特　集「系列パターンマイニングの最近の動向」

多様なデータに対する系列パターンマイニングの適用

Application of Sequential Pattern Mining to Various Data Sets

櫻井　茂明　東芝ソリューション株式会社
Shigeaki Sakurai　Toshiba Solutions.
Sakurai.Shigeaki@toshiba-sol.co.jp

Keywords: sequential interestingness, time constraint, constraint pattern, sales force automation system, medical examination.

1. はじめに

コンピュータ環境およびネットワーク環境の発展に伴って，多量のデータが簡便に収集，蓄積されるようになった．このため，これらデータを分析する技術が活発に研究開発されている．このような大規模データは近年ますます巨大化しており，ビッグデータと呼ばれる流行語が，IT分野では生み出されている．多くの企業では，ビッグデータの分析に向けた技術開発に取り組んでいる．

ビッグデータは多様なデータによって構成されているが，TwitterやYouTubeの隆盛，スマートグリッドやスマートコミュニティへの期待の高まりに見られるように，時系列的に収集されるデータが今後ますます増大するものと考えられる．時系列的に収集されるデータを分析する技術は，当初，統計学や信号処理の分野で研究がなされており，時系列的に与えられる数値データを対象としていた．1990年代の後半に入ると，相関ルールの基となる頻出パターンの発見問題が，時系列的に拡張されるようになり，個々の商品のような離散的なアイテムおよびアイテムの集合が時系列的に並べられたデータから，系列パターンを発見する技術が研究開発されるようになった．このような系列パターンの発見技術は，商品の購買履歴の分析にその端を発するが，近年は，テキストを単語のつながりとみなした系列データの分析や，定期的に検査されるヘルスケアデータを離散化した系列データの分析へとその適用対象を広げている．また，より分析者のニーズに合った系列パターンを発見するために，系列パターン発見研究の初期で対象としていた，その頻出性に基づいた系列パターン以外の特徴的な系列パターンを発見することも求められるようになってきている．

このような背景のもと，本解説では，多様な系列データの中から特徴的な系列パターンを発見する技術として，系列パターンに含まれる部分系列パターン間の関係を評価する評価基準である系列興味度 [Sakurai 08a] を紹介する．また，系列データの分析に分析者の背景知識を導入する枠組みとして，時間制約 [Sakurai 08d] および制約パターン [Sakurai 08b] に基づいた方法を紹介する．一方，これら技術に基づいた分析例として，営業員が日々記載する営業日報を分析し，営業活動の改善に役立てる例 [櫻井 06]，および従業員に対して定期的に実施されている健康診断データを分析し，従業員の生活改善に役立てる例 [櫻井 07] を紹介する．

2. 系列パターンと系列データ

本解説で対象とする系列データとは，複数のアイテム集合が時系列的に並べられたアイテム集合の系列のことである．このとき，各アイテム集合には，同一のアイテムはせいぜい一つしか含まれていないことが仮定されている．本解説では，複数の系列データに現れる特徴的な部分系列を系列パターンとして発見する．[Agrawal 95] で問題設定された系列パターンの発見問題は，[Agrawal 94] で問題設定された相関ルールの発見問題において必要となる，頻出パターンの発見問題の自然な拡張となっている．このため，特徴的かどうかの判定には，式（1）で定義される支持度（*supp*）が利用されており，頻出する系列パターンを特徴的な系列パターンとして発見することを試みている．すなわち，与えられた部分系列の支持度が，指定した最小支持度以上となる部分系列を，特徴的な系列パターンとして発見している．

$$supp(s)=\frac{s\text{を含む系列データ数}}{\text{系列データ数}} \quad (1)$$

ただし，s は部分系列を表すとする．

また，相関ルールの発見問題の場合と同様に，発見された系列パターンに対して，その部分系列パターン間の信頼度を評価することにより，系列パターンのルール化も行っている．このとき，系列パターンにおける信頼度（*conf*）は式（2）に示すように定義されている．本式においては，s_p が系列パターン s に含まれる部分系列パタ

ーンを表している．

$$conf(s|s_p)=\frac{s\text{を含む系列データ数}}{s_p\text{を含む系列データ数}} \quad (2)$$

3. 系列パターンの発見法

前章で紹介した系列パターンは，系列パターンを構成するアイテムが成長するに従って，支持度が単調に減少するアプリオリ性を利用することにより，指定した最小支持度以上のすべての系列パターンを効率良く発見することを可能としている．このような発見法として，複数の方法 [Agrawal 95, Ayres 02, Pei 01, Zaki 01] が提案されているが，本解説では，当初提案された発見済みの系列パターンから，より大きな候補を生成することにより，すべての系列パターンを発見する方法 [Agrawal 95] を簡単に紹介する．候補に基づいた方法では，時系列データを構成する各アイテムに対して，当該アイテムを含む時系列データの数を算出して支持度を計算し，その支持度が最小支持度以上となるすべてのアイテムの発見を行う．この発見されたアイテムが1次頻出アイテム集合となる．次に，発見された1次アイテム集合を組み合わせることにより，二つのアイテムで構成された候補を生成し，その支持度が最小支持度以上となる候補を2次頻出アイテム集合として発見する．同様に，2次頻出アイテム集合を組み合わせることにより，三つのアイテムで構成された候補を生成し，その支持度が最小支持度以上となる候補を3次頻出アイテム集合として発見する．このような，アイテム集合の成長を順次実施していくことにより，すべての頻出アイテム集合の発見を行う．ここまでの頻出アイテム集合の発見は，系列データを単位に頻度を計算することを除いては，[Agrawal 94] における頻出パターンの発見法と同一の方法となっている．

候補に基づいた方法では，この頻出アイテム集合の中から二つのアイテム集合を取り出して組み合わせることにより，二つのアイテム集合が時系列的に並んだ系列パターンの候補（2次候補系列パターン）を生成する．このとき，取り出したアイテム集合の並べ方を変えることにより，2種類の候補が生成されることに注意する必要がある．候補に基づいた方法では，生成された候補を含む系列データの支持度を算出し，その支持度が最小支持度以上となる場合に，当該候補を2次頻出系列パターンとして判定する．同様に，2次頻出パターンを組み合わせることにより，三つのアイテム集合が時系列的に並んだ3次候補系列パターンを生成する．また，生成された候補の支持度を算出し，3次頻出系列パターンになるかどうかの判定を行う．以上のような系列の成長を繰り返していくことにより，すべての系列パターンの発見を行う．図1は，k 次頻出時系列パターンから $(k+1)$ 次候補系列パターンを生成する様子を示している．図においては，各丸が一つのアイテムを表しており，同一の時刻に発生したとみなされるアイテムが矢印によって区切られている．また，同じ模様の丸は同一のアイテムであることを示している．図からわかるように，前方にある $k-1$ 個のアイテム集合が同一であり，最後尾のアイテム集合が異なる二つの k 次頻出系列パターンを組み合わせることにより，候補の生成が行われていることに注意する必要がある．ただし，1次頻出系列パターンから2次候補系列パターンを生成する際には，前方のアイテム集合が存在しないため，任意の1次頻出系列パターンを組み合わせることができる．また，図の例では，上部にある k 次頻出系列パターンの最後尾のアイテム集合を共通する部分の次に配置することにより，候補を生成しているが，下部にある k 次頻出系列パターンの最後尾のアイテム集合を共通する部分の次に配置する候補も生成できることに注意する必要がある．

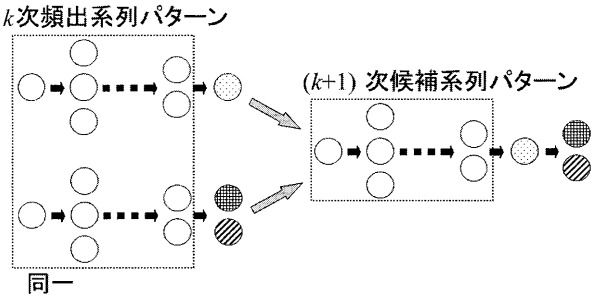


図1 候補系列パターンの生成

4. 特徴的な系列パターンの発見

系列パターンの発見法が発見する頻出系列パターンは，頻出しているという意味では特徴的なパターンであるものの，頻出系列パターンは，分析者にとっては既知の系列パターンであることも多く，必ずしも分析者に新たな知見を与える系列パターンにはなっていなかった．そこで，本章では，より有益な知見を与える系列パターンの発見を目指して導入した，系列パターンの評価基準および，系列パターンを分析者の知見に従って絞り込む制約条件の導入法を紹介する．

4·1 系列興味度

特定の系列パターンの中に，相対的な頻度がそれほど高くない部分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．ただし，相対的な頻度が高くないとは，各系列パターンの頻度を考えた場合に，その頻度が系列パターンの頻度分布において，下位のほうに位置することを意味していることとする．このような系列パターンが与えられているとすれば，相対的な頻度がそれほど高くない部分系列パターンが与えられた場合に，系列パターンに含まれる残りのアイテムの出現を精度良く予測するのに活用することができる．このため，このよう

な系列パターンは，ある種の特徴的な系列パターンと考えることができ，分析者にとって価値あるものであると考えられる．このとき，相対的な頻度がそれほど高くないことをいかにして表現するかが問題となる．本解説では，系列パターンに含まれる部分系列パターンの頻度の逆数の最小値に基づいてこの程度を評価した，系列興味度 [Sakurai 08a] を紹介する．具体的には，系列興味度は，式（3）に示すように定義される．

$$inst(s)=\min_{s_p\subseteq s}\left\{\left(\frac{1}{f(s_p)}\right)^{\alpha}\right\}\times\frac{(f(s))^{(1+\alpha)}}{N} \qquad (3)$$

ただし，α（$\geqq 0$）を系列興味度パラメータ，s_p を系列パターン s の部分系列パターン，$f(s)$ を系列パターン s が含まれる系列データ集合に含まれる系列データの数，N を系列データの総数とする．本式により，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ほど高くない部分系列パターン s_p とともに現れやすい系列パターンを発見することができる．

このように定義される系列興味度には，次のような性質が存在する．

性質 1：系列興味度においては，アプリオリ性 [Agrawal 95, Srikant 96] が成立する．すなわち，系列パターンの系列興味度はその部分系列パターンの系列興味度以下となる．

性質 2：系列興味度においては，従来の指標である支持度および信頼度との間に次のような関係が成立する．

$$inst(s)=\min_{s_p\subseteq s}\{(conf(s|s_p))^{\alpha}\}\times supp(s) \qquad (4)$$

かつ

$$inst(s)=\min_{it\in s}\{(conf(s|it))^{\alpha}\}\times supp(s) \qquad (5)$$

ただし，it は系列パターン s を構成するアイテムであるとする．

性質 2 からわかるように，系列興味度パラメータ 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．

性質 3：系列興味度パラメータ α が 0 の場合，系列興味度の値は支持度の値と一致する．また，系列パターンに含まれるアイテムの個数が 1 個の場合，系列興味度の値は支持度の値と一致する．

性質 4：二つの系列興味度パラメータ α_1 および α_2 に $\alpha_1 > \alpha_2$ の関係があるとする．このとき，α_1 に基づいた系列興味度のほうが α_2 に基づいた系列興味度よりも，支持度の値が小さな系列パターンを出力する確率が高くなる．

上述した系列興味度の性質により，系列興味度によって発見される系列パターンを構成するアイテムには，その出現頻度がそれほど高くないアイテムを含むことを期待することができる．このため，支持度に基づいた系列パターンの発見よりも，多様な系列パターンを発見することが期待できる．また，系列興味度によって獲得された系列パターンは部分系列パターンが与えられた場合に，その残りの部分系列パターンを高い確率で予測することができる系列パターンとなっている．このため，系列パターンを予測に利用可能という点で，信頼度によって発見される系列パターンと類似の性質をもった系列パターンを効率的に発見することが期待できる．

4・2 時間制約

前節までに説明してきた特徴的な系列パターンおいては，出現するアイテムの順序関係だけを考慮しており，時間的に掛け離れた，実際には意味のないアイテムの並びも，系列パターンの頻度として積算する可能性がある．このため，意味のないアイテムの並びを含んだ系列パターンを，特徴的な系列パターンとして抽出する危険性がある．この問題に対して，[Srikant 96] では，隣接するアイテムが指定された時間間隔に収まるように，アイテム間で満たすべき時間制約として，最小時間間隔および最大時間間隔を導入している．しかしながら，アイテム間の時間的な関係には多様な関係が存在しており，これら時間間隔だけでは，必ずしもアイテム間に柔軟な時間制約を導入することはできない．そこで，本解説では，柔軟な時間制約の導入を実現する一つの方法として，[Sakurai 08d] に提案されている七つの時間制約を紹介する．図 2 は，本制約のイメージを表しており，白抜きの各丸印が通常のアイテム，アミがけされた丸印が分析者によって指定された特定のアイテムを表している．

(1) 始端アイテム，終端アイテム間の時間制約

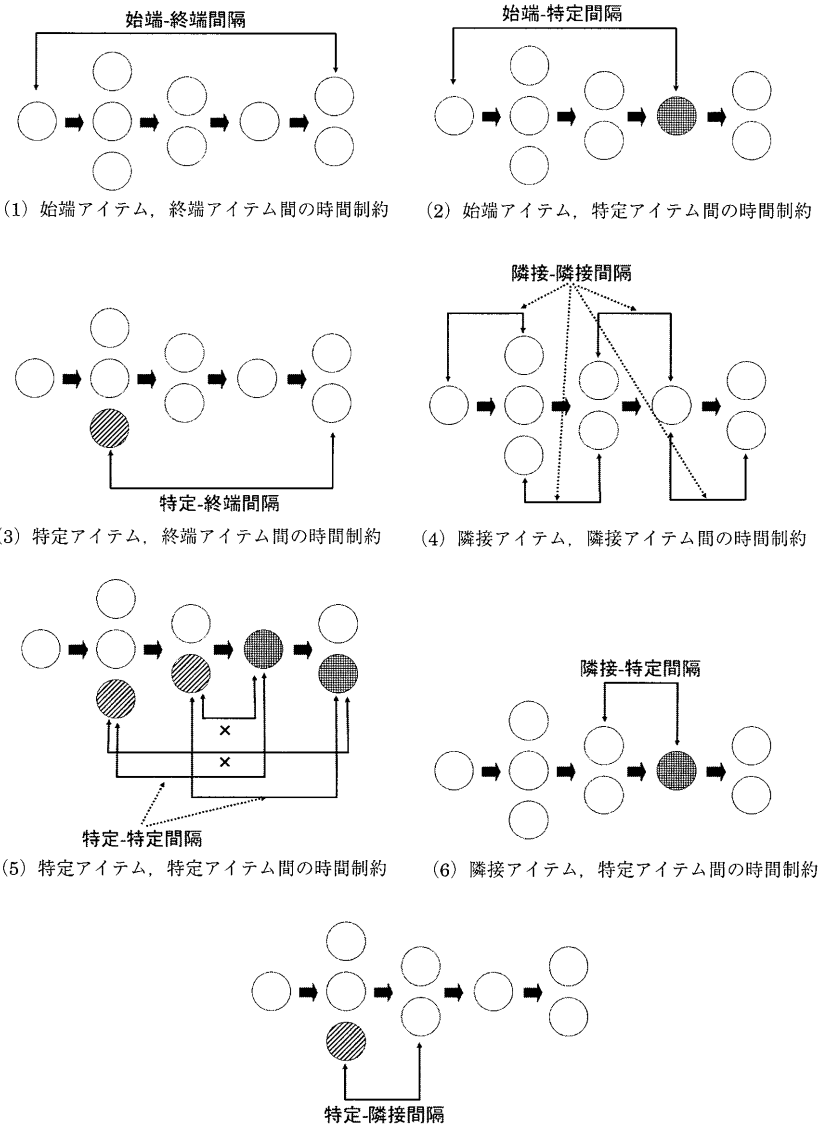
本制約は，系列パターンを構成する先頭のアイテム集合に含まれるアイテムと，最後尾のアイテム集合に含まれるアイテムとの間の時間間隔（始端 - 終端間隔）の最小値および最大値を指定する．これにより，始端 - 終端間隔が最小値と最大値の間に含まれる系列パターンのみを抽出することができる．

(2) 始端アイテム，特定アイテム間の時間制約

本制約では，系列パターンの先頭のアイテム集合に含まれるアイテムから，分析者が指定する特定のアイテムまでの時間間隔（始端 - 特定間隔）の最小値および最大値を指定する．これにより，始端 - 特定間隔が最小値と最大値の間に含まれる系列パターンを抽出することができる．

(3) 特定アイテム，終端アイテム間の時間制約

本制約では，分析者が指定する特定のアイテムから，系列パターンの最後尾のアイテム集合に含まれるアイテムまでの時間間隔（特定 - 終端間隔）の最小値および最大値を指定する．これにより，特定 - 終端間隔が最小値と最大値の間に含まれる系列パターンを抽出することが



(1) 始端アイテム，終端アイテム間の時間制約　(2) 始端アイテム，特定アイテム間の時間制約

(3) 特定アイテム，終端アイテム間の時間制約　(4) 隣接アイテム，隣接アイテム間の時間制約

(5) 特定アイテム，特定アイテム間の時間制約　(6) 隣接アイテム，特定アイテム間の時間制約

(7) 特定アイテム，隣接アイテム間の時間制約

図2　時間制約

できる．

(4) 隣接アイテム，隣接アイテム間の時間制約

本制約では，系列パターンにおいて隣接する，任意のアイテム間に対して，その時間間隔（隣接 - 隣接間隔）の最小値および最大値を指定する．これにより，系列パターンのすべての隣接するアイテムが，指定した最小値と最大値の範囲で発生する，系列パターンを抽出することができる．

(5) 特定アイテム，特定アイテム間の時間制約

本制約は分析者が指定する二つの特定のアイテム間に対して，その時間間隔（特定 - 特定間隔）の最小値および最大値を指定する．これにより，特定 - 特定間隔が指定した最小値と最大値の間に含まれる系列パターンを抽出することができる．ただし，一つの系列パターンの中に特定のアイテムの組が複数存在する場合には，その出現順序に従ったアイテムの組だけを考えることにする．すなわち，2種類のアイテム a，b が a，a，b，b の順に系列パターンに出現しているとすれば，最初の a と最初の b および2番目の a と2番目の b に対してのみ本制約を適用する．このような限定を置くことにより，系列パターンの中で循環するような部分系列パターンが，本制約を満たす系列パターンを抽出することができる．

(6) 隣接アイテム，特定アイテム間の時間制約

本制約は分析者が指定した特定のアイテムとその前方に隣接しているアイテムとの間に対して，その時間間隔（隣接 - 特定間隔）の最小値および最大値を指定する．これにより，隣接 - 特定間隔が指定した最小値と最大値の間に含まれる系列パターンを抽出することができる．

(7) 特定アイテム，隣接アイテム間の時間制約

本制約は分析者が指定した特定のアイテムとその後方に隣接しているアイテムとの間に対して，その時間間隔（特定 - 隣接間隔）の最小値および最大値を指定する．

これにより，隣接 - 特定間隔が指定した最小値と最大値の間に含まれる系列パターンを抽出することができる．

上記に紹介した時間制約は，候補となる系列パターンを含む系列データの頻度を算出する際に評価され，当該系列パターンに関連するすべての時間制約が成立する場合に，その頻度が1積算される．この時間制約を導入した系列パターンは，時間的に意味のある範囲にあるアイテムの並びで構成された系列パターンを発見することができる．このような時間制約は，分析者の対象データに対する背景知識に基づいて指定されることになるため，時間制約を考慮しない場合よりも，分析者にとって興味のある系列パターンを容易に発見することが期待できる．また，特別の知見を持ち合わせていないアイテムに関しては，時間制約を導入しなくてもよいため，時間制約を導入したとしても，特徴的な系列パターンを取りこぼす危険性はないものと期待できる．

4·3 制約パターン

前節までに対象としてきたアイテムの場合，個々のアイテムには，意味的な関係は特に設定されていなかった．しかしながら，系列パターンの発見問題の適用範囲が広がるにつれて，表構造で構成されたデータがその適用対象に含まれるようになってきた．このような表構造データの場合，各アイテムは属性と属性値で構成されており，同じ属性をもつアイテム同士は，異なる属性をもつアイテム同士よりも，意味的な深い関係をもっている．この意味的な関係を利用することにより，分析者にとって特徴的な系列パターンを効率的に発見することが期待できる．そこで，以下においては，この枠組みを紹介する．

この枠組みの中には，同一の属性で構成されたアイテムは，系列パターンを構成するアイテム集合の中には一つしか存在しないといった，表構造データの特徴に基づいた制約のほかにも，分析者は，同一の属性の変化に興味があるとの仮定のもとに提案されている属性制約 [Sakurai 08c] といった制約が存在する．本解説では，分析者の知見をより積極的に利用することにより，系列パターンの絞り込みを可能とする，制約パターン [Sakurai 08b] に基づいた方法を紹介する．

制約パターンは，属性値の変化に着目した制約条件であり，分析者にとって興味がある属性値の変化を，制約パターンとして入力する．系列パターンの発見法は，この制約パターンを参照することにより，制約パターンに合致する系列パターンだけを発見する．形式的には，m 個のアイテム集合で構成された m 次制約パターンは，式（6）のように記述することができる．特に，$m=1$ の場合には，この制約パターンを制約アイテム集合と呼ぶことにする．

$$\{C_1, C_2, \cdots, C_m\},$$
$$C_i = (x_1 : a_{i1},\ x_2 : a_{i2},\ \cdots,\ x_n : a_{in}) \qquad (6)$$

本式においては，C_i が同一の単位時間内に発生するアイテムの集合を表しており，$i < j$ の場合には，C_i は C_j よりも時系列的に後に現れることを表している．また，x_j（$j = 1, 2, \cdots, n$）は任意の属性を表しており，異なる x_j に対しては，異なる属性が割り当てられることを表している．このほか，a_{jk} は x_j に対応する属性の属性値を表している．

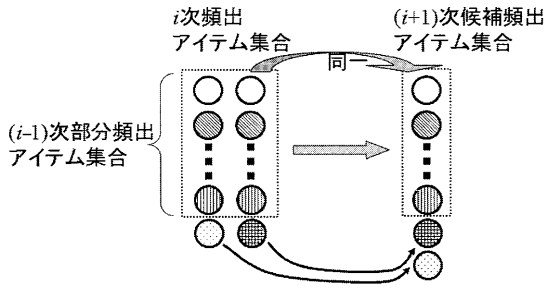


図3 候補アイテム集合の生成

例えば，表構造データの属性として，「血圧」，「脈拍」が与えられており，各属性値として，「正常」，「注意」，「異常」が与えられている場合を考えてみることにする．このとき，「血圧:正常」，「血圧:注意」，「血圧:異常」，「脈拍:正常」，「脈拍：注意」，「脈拍：異常」といったアイテムを考えることができる．このとき，$C_1 = (x_1$：正常，x_2：異常），$C_2 = (x_1$：異常，x_2：異常）といった制約パターンが与えられているとすれば，{(血圧:正常，脈拍:異常)，(血圧：異常，脈拍：異常)} といった部分系列パターンを含んだ系列パターンだけが，抽出されることになる．

制約パターンにおける属性値集合を構成する属性の数が少ない場合には，すべての属性の組合せに基づいて候補系列パターンを生成し，頻出しているかどうかを判定したとしても，それほど多くの時間を必要とせずに，制約パターンに一致する系列パターンを発見することができる．しかしながら，制約パターンに指定する属性の数が多い場合や，もともとの属性の数が多い場合には，その組合せは膨大なものとなる．この組合せは，属性の数に対して，指数関数的に増大していくため，制約条件を満たすすべての組合せを考慮した候補系列パターンを生成して，その頻度を評価する方法は，現実的な方法とはいえない．そこで，系列パターンの発見法における候補系列パターンの生成時に，与えられた制約パターンを考慮することにより，効率的に系列パターンを発見する方法を紹介する．

ここで，候補頻出アイテム集合の生成方法に着目してみると，図3に示すように，前方の $i-1$ 個のアイテムが一致する二つのアイテム集合に対して，残りの1個のアイテムをアイテム間の順序が保たれるように追加することにより，候補頻出アイテム集合を生成している．また，この候補頻出アイテム集合が頻出していると判定された場合に，当該候補頻出アイテム集合は頻出アイテム集合と判定される．

一方，新たに生成した頻出アイテム集合が制約アイ

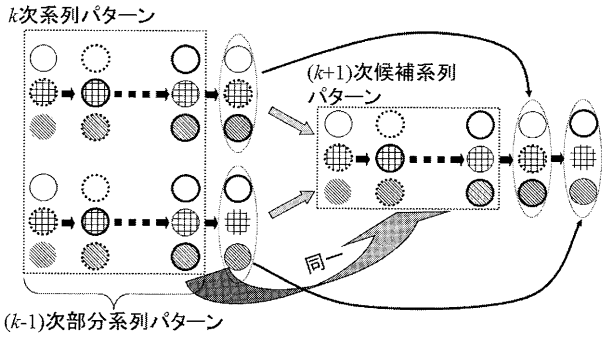


図4 候補系列パターンの生成

テム集合に一致しているとすれば，もとになった二つの頻出アイテム集合は部分制約アイテム集合に一致している．逆にいえば，部分制約アイテム集合に一致していない頻出アイテム集合を組み合わせて，候補頻出アイテム集合を生成したとしても，制約アイテム集合に一致する頻出アイテム集合を生成することはできない．このため，部分制約アイテム集合に一致しない頻出アイテム集合を生成する必要はない．したがって，候補アイテム集合の生成時に，部分制約アイテム集合に一致しているかどうかを判定することにより，不要な頻出アイテム集合の生成を回避することができる．

一方，候補系列パターンの生成方法に着目してみると，図4に示すように，前方の$k-1$個のアイテム集合が一致する二つの系列パターンに対して，残りの1個のアイテム集合を順に付加した系列を候補系列パターンとして生成している．また，この候補系列パターンが頻出していると判断された場合に，当該候補系列パターンは系列パターンと判定される．

したがって，候補頻出アイテム集合の生成の場合と同様に，候補系列パターンの生成時に，部分制約パターンに一致しているかどうかを判定することにより，不要な系列パターンの生成を回避することができる．

次に，与えられた制約パターンから部分制約アイテム集合および部分制約パターンを生成する方法を紹介する．ここで，候補頻出アイテム集合および候補系列パターンの生成に着目してみると，候補頻出アイテム集合のもとになった頻出アイテム集合への分解および候補系列パターンのもとになった系列パターンへの分解は，一意に決定することができる．また，必要となる部分制約アイテム集合および部分系列パターンは，このもとになる頻出アイテム集合および系列パターンに対して適用される制約条件になっている．このため，与えられた制約パターンに対して候補生成の逆操作となる分解を実施することにより，必要となる部分制約アイテム集合および部分制約パターンを一意に決定することができる．すなわち，式(6)の制約パターンが与えられている場合には，式(7)に示す二つの部分制約パターンを生成することができる．

$$(C_1, C_2, \cdots C_{m-2}, C_{m-1}),$$
$$(C_1, C_2, \cdots C_{m-2}, C_m) \qquad (7)$$

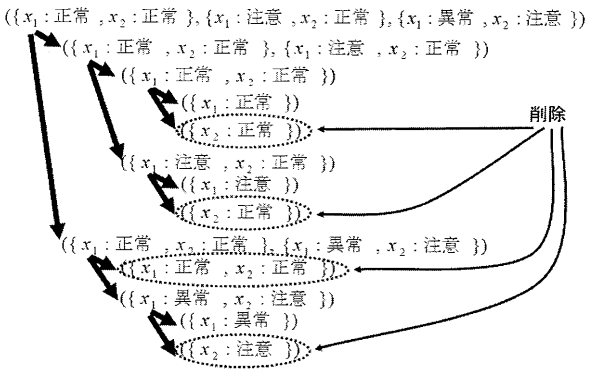


図5 制約パターンの分解

また，生成された部分制約パターンを新たな制約パターンとして，制約パターンの分解操作を部分制約パターンの長さが1になるまで繰り返すことにより，必要なすべての部分制約パターンを得ることができる．一方，生成された1次部分制約パターン（C_i）を制約アイテム集合とみなすことにより，式(8)に示す二つの部分制約アイテム集合を生成することができる．

$$\{x_1 : a_{i1}, \cdots, x_{n-2} : a_{in-2}, x_{n-1} : a_{in-1}\}$$
$$\{x_1 : a_{i1}, \cdots, x_{n-2} : a_{in-2}, x_n : a_{in}\} \qquad (8)$$

また，生成された部分制約アイテム集合を新たな制約アイテム集合とみなして，制約アイテム集合の分解操作を部分制約アイテム集合に含まれるアイテム数が1になるまで繰り返すことにより，必要なすべての部分制約アイテム集合を得ることができる．

図5は，3次制約パターン（{x_1：正常，x_2：正常}，{x_1：注意，x_2：正常}，{x_1：異常，x_2：注意}）が分解される様子を示している．制約アイテムの場合においては，同一の属性値となる制約アイテムは，一つだけを残して削除されることに注意する必要がある．また，別解の途中段階で同一の部分制約パターンが生成される場合には，一つの部分制約パターンだけを残して，残りの部分制約パターンは削除されることにも注意する必要がある．

5. 適　用　例

本章では，ここまでに紹介してきた系列興味度および制約条件に基づいた方法を，実データに適用した具体例を紹介する．

5·1 営業日報分析

本節では，Sales Force Automation（SFA）システムによって収集された，BtoB領域の営業員によって記述されている営業日報分析に，系列パターン発見技術を適用した例[櫻井 06]を紹介する．

図6は，対象とする営業日報の例を示しており，各営業日報は，時間情報，属性情報，テキスト情報といった要素から構成されている．

本営業日報に対して，図7の流れに従った処理を行う

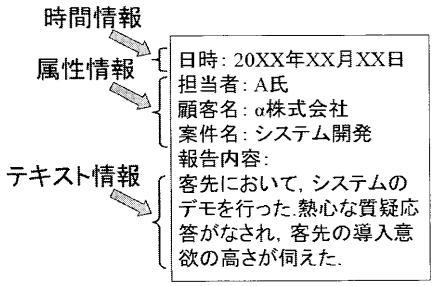


図6 営業日報の例

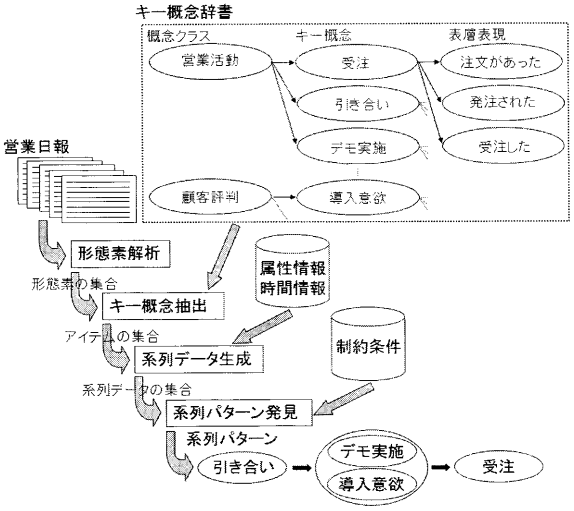


図7 営業日報からの系列パターン発見の流れ

ことにより，営業日報を特徴付ける系列パターンを発見する．本分析では，系列パターン発見の前処理として，「形態素解析」，「キー概念辞書抽出 [Ichimura 01]」，「系列データ生成」の各処理を実施することにより，系列データの生成を行っている．ここで，キー概念辞書は，概念クラス，キー概念，表層表現と呼ばれる3階層からなる辞書であり，表層表現には，実際に営業日報内で記載される表現が記述されている．同一の意味をもつ表層表現をまとめたものがキー概念となっており，系列データにおけるアイテムに対応している．一方，概念クラスは，類似の概念をもつキー概念をまとめたものとなっている．概念クラスは，営業日報に含まれる概念の数に関するOLAP分析には利用されているものの，系列パターンの発見においては，特に利用されてはいない．

営業日報からの系列パターンの発見では，形態素解析された営業日報のテキスト情報に対して，キー概念辞書を適用することにより，まずは，各営業日報を特徴付けるアイテム集合を抽出する．次に，このようにして抽出したアイテム集合を，営業日報に含まれる属性情報が同一のものでグルーピングし，グルーピングしたアイテム集合を，時間情報によって並び換えることにより，系列データを生成する．このような系列データを系列パターンの発見法に適用することにより，図7の最下部に示すような系列パターンを発見することができる．図の系列パターンの場合，顧客から「引き合い」があった後に，「デモ実施」したところ，顧客の高い「導入意欲」が伺え，その後に，「受注」に至ったという，受注に至るまでの成功の道筋が示されている．このような系列パターンを利用することにより，営業員が自身で抱えている案件のステータスを系列パターンと比較し，系列パターンに付随する営業日報を参照することにより，その後の営業活動の指針を得ることができる．

具体的には，五つの営業部門に導入されているSFAシステムによって収集された約28 000件の営業日報に対して，実際の分析を試みている．この際，案件が受注となったか失注となったかが，長くても半年で決まるといった背景知識により，始端アイテム，終端アイテム間の時間制約を導入している．また，何らかの「不評」が発生したにもかかわらず，その後に「受注」に至った場合に，分析者は興味があるとの想定のもと，「不評」，「受注」といった特定のアイテム間に対して，特定アイテム，特定アイテム間の時間制約を導入しており，この特定-特定間隔を種々変えることにより，系列パターンの発見を行っている．

このような発見に対して，時間制約を導入することにより，時間制約を導入しない場合と比べて，営業員が参照する必要のある営業日報を，70～90%程度，削減できることが確認されている．

5・2 定期健康診断データ分析

本節では，従業員の定期健康診断データから，系列パターンを発見する例 [櫻井 07] を紹介する．図8は，本データに基づいた分析のアウトラインを示している．図からわかるように，各年度における個別の従業員の検診結果を年度順に並べた後で，連続する年度における各従業員の各属性の値の変化（改善，安定，悪化）を抽出し，その変化を並べたものが系列データとなっている．検診結果の場合，検査項目が属性，その検査結果が属性値で与えられた表構造データとなっているため，属性と属性

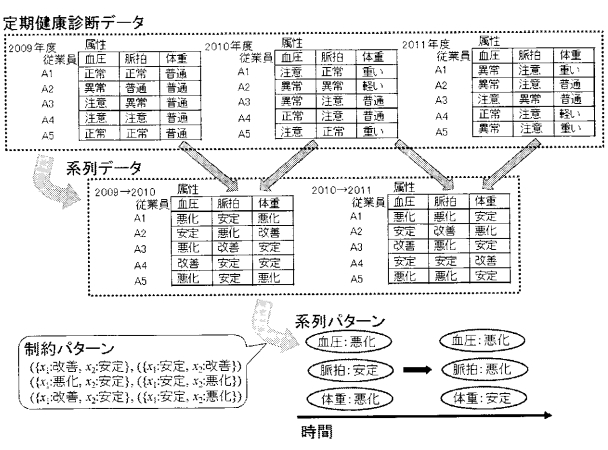


図8 定期健康診断データからのパターン発見

値の組合せがアイテムとなっている．また，同一の年度間の変化を，従業員ごとにまとめたものがアイテム集合となっている．このようなデータに対して，制約パターンを適用することにより，分析者の興味にあった系列パターンの絞り込みを行うことができる．

実データを用いた実験では，男性社員の検査結果約19万件を利用している．本データは，個人の特定ができないように暗号化されており，その暗号化キーは産業医によって管理されている．本データを20代，30代，40代，50才以上といった四つのデータ集合に分割し，年代ごとの四つの系列データの集合を生成する．制約パターンとしては，二つのアイテムの同時改善，1年遅れた改善，2年遅れた改善といった三つの制約パターンからなる制約up，改善を悪化に変えた三つの制約パターンからなる制約dw，二つのアイテムの逆方向の同時変化，1年遅れた逆方向変化，2年遅れた逆方向変化を表す五つの制約パターンからなる制約udを利用している．実験では，本制約パターンを導入した場合と導入しなかった場合で，発見される系列パターンの個数を，年代ごとの各時系列データに対して評価している．その結果，制約パターンを導入しなかった場合に発生していた，長い系列パターンの発見が記憶容量の関係で発見できなくなるといった問題点や，分析者が見きれないほど多数の系列パターンが発見されるといった問題点を回避できることが確認されている．発見された系列パターンは，検査結果に基づく，従業員の生活改善に役立てられることになる．

6. ま と め

本解説では，多様な系列データの中から，分析者にとって興味のある系列パターンを発見するための方法として，系列パターンの評価基準である系列興味度とその理論的な性質を紹介するとともに，分析者の背景知識を利用した枠組みとしての時間制約および制約パターンを紹介した．また，これら技術を適用した応用例として，営業日報分析，定期健康診断データ分析について紹介した．

本解説では，取り上げなかったが，数値系列データの分析や，系列データのリアルタイム分析，タイプの異なる複数の系列データの複合分析などに対しても，高い分析ニーズがある．このため，本分野は今後ますます，研究が進むものと考えられる．

◇ 参 考 文 献 ◇

[Agrawal 94] Agrawal, R. and Srikant, R.: Fast algorithms for mining association rules, *Proc. 20th Int. Conf. Very Large Data Bases*, pp. 487-499 (1994)

[Agrawal 95] Agrawal, R. and Srikant, R.: Mining sequential patterns, *Proc. 11th Int. Conf. Data Engineering*, pp. 3-14 (1995)

[Ayres 02] Ayres, J., Gehrke, J. E., Yiu, T. and Flannick, J.: Sequential pattern mining using bitmaps, *Proc. 8th Int. Conf. Knowledge Discovery and Data Mining*, pp. 429-435 (2002)

[Ichimura 01] Ichimura, Y., Nakayama, Y., Miyoshi, M., Akahane, T., Sekiguchi, T. and Fujiwara, Y.: Text mining system for analysis of a salesperson's daily reports, *Proc. Pacific Association for Computational Linguistics 2001*, pp. 127-135 (2001)

[Pei 01] 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, *Proc. 2001 Int. Conf. Data Engineering*, pp. 215-224 (2001)

[櫻井 06] 櫻井茂明，植野 研：時間情報の付随したテキストデータの分析法，知能と情報，Vol. 28, No. 2, pp. 290-298 (2006)

[櫻井 07] 櫻井茂明，北原洋一，折原良平：制約パターンに基づいた時系列パターンの発見法，第66回人工知能基本問題研究会予稿集，pp. 35-40 (2007)

[Sakurai 08a] Sakurai, S., Kitahara, Y. and Orihara, R.: A Sequential pattern mining method based on sequential interestingness, *Int. J. Computational Intelligence*, Vol. 4, No. 4, pp. 252-260 (2008)

[Sakurai 08b] Sakurai, S., Kitahara, Y. and Orihara, R.: Discovery of sequential patterns based on constraint patterns, *Int. J. Computational Intelligence*, Vol. 4, No. 4, pp. 275-281 (2008)

[Sakurai 08c] Sakurai, S., Kitahara, Y., Orihara, R., Iwata, K., Honda, N. and Hayashi, T.: Discovery of sequential patterns coinciding with analysts' interests, *J. Computers*, Vol. 3, No. 7, pp. 1-8 (2008)

[Sakurai 08d] Sakurai, S., Ueno, K. and Orihara, R.: Discovery of time series event patterns based on time constraints from textual data, *Int. J. Computational Intelligence*, Vol. 4, No. 2, pp. 144-151 (2008)

[Srikant 96] Srikant, R. and Agrawal, R.: Mining sequential patterns: Generalizations and performance improvements, *Proc. 5th Int. Conf. Extending Database Technology*, pp. 3-17 (1996)

[Zaki 01] Zaki, M. J.: SPADE: An efficient algorithm for mining frequent sequences, *Machine Learning*, Vol. 42, No. 1, pp. 31-60 (2001)

2011年12月12日 受理

著 者 紹 介

櫻井 茂明（正会員）

1991年東京理科大学理学研究科数学専攻修士課程修了．同年（株）東芝入社，ソフトウェアシステム技術研究所配属．1998年から2年間新情報処理開発機構つくば研究センタ出向．2000年（株）東芝帰任．2010年4月から，東芝ソリューション（株）IT技術研究所出向．現在，同研究所研究主務．2009年6月から，東京工業大学大学院総合理工学研究科連携教授を兼務．2004年技術士（情報工学部門）登録．博士（工学）．日本知能情報ファジィ学会理事（会誌担当）．