が 2 となる特徴的なイベント集合の中からひとつのイベントが一致するふたつのイベント集合を取り出し、組み合わせることにより、イベントの個数が 3 となる候補イベント集合を生成する。このようにして生成されるすべての候補イベント集合に対して、イベントの個数が 2 となるイベント集合の場合と同様な処理を実施することにより、イベントの個数が 3 となるすべての特徴的なイベント集合を発見する。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 $\{ev_1, \cdots, ev_{i-2}, ev_{i-1}\}$, $\{ev_1, \cdots, ev_{i-2}, ev_i\}$ から、イベントの個数が i となる候補イベント集合 $\{ev_1, \cdots, ev_{i-2}, ev_{i-1}, ev_i\}$ 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。図 3 は、i 個のイベントを持つ候補イベント集合が生成される様子を図示している。このとき、この候補イベント集合の系列興

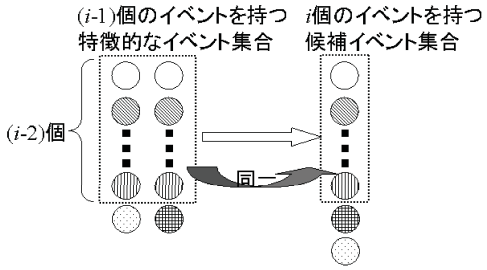


図 3　候補イベント集合の生成

味度がしきい値以上になるかどうかを判定するには、その頻度 $f_s(\{ev_1, \cdots, ev_{i-2}, ev_{i-1}, ev_i\})$ 及び候補イベント集合に含まれるイベントの最大頻度 $\max_{ev_k \in \{ev_1, \cdots, ev_{i-2}, ev_{i-1}, ev_i\}} \{f_s(ev_k)\}$ を計算する必要がある。ここで、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を考えたとすれば、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ので、式 (14) の関係が成立する。

$$\begin{aligned} &\max_{ev_k \in \{ev_1, \cdots, ev_{i-1}, ev_i, ev_{i+1}\}} \{f_s(ev_k)\} \\ &= \max[\max_{ev_k \in \{ev_1, \cdots, ev_{i-1}, ev_i\}} \{f_s(ev_k)\}, \\ &\qquad \max_{ev_k \in \{ev_1, \cdots, ev_{i-1}, ev_{i+1}\}} \{f_s(ev_k)\}] \end{aligned} \quad (14)$$

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、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。このような特徴的なイベント集合の発見を、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ことができる。このようにして、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が 1 次時系列パターンとなる。

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頻度を計算する。この頻度、候補時系列パターンに含まれるイベントの最大頻度、系列興味度パラメーターの値から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。このような処理を 1 次時系列パターンのすべての組み合わせに対して実施することにより、すべての 2 次時系列パターンを生成することができる。ここで、1 次時系列パターンの組み合わせだけを対象としてもすべての 2 次時系列パターンを発見することができるのは、系列興味度においてアプリオリ性が成立するためである。

次に、この 2 次時系列パターンの中から第 1 の要素が一致するふたつの時系列パターンを抽出して組み合わせることにより、3 次候補時系列パターンを生成する。この 3 次候補時系列パターンに対して、2 次候補時系列パターンの場合と同様に当該の 3 次時系列パターンの系列興味度を計算することにより、当該の 3 次候補時系列パターンの系列興味度が最小系列興味度以上になるかどうかを判定することができる。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 $(s_p, el_1), (s_p, el_2)$ 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 (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, el_2 をそれぞれ時系列パターンの要素とする。図 4 は、k 次候補時系列パターンが生成される様子を図示している。このとき、

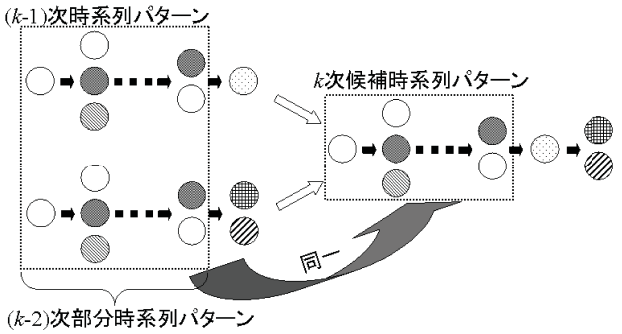


図 4　候補時系列パターンの生成

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には、その頻度 $f_s((s_p, el_1, el_2))$ を計算する必要がある。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 $\max_{ev \in (s_p, el_1, el_2)} (f_s(ev))$ を計算する必要がある。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を考えたとすれば、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ので、式 (15) の関係が成立する。

$$\max_{ev \in (s_p, el_1, el_2)} \{f_s(ev)\} = \max[\max_{ev \in (s_p, el_1)} \{f_s(ev)\}, \max_{ev \in (s_p, el_2)} \{f_s(ev)\}] \quad (15)$$

このため、基になった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処理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効率的に発見することができる。

以上により、図 5 に示す擬似コードに従うことにより、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 5 のアルゴリズムにおいては、SeqDB を時系列データ集合、MinInst を最小系列興味度、L_{1i} 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切り出す関数、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切り出す関数、$\bowtie_{seq}$ 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 $\bowtie_{seq}$ 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。

今回提案する系列興味度に基づいた時系列パターン発見アルゴリズムでは、AprioriAll ライクに発見アルゴリズムを構成している。しかしながら、系列興味度の枠組は、PrefixSpan [6] のような射影型の時系列パターン発見アルゴリズムにも適用することができる。計算速度の観点からは、射影型の発見アルゴリズムの方が有利であるとの報告がなされており、場合によっては、今後射影型の発見アルゴリズムを構成することが必要になると考えられる。

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。各データは、顧客名、担当者名、所属、活動日、案件名、ソリューション名、活動内容といった項目から構成されており、活動内容は自然言語で記述されたテキストデータである。このような SFA データに対して、顧客名、案件名が一致する営業日報ごとにグループを作成し、作成したグループごとに活動日の順に営業日報の並べ替えを行う。また、各営業日報の活動内容の中からキー概念辞書 [3] を利用することにより、顧客の印象や顧客に対して実施した活動などを表現したイベントを抽出する。こ

//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=$\frac{\text{freq}}{|\text{SeqDB}|}$;
  If inst $\geqq$ 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}!=ϕ; 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_freq(evs, SeqDB, 1);
        tinst $= \max$ (sf[evs_1], sf[evs_2]);
        inst=$(\frac{\text{freq}}{\text{tinst}})^{\alpha} \times \frac{\text{freq}}{|\text{SeqDB}|}$;
        If inst $\geqq$ MinInst;
        Then store tinst to sf[evs]; add evs to L_{1i};
$L_1 = \bigcup_i L_{1i}$;
//時系列パターン発見;
For(k=2; L_{k-1}!=ϕ; k++)
  $L_k = \phi$;
  For each sequence $es_1 \in L_{k-1}$
    For each sequence $es_2 \in L_{k-1}$
      If subseq($es_1, k-2$)==subseq($es_2, k-2$);
      Then $cs = es_1 \bowtie_{seq} es_2$;
        freq=calc_freq(cs, SeqDB, k);
        tinst $= \max$ (sf[es_1], sf[es_2]);
        inst=$(\frac{\text{freq}}{\text{tinst}})^{\alpha} \times \frac{\text{freq}}{|\text{SeqDB}|}$;
        If inst $\geqq$ MinInst;
        Then store tinst to sf[cs]; add cs to L_k;

図 5　系列興味度に基づいた時系列パターン発見アルゴリズム

こで、キー概念辞書とは分析対象となるタスクごとに専門家によって生成される一種のシソーラスである。本キー概念辞書は、実際にテキストデータに記述される表現を正規表現で記述した表層表現、同一の意味を持つ表層表現をまとめたキー概念、関連するキー概念をまとめた概念クラスからなる 3 階層の木構造形式によって記述されている。本キー概念辞書によって抽出されるキー概念がテキストデータを特徴付けるイベントとして利用される。今回の実験データの場合には、概念クラス数 3、キー概念数 61、表層表現数 835 といった規模のキー概念辞書を利用しており、61 種類のイベントによってテキストデータは特徴付けられている。各グループにおいては、抽出されたイベントをテキストデータの順序に従って並べることにより、ひとつのグループからひとつの時系列データが生成される。ただし、同じ活動日の営業日報から抽出されたイベントは同時に発生したとみなすことにより、同一の要素に含まれるイベントとみなすことにする。また、ひとつの要素において、同一のイベントが複数含まれる場合には、同じことが繰り返し記述されているとし、ひとつのイベントに集約することにする。以上のようにして実験データを生成することにより、与えられた営業日報から 6,434 件の時系列データを生成することができる。また、これら時系列データの中には、24,249 個の要素、57,133 個のイベントが含まれている。

3.2 実験方法

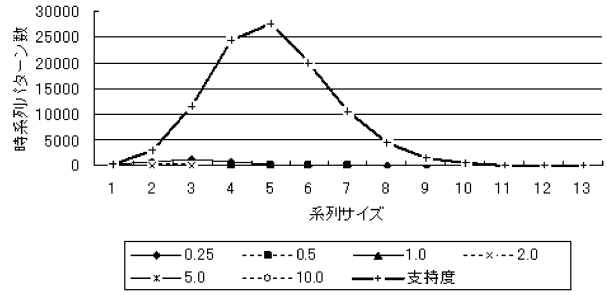
第 1 の実験として、最小系列興味度及び最小支持度として、1.0%及び 2.0%の 2 種類を利用するとともに、系列興味度パラメーターを 0.25、0.5、1.0、2.0、5.0、10.0 と変化させる。また、支持度を利用して時系列パターンを発見する。次に、第 2 の実験として、最小支持度が 3.0%の場合に抽出される 1 次時系列パターンの数と系列興味度によって発見される 1 次時系列パターンの数が一致するように、各系列興味度パラメーターの最小系列興味度を調整し、調整された最小系列興味度を用いて時系列パターンを発見する。以上の 2 種類の実験を実施した上で、発見される時系列パターンの違いと最小系列興味度及び最小支持度との間の関係を評価し、提案する系列興味度の効果を検証する。

3.3 実験結果

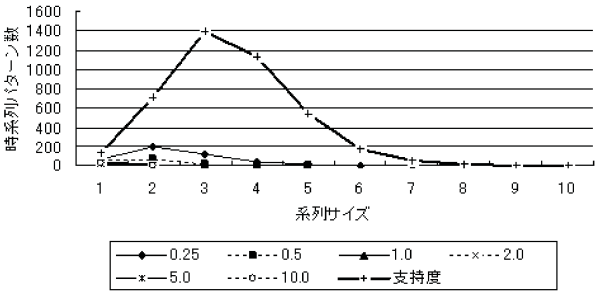
表 1 に第 2 の実験の場合に調整された、各系列興味度パラメーターに対応する最小系列興味度を示す。ただし、各値は指数表示によって表されており、Ex は 10^x を表している。また、図 6(a)〜(f) に第 1 及び第 2 の実験によって発見された時系列パターンの個数を比較した結果を示す。ただし、図 6(a)、(b) は、第 1 の実験の場合における系列サイズの変化に伴って発見される時系列パターンの数の推移を示している。これらの図においては、x 軸が系列サイズ、y 軸が発見される時系列パターンの数を示している。図 6(c) は、系列興味度及び支持度が、2.0%から 1.0%に小さくなった場合に、発見される時系列パターンの数が減少する割合を示している。この図においては、x 軸が系列サイズ、y 軸が時系列パターンの減少率を示している。図 6(d) においては、第 2 の実験の場合における系列サイズの変化に伴って発見される時系列パターンの数の推移を示している。この図においては、x 軸が系列サイズ、y 軸が発見される時系列パターンの数を示している。図 6(e)、(f) は、第 2 の実験において発見された時系列パターンに対して、支持度によって発見された時系列パターンを基準として、共通なものとして発見された時系列パターン及び系列興味度のみによって発見された時系列パターンの数が、系列興味度パラメーターの変化に伴って推移する様子を示している。これらの図においては、x 軸が系列興味度パラメーターの値、y 軸が発見される時系列パターンの数を示している。ただし、系列サイズが 4 以上の場合には、系列興味度のみによって発見される時系列パターンは存在していなかったため、図 6(f) には当該系列サイズに対応するグラフは記述されていない。

3.4 考　　察

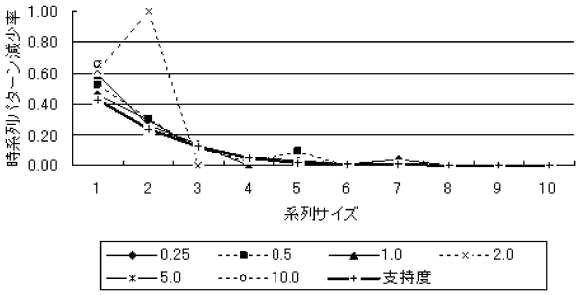
系列興味度の特徴: 系列興味度は支持度を時系列パターンの中で最大頻度を与えるイベントを条件とする信頼度によって補正された値として定義することができる。このため、その値は



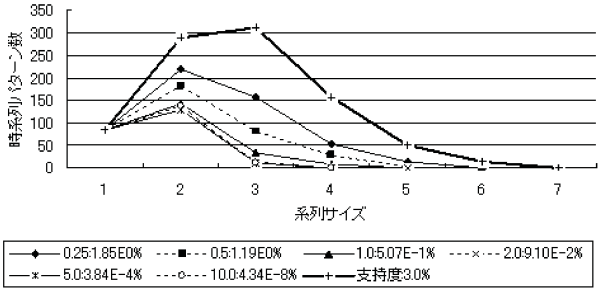
(a) 時系列パターン数 (1.0%)



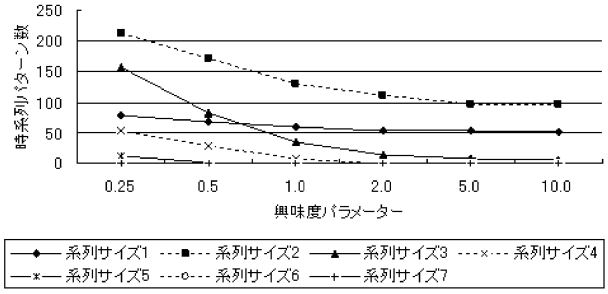
(b) 時系列パターン数 (2.0%)



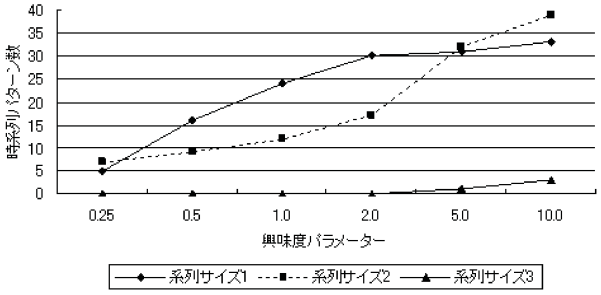
(c) 時系列パターン減少率



(d) 時系列パターン数 (系列サイズ 1 同数)



(e) 共通時系列パターン数



(f) 系列興味度単独時系列パターン数

図 6　実 験 結 果

表 1　系列興味度パラメーターの値

パラメーター	最小系列興味度	パラメーター	最小系列興味度
0.25	1.85E+0%	2.0	9.10E−2%
0.5	1.19E+0%	5.0	3.84E−4%
1.0	5.07E−1%	10.0	4.34E−8%

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従って、その補正値は単調に減少する。このため、系列興味度の値も単調に減少する。従って、同じ値を時系列パターンの発見の際に利用する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支持度によって発見される時系列パターンの数よりも少なくなる。図 6(a)、(b) はその性質を示しており、系列興味度パラメーターの値が大きくなるに従って、発見される時系列パターンの数は減少している。

また、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。ここで、表 1 に着目してみると、系列興味度パラメーターの値と最小系列興味度の次数が近い値を示している。このため、従来の支持度と同じような感覚で系列興味度を利用するには、考えていた最小支持度を $\frac{1}{\alpha}$ 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、支持度を信頼度の α 乗によって補正していることに起因した現象と考えられる。

次に、系列興味度及び支持度を小さくした場合に発見される時系列パターンの数が減少する割合に着目してみる。図 6(c) に示すように、系列興味度パラメーター 2.0、系列サイズ 2 の場合を除いて、各指標において同程度の割合で発見される時系列パターンの数は減少している。この原因としては、どちらの指標ともに時系列イベントパターンの増加に対して、単調に減少するといった基本的な性質が一致することが考えられる。一方、系列興味度パラメーター 2.0、系列サイズ 2 の場合には、系列興味度 1%及び 2%のいずれの場合においても、1 個の時系列パターンしか発見されておらず、例外的な値になったとものと考えられる。このため、最小支持度と同程度の感覚で最小系列興味度を小さくすることができると考えられる。

発見されるパターンの特徴: 図 6(d) が示すように、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、より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支持

度の場合に比べて少なくなっている。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従って顕著に現れている。系列サイズが大きくなるに従って、時系列パターンに含まれるイベントの種類も多くなるため、当該時系列パターンの中で最大となるイベントの頻度も大きくなり、対応する最小信頼度の値が小さくなる。このように、支持度と最小信頼度のふたつの減少要因が存在するため、系列サイズが大きくなるに従って系列興味度において発見される時系列パターンの数が減少したものと考えられる。この現象は、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すると考えられるため、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ものと考えられる。

また、支持度によって発見される時系列パターンの場合、頻出するイベントの組み合わせを若干変えた類似の時系列パターンが発見される傾向にある。これに対して、系列興味度の場合には、そのような時系列パターンの数が減少する一方で、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。この傾向は系列興味度パラメーターの値が増大するに従って顕著になっており、図 6(e)、(f) に示すように、共通する時系列パターンの数が減少する一方、系列興味度単独で出現する時系列パターンの数が増加している。この原因は、系列興味度パラメーターの値が大きくなるに従って、対応する最小信頼度の影響が大きくなるためと考えられる。この結果は、支持度の大小と系列興味度の大小が逆転する確率が高まるといった理論的な結果とも一致している。このように、系列興味度の場合には、支持度の場合よりもバリエーションに富んだイベントに基づいた時系列パターンを発見することができる。従って、頻度がそれ程多くないために見逃されていた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ため、従来の支持度に比べて余分な計算量が必要となる。しかしながら、図 5 の時系列パターン発見アルゴリズムに示すように、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。このため、支持度と同程度の計算量で時系列パターンを発見することができる。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、定義された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することが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、系列興味度を利用した特徴的な時系列パターンの発見法を提案した。提案する発見法は、支持度及び信頼度がともにある程度高い時系列パターンを、従来の支持度に基づいた時系列パターンの発見法と同程度の計算量で発見することができる。加えて、提案する発見法を SFA システムから得られた時系列データに適用しその効果を検証した。

今後の課題としては、本論文では系列興味度によって、従来よりも幅広くイベントを考慮して特徴的な時系列パターンを発見することができることを示したものの、発見された時系列パターンが分析者にとって興味ある時系列パターンになっているかどうかを十分には検証していない。このため、系列興味度のみによって発見された時系列パターンを分析者に提示する等して、今後その妥当性を検証する必要がある。また、現在取り扱っている時系列データは、テキストデータからイベントを発見して特徴的な時系列パターンを発見しているが、他の応用として分析している健診データ等では、テキストデータに加えて数値データをも時系列データの対象としている。このため、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していきたい。加えて、時間情報の付随したテキストデータの分析においては、テキストデータを特徴付けるイベントを設定すること自体が困難な場合もある。このため、テキストデータの形態素解析結果、構文解析結果等を利用することにより、イベントを設定することなしに時間情報の付随したテキストデータを分析する方法を検討していきたい。

文　献

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

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

[3] 市村 由美, 鈴木 優, 酢山 明弘, 折原 良平, 中山 康子, 「日報分析システムと分析用知識記述支援ツールの開発」, 信学論, J86-D-II, 2, 310-323 (2003).

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

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

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

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

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

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

[10] 櫻井 茂明, 植野 研, 酢山 明弘, 折原 良平, 「時系列イベントパターンマイニングにおける時間制約の導入」, DEWS2005, 6C-o1 (2005).

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

[12] 吉田 由紀子, 「意外性に基づく状態列パターンの評価装置」, 特開 2004-178515.