解 説

特徴的な系列パターンを発見するための技術†

櫻井 茂明*

1．はじめに

コンピュータ環境及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益々巨大化しており，Big Data と呼ばれるバズワードが生み出されている．多くの企業では，Big Data の利活用に取り組んでおり，その分析技術も急速に研究開発されている．

Big Data を構成するデータには多様なものが存在しているが，M2M から発信されるデータやライフログに関するデータが今後爆発的に増大することが予想されており，時系列的に収集されるデータが今後益々増大すると考えられている．データマイニング分野におけるこのようなデータの分析技術としては，離散的なアイテムやアイテム集合を時系列的に並べたデータから，系列パターンを発見する問題として発展してきている．当初の研究においては，リテール分野への適用がなされており，個々の商品を離散的なアイテムとみなした系列データからの系列パターン発見が試みられていた．近年においては，テキストを単語のつながりとみなしたデータ，定期的に実施される健康診断の健診値や問診結果を離散化したデータ，機械から定期的に発信されるログデータといった多様な系列データへの適用が試みられており，その適用対象は急速に広まっている．一方，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，系列パターンの頻出性以外の基準に基づいた，特徴的な系列パターンの発見も試みられている．加えて，対象とするデータをより詳細に表現するために，系列データや系列パターンそのものの拡張も行われている．

† Techniques for Discovery of Characteristic Sequential Patterns
Shigeaki SAKURAI
* （株）東芝クラウド＆ソリューション社ビックデータ・クラウドテクノロジーセンター
Big Data Cloud Technology Center, Toshiba Corporation Cloud & Solutions Company

このような背景の下，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，時間制約[Sakurai et al. 08]，アイテム間制約[櫻井 他 13]，結論部制約に基づいた方法を紹介し，その適用例として営業日報データの分析に関して紹介する．加えて，本解説で紹介した系列パターンの発見技術の関連研究を紹介する．

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．リテール分野の例を考えた場合，アイテムは個別の商品を表しており，アイテム集合が 1 枚のレシートに記述されている商品の一覧に対応している．このようなレシートを，顧客ごとに集めて，時間順に並べたものがひとつの系列データとなる．ただし，系列データにおいては，各アイテム集合には，同一のアイテムはせいぜいひとつしか含まれていないことが仮定されている．このため，先のレシートの例の場合，購入された商品の個数や金額に関する情報は考慮せずに，商品が購入されたかどうかだけに着目したデータになっている．

本解説では，このような系列データを多数集めた場合に，その中から，特徴的な部分系列を系列パターンとして発見する方法を紹介する．ここで，系列パターンが特徴的であるかどうかを評価するには，何らかの評価指標が必要となるが，その指標を算出するための基礎として，系列の含意関係に関してまずは説明する．

系列 s_1 が系列 s_2 を含意するとは，系列 s_2 を構成するすべてのアイテム集合が，その順序関係を保ちつつ，系列 s_1 を構成するいずれかのアイテム集合の部分集合になっていることと定義される．例えば，顧客Aの 3 日間における購入品履歴を示す系列が，式(1)のように与えられているとする．本系列においては，同じレシート内のアイテムが {}で括られており，→によって，次の日のレシートと区別されるように記述されている．すなわち，本系列は，顧客Aが 1 日目に「卵」，「バター」，「パン」を購入し，2 日目に「シリアル」，「牛乳」を購入し，3 日目に「ご飯」，「納豆」，

表1　購入品履歴

ID	系列
A	{ 卵, バター, パン } → { シリアル, 牛乳 } → { ご飯, 納豆, 卵 }
B	{ 卵, パン } → { シリアル, 牛乳 } → { 納豆 }
C	{ バター, パン } → { ご飯, 卵 }
D	{ 卵, バター, パン, シリアル } → { 牛乳 } → { 卵 }
E	{ 卵, ジャム, パン } → { シリアル, 牛乳 }
F	{ ご飯, 納豆, 卵 } → { シリアル, 牛乳 }

「卵」を購入したことを表している．

{卵，バター，パン}
→{シリアル，牛乳}→{ご飯，納豆，卵}　(1)

一方，他の顧客の購入品履歴が表1に示すように与えられているとする．このとき，顧客Bの系列（系列B）の1～3番目のアイテム集合は，それぞれ，顧客Aの系列（系列A）の1～3番目のアイテム集合の部分集合となっており，系列Bは系列Aに含意されているといえる．同様に，系列Cの1番目，2番目のアイテム集合は，それぞれ，系列Aの1番目，3番目のアイテム集合の部分集合となっており，系列Bは系列Aに含意されているといえる．このとき，系列Aの2番目のアイテム集合に対応するものが，系列Cには存在していない場合にも，含意関係が成立することに注意する必要がある．

これに対して，系列Dに着目してみると，1番目のアイテム集合は，構成するアイテム自体は系列Aに出現しているものの，1番目のアイテム集合を部分集合とするようなアイテム集合は，系列Aに存在してはいない．このため，系列Dは系列Aに含意されてはいない．同様に，系列Eにおいては，1番目のアイテム集合を構成するアイテムである「ジャム」が，系列Aには含まれていないため，系列Eは系列Aに含意されてはいない．また，系列Fは，1番目，2番目のアイテム集合を部分集合としてもつ，アイテム集合が系列Aに存在してはいるものの，3番目，2番目のアイテム集合であり，順序関係が系列Aとは異なっている．このため，系列Fは系列Aに含意されてはいないことになる．

上記に説明した系列の含意関係を利用することにより，系列パターンが特徴的であるかどうかを評価する基準として，式(2)及び式(3)で定義される支持度及び信頼度が定義されている．

$$\text{支持度}(s) = \frac{s\text{が含意される系列データ数}}{\text{系列データ数}} \qquad (2)$$

$$\text{信頼度}(s|s_p) = \frac{s\text{が含意される系列データ数}}{s_p\text{が含意される系列データ数}} \qquad (3)$$

ただし，sが系列パターンを表すとし，s_pが系列パターンsに含意される部分系列パターンを表すとする．以上の定義から分かるように，支持度は相対頻度を表しており，系列パターンの頻出性を評価する基準となっている．また，信頼度は，s_pが与えられた場合において，sが与えられる条件付き確率を表しており，s_pを前提部，sからs_pを除いた部分系列$s-s_p$を結論部とするルールの確からしさを表す基準となっている．

3．系列パターンの発見法

系列パターンの発見法としては，指定した最小支持度以上のすべての系列パターン（頻出系列パターン）を効率よく発見する複数の方法[Agrawal and Srikant 95，Ayres et al. 02，Pei et al. 01，Zaki 01]が提案されている．これらの方法では，系列パターンを構成するアイテムの数が増えるにしたがって，その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，効率のよい系列パターンの発見を実現している．発見された頻出系列パターンは，信頼度を利用してさらなる絞り込みを行うことにより，ある種の特徴的な系列パターンを発見することができる．このような2段階の方法がとられるのは，支持度がアプリオリ性を満たす一方，信頼度はアプリオリ性を満たさないためである．本解説では，頻出系列パターンの発見法のうち，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，すべての頻出系列パターンを発見する候補に基づいた方法[Agrawal and Srikant 95]を簡単に紹介する．

候補に基づいた方法では，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次頻出アイテム集合を組み合わせることにより，ふた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する．例えば，1次頻出アイテム集合として，{卵}，{バター}といったふたつの1次頻出アイテム集合が与えられている場合，これらを組み合わせることにより，式(4)に示す候補を生成する．

{卵，バター}　(4)

この候補の支持度が最小支持度以上となる場合に，{卵，バター}は2次頻出アイテム集合となる．

次に，2次頻出アイテム集合を組み合わせることにより，3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する．例えば，2次頻出アイテム集合として，{卵，バター}，{卵，パン}といったふたつの2次頻出アイテム集合が与えられている場合，これらを組み合わせることにより，式(5)に示す候補を生成する．

{卵，バター，パン}　　(5)

このとき，組み合わせるアイテム集合としては，最初のアイテム(卵)が共通しているものを組み合わせる必要があることに注意する必要がある．このような組み合わせを考慮するだけで，すべての3次頻出アイテム集合を発見することができるのは，支持度がアプリオリ性を満たすためである．この候補の支持度が最小支持度以上となっている場合に，{卵，バター，パン}は3次頻出アイテム集合となる．

一般には，前方の$(i-1)$個のアイテムが一致するふたつのi次頻出アイテム集合を組み合わせることにより，$(i+1)$個のアイテムで構成されるアイテム集合を候補として生成する．また，その候補の支持度が最小支持度以上となる場合に，$(i+1)$次頻出アイテム集合となる．図1は，$(i+1)$個のアイテムで構成される候補が生成される様子を示している．本図においては，各丸がひとつのアイテムを表している．また，各丸に付与されている模様が同じものが，同じアイテムを表している．

候補に基づいた方法では，以上のようなアイテム集合の成長を，$(i+1)$次頻出アイテム集合の個数が1個以下になるまで順次繰り返すことにより，すべての頻出アイテム集合(1次頻出系列パターン)を発見することができる．

このようにしてすべての頻出アイテム集合が発見されたとすれば，系列の延伸が行われることになる．具体的には，生成済みの頻出アイテム集合の中からふた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，ふたつの頻出アイテム集合が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．例えば，{卵}，{シリアル，牛乳}といったふたつの頻出アイテム集合が与えられているとすれば，式(6)及び式(7)に示す2種類の候補を生成することができる．

{卵}→{シリアル，牛乳}　　(6)
{シリアル，牛乳}→{卵}　　(7)

生成された候補はその支持度が算出され，最小支持度以上となる場合に，2次頻出系列パターンとなる．

すべての2次頻出系列パターンが発見されたとすれば，ふたつの2次頻出系列パターンを組み合わせることにより，3つのアイテム集合が系列的に並んだ3次候補系列パターンを生成する．このとき，ふたつの2次頻出系列パターンは，その前方の1次頻出系列パターンが一致するものを組み合わせる必要があることに注意する必要がある．例えば，{卵}→{シリアル，牛乳}，{卵}→{納豆}といったふたつの2次頻出系列パターンが与えられているとすれば，式(8)及び式(9)に示す2種類の候補を生成することができる．

{卵}→{シリアル，牛乳}→{納豆}　　(8)
{卵}→{納豆}→{シリアル，牛乳}　　(9)

一般には，前方の$(k-1)$個のアイテム集合が一致するふたつのk次頻出系列パターンを組み合わせることにより，$(k+1)$個のアイテム集合で構成される系列パターンを候補として生成する．また，その候補の支持度が最小支持度以上となる場合に，$(k+1)$次頻出系列パターンとなる．図2は，k次頻出系列パターンから$(k+1)$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，図1の場合と同様に，各丸がひとつのアイテムを表しており，同一の時刻に発生したと

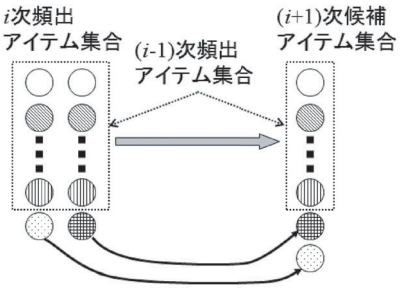


図1　候補アイテム集合の生成

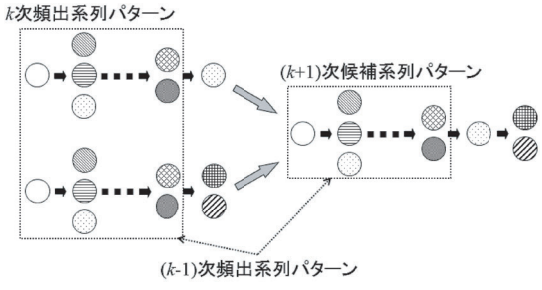


図2　候補系列パターンの生成

みなされるアイテムが矢印によって区切られている．以上のような系列の成長を，(k+1)次頻出系列パターンが 1 個も発見されなくなるまで繰り返すことにより，すべての系列パターンの発見を行う．ここで，アイテム集合の成長の場合には，同一のアイテム集合を組み合わせることができない反面，系列の成長においては，同一の系列の組み合わせができるため，成長の終了条件が異なっていることに注意する必要がある．

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はあるものの，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かった．このため，発見されるパターンが，必ずしも分析者に新たな知見を与えるものにはなっていなかった．また，信頼度によって系列パターンを絞り込むことができるものの，十分な絞り込みを行うにはやや力不足であった．そこで，本節では，系列パターンを様々な制約によって絞り込むことにより，より特徴的な系列パターンを発見するために研究開発されたいくつかの手法を紹介する．

4.1 時間制約

頻出系列パターンの発見においては，出現するアイテムの順序関係だけを考慮している．このため，時間的に掛け離れた，実際には意味の無いアイテムの並びも，系列パターンの頻度として積算しており，意味の無いアイテムの並びを含んだ系列パターンを，発見する危険性があった．この問題に対して，[Srikant and Agrawal 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，最小時間間隔及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，これらの時間間隔だけでは，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，本節では，柔軟な時間制約の導入を実現するひとつの方法として，[Sakurai et al. 08] に提案されている 7 つの時間制約を紹介する．図 3 は，本制約のイメージを表している．図においては，白抜きの各丸印が通常のアイテム，網掛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)の最小値及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)の最小値及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)の最小値及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)の最小値及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

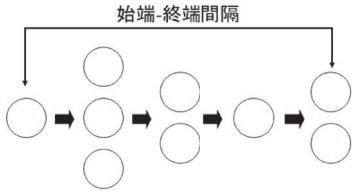
本制約は分析者が指定するふたつの特定のアイテム間に対して，その時間間隔(特定−特定間隔)の最小値及び最大値を指定する．これにより，特定−特定間隔が指定した最小値と最大値の間に含まれる系列パターンを抽出することができる．ただし，ひとつの系列パターンの中に特定のアイテムの組が複数存在する場合には，その出現順序に従ったアイテムの組だけを考えることにする．すなわち，2 種類のアイテム a，b が a，a，b，b の順に系列パターンに出現しているとすれば，最初の a と最初の b 及び 2 番目の a と 2 番目の b に対してのみ本制約を適用する．このような限定を置くことにより，循環するような部分系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

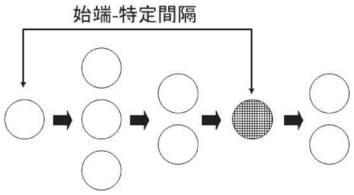
本制約は分析者が指定した特定のアイテムとその前方に隣接しているアイテムとの間に対して，その時間間隔（隣接−特定間隔)の最小値及び最大値を指定する．これにより，隣接-特定間隔が指定した最小値と最大値の間に含まれる系列パターンを抽出することができる．

(7)特定アイテム，隣接アイテム間の時間制約

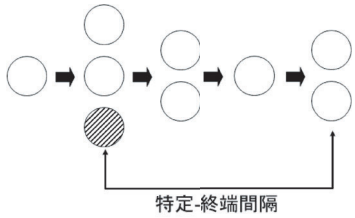
本制約は分析者が指定した特定のアイテムとその後



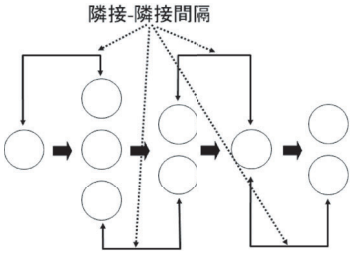
(1) 始端アイテム, 終端アイテム間の時間制約



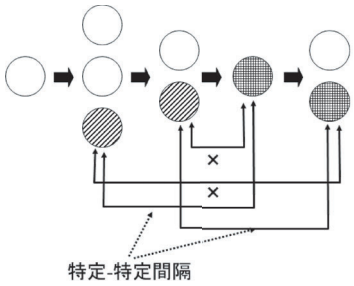
(2) 始端アイテム, 特定アイテム間の時間制約



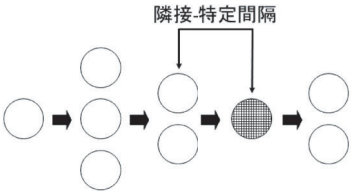
(3) 特定アイテム, 終端アイテム間の時間制約



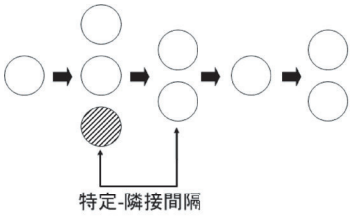
(4) 隣接アイテム, 隣接アイテム間の時間制約



(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 3 時間制約［櫻井 12］

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）の最小値及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る．すなわち，当該系列パターンが系列データに含意される場合に，当該系列パターンに関連するすべての時間制約が，系列データを参照することにより評価される．また，すべての時間制約が成立する場合に，当該系列パターンの頻度が 1 積算される．このような評価がすべての系列データに対して行われることにより，時間的に意味のある系列パターンだけを発見することができる．

4.2 アイテム間制約

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本問題における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，同じ属性を持つアイテム同士は，異なる属性を持つアイテム同士よりも，意味的に深い関係を持つといえる．この意味的な関係を制約として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できる．以下においては，属性と属性値から構成されるアイテムを考え，このアイテム間の関係を制約として指定するアイテム間制約[櫻井他 13]を紹介する．

アイテム間制約は，一般に式(10)のように記述することができる．

$$C_1 \rightarrow C_2 \rightarrow \cdots \rightarrow C_n \tag{10}$$

ただし，$n(>0)$ はアイテム間制約によって指定される系列の長さを表すとし，C_i は時系列的に i 番目に出現するアイテム集合を表すとする．C_i は一般に，$m_i(>0)$ 個のアイテムから構成されており，各アイテムは属性と属性値によって構成されている．従って，式(11) によって C_i を記述することができる．

$$C_i = \{A_{i1} : a_{i1},\ A_{i2} : a_{i2},\ \cdots,\ A_{im_i} : a_{im_i}\} \tag{11}$$

式(11) においては，A_{ij} 及び a_{ij} が属性及び属性値を表しており，$A_{ij} : a_{ij}$ によってアイテムが表現されている．アイテム間制約においては，属性及び属性値の両方に対して具体的な値を指定することも可能であるが，属性のみを指定したり，属性値のみを指定したりすることもできる．系列パターンの発見法は，頻出系列パターンに対して，指定されたアイテム間制約を適用することにより，制約を満たす頻出系列パターンを特徴的な系列パターンとして発見する．このとき，アイテム間制約は同時に複数指定することも可能であり，複数指定されている場合には，いずれかのアイテム間制約を満たす頻出系列パターンが，特徴的な系列パターンとして発見される．

例として，表 2 に示す 5 つの属性とその属性値からなる表構造データが日単位に収集されており，このような系列データから，アイテム間制約を利用して特徴的な系列パターンを発見する場合を考えてみることにする．ただし，本表においては，気温と湿度の場合における属性値「普通」と，人手と交通量の場合における「普通」とを区別するために，属性値「普通」に対して添え字が付与されている．このとき，式(12)に示すアイテム間制約が指定されているとすれば，ある日において，天気が晴れであり，気温に関してはいずれの属性値でもよく，その後の日において，人手あるいは交通量が多くなるような頻出系列パターンを，特徴的な系列パターンとして発見することができる．

表 2　表構造データ

属性	属性値
天気	晴れ, 雨, 曇り, 雪
気温	高い, 普通 $_1$, 低い
湿度	高い, 普通 $_1$, 低い
人手	多い, 普通 $_2$, 少ない
交通量	多い, 普通 $_2$, 少ない

$$\{\text{天気：晴れ，気温：}*\} \rightarrow \{* \text{：多い}\} \tag{12}$$

4.3 結論部制約

アイテム間制約は，アイテム間に対して柔軟に制約を指定することができるため，ある種の特徴的な系列パターンを発見することができる．また，アイテム間制約を満たさない系列データを事前に削除することができるため，系列パターンの発見に必要となる時間の削減に対しても一定の効果がある．しかしながら，複数のアイテム間制約を同時に指定した場合には，削除可能な系列データの数は限られたものとなり，時間削減効果は限定的なものとなる．

一方，アイテム間制約の特殊な例として，アイテム間制約の始端(C_1) や終端(C_n) が系列パターンにおいて，出現する位置を指定することもできる．系列パターンにおいては，終端より前に出現している部分系列パターンを前提部，終端を結論部とみなすことにより，発見された系列パターンを一種のルールとみなすことができる．また，収集された系列データが，その前提部と合致するかどうかを判断することにより，次に発生すると考えられるアイテム集合を予測することができる．このため，終端に出現するアイテム集合を指定したいとのニーズは比較的高いものになっている．ここで，例として，式(13)に示す系列パターンが発見されているとする．

$$\begin{aligned}&\{\text{天気：晴れ，気温：普通}_1\}\\ &\quad\rightarrow\{\text{人手：多い}\}\rightarrow\{\text{交通量：多い}\}\end{aligned} \tag{13}$$

このとき，現時点において，{天気：晴れ，気温：普通$_1$}→ {人手：多い} といった系列データが収集されているとすれば，次の時点において，{交通量：多い} が発生することを予測することができる．交通量の予測を行いたいといった観点では，終端に出現するアイテムを{交通量：*}と指定することにより，分析者は目的とする系列パターンを発見することができる．

終端に出現するアイテムやアイテム集合を指定した

場合，アイテム集合の成長や系列パターンの成長に制限をかけることができるため，系列パターンの発見をより効率よく行うことができる．以下においては，終端を固定した場合のアイテム間制約を特に結論部制約と呼ぶことにする．ここでは，結論部制約の最も単純な場合として，終端に特定のアイテムが指定されている場合を考えるとし，結論部制約の利用により，系列の成長に基づいた計算量が削減される効果について説明する．

1次頻出系列パターン全体を考えた場合，この中に，結論部制約によって指定されたアイテム(以下，終端アイテム)だけからなる1次頻出系列パターンが存在する必要がある．このようなパターンが存在しない場合には，1次頻出アイテム集合の発見段階で，頻出系列パターンの発見そのものが終了することになる．次に，2次頻出系列パターンについて考えてみると，結論部制約が指定されていない場合は，任意のふたつの1次頻出系列パターンを組み合わせて，2次候補系列パターンを生成する必要がある．これに対して，結論部制約が指定されている場合には，結論部に指定された終端アイテムが終端となる2次候補系列パターンだけを生成すれば十分である．なぜなら，頻出系列パターンの終端が終端アイテムとなっていない頻出系列パターンは，終端制約を満たす頻出系列パターンにならないばかりか，候補生成においても利用される可能性がないからである．ただし，この後に説明する系列パターンの後方延伸に基づいて，より高次の系列パターンを生成していくことにする．従って，結論部制約を利用することにより，2次候補系列パターンの個数を大幅に減らすことができ，頻出系列パターン発見における計算量を大幅に減らすことができる．図4は，結論部制約を指定した場合において，生成する必要がある2次候補系列パターンを示している．図においては，白抜きの丸が結論部制約によって指定される終端アイテムを表しており，生成する必要のない2次候補系列パターンに，バツ印が付けられている．

次に，2次以降の候補系列パターンを生成する後方延伸について説明する．従来の候補系列パターンの生成においては，最後のアイテム集合を除いた部分頻出系列パターンが一致するふたつの頻出系列パターンを組み合わせることにより，候補系列パターンの生成を行っていた．これに対して，後方延伸においては，最初のアイテム集合を除いた部分頻出系列パターンが一致するふたつの頻出系列パターンを組み合わせることにより，候補系列パターンの生成を実現している．このような後方延伸は，延伸する方向が変わるだけであり，すべての可能性のある候補を網羅することができる．このため，いままでの候補系列パターンの生成と等価な結果を得ることができる．一方，結論部制約が指定されている場合には，終端アイテムが決定されているため，後方延伸の対象となるのは頻出系列パターンの最後のアイテム集合が，終端アイテムに一致するものだけである．図5は結論部制約によって終端アイテムが指定されている場合における後方延伸の様子を示している．本図により，終端アイテムが頻出系列パターンの終端に含まれているのを確認することができる．

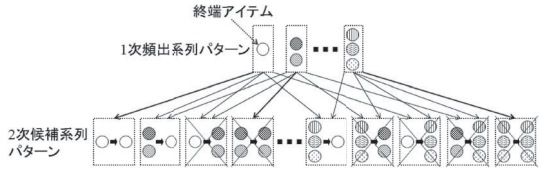


図4　削除対象系列パターン

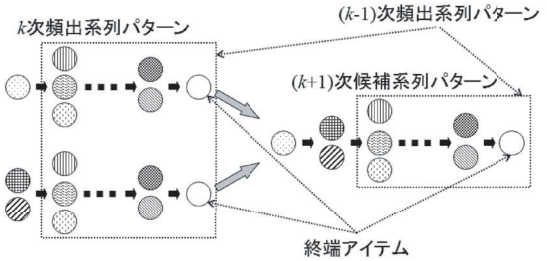


図5　系列パターンの後方延伸

5．営業日報データの分析

本節では，特徴的な系列パターンの発見法を，実データに適用した場合の例として，営業日報データを対象とした分析事例を簡単に紹介する．営業日報データの分析においては，Sales Force Automation(SFA)システムによって収集された営業日報を対象とした系列パターンの発見が試みられている[櫻井・植野 06]．

本データは，BtoB 領域の営業員によって記述された営業日報となっており，各営業日報には，日時，担当者，顧客名，案件名，報告内容といった内容が記述されている．これら営業日報に対して，形態素解析，キー概念抽出といった処理を実施することにより，キー概念をアイテムとする，系列データを生成する．ここで，形態素解析とは，報告内容に記載されている日本語文章を，品詞付きの単語集合に分解する処理である．また，キー概念抽出[Ichimura et al. 2001]とは，品詞付き単語やそれらの組み合わせが意味する内

容を，キー概念として取り出す処理である．営業日報データの場合，「受注」，「引き合い」，「デモ実施」，「導入意欲」のようなキー概念が報告内容から取り出されることになる．

営業日報からの系列パターンの発見では，ひとつの営業日報から抽出されるキー概念の集合がアイテム集合に相当しており，同一の担当者，顧客名，案件名で収集し，日時によって並べかえた営業日報から生成されるものが，ひとつの系列データに相当している．生成した多数の系列データを，系列パターンの発見法に適用することにより，式(14)に示すような系列パターンを発見する．

{引き合い}→{デモ実施, 導入意欲}→{受注}　　(14)

本式における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」に到ったという，受注に到るまでの成功の道筋が表現されている．このような系列パターンを活用することにより，営業員が自身の案件のステータスを系列パターンと比較するとともに，系列パターンの基になった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．

本分析事例では，５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，営業員にとって最も興味があると考えられる，「受注」や「失注」を結論とする系列パターンに着目している．また，案件が最終的に「受注」となったか「失注」になったかが，長くても半年で決まるといった背景知識を利用して，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．この他，何らかの問題が発生したにも関わらず，最終的に「受注」に到った場合に，分析者は興味があるとの想定の下，何らかの問題を示すキー概念である「不評」と「受注」との間に対して，特定アイテム，特定アイテム間の時間制約を導入している．これら制約を利用することにより，制約を導入しない場合に比べて，営業員にとって参照する必要がある営業日報を，70%から90%程度削減できることを確認している．

6．関連研究

本解説で対象としている系列データは，離散的なアイテムやアイテム集合を並べたものになっているが，株価，電気の使用量，室内・室外の温度といった，数値的なデータが時系列的に与えられるデータも多数存在している．このようなデータを系列パターンの発見問題として扱うためには，数値の系列データを事前に離散化する必要がある．数値の系列データを記号情報に変換する代表的な手法としては，[Lin et al. 03] に提案されているSAX（Symbolic Aggregate approXimation）法が知られている．SAX法では，数値の系列を指定した時間幅で区分された区間を平均値で代表することにより，区分的な数値の系列に変換し，この変換した数値を記号に変換する．このとき，正規化された数値系列データがGaussian 分布に従うことを仮定することにより，各記号の出現頻度が等しくなるように，各記号に変換される数値のしきい値を決定している．SAX法は，様々な拡張が行われており，[Lkhagava et al. 06] では，区分された区間における数値の最小値と最大値を利用して離散化を行う方法が提案されている．また，[Malinowski et al. 13]や[Zalewski et al. 12] では，区分された区間における変化の方向性を利用した離散化を行う方法が提案されいる．

一方，発見対象とする系列データや，発見する系列パターンを拡張する方法として，時間情報に着目した手法が提案されている．[Giannotti et al. 06] では，時間情報の付随した系列データを対象とすることにより，系列パターンを構成するアイテム集合間に，時間情報を付与した系列パターンを発見する方法を提案しており，[Vautier et al. 05] では，時間間隔に関する情報を付与した系列パターンを発見する方法を提案している．また，[Höppner 01] では，アイテムが出現する時間間隔を持った系列データから，Allen のインターバル論理[Allen 83] によって記述される，「前」，「後」，「出会う」などのアイテム間の関係をパターンとして発見する方法を提案している．加えて，[Jiang et al. 05] では，時間粒度の階層である年，月，日に基づいたカレンダー制約をファジィ理論によって拡張し，その制約に基づいて非周期的な相関ルールを発見する方法を提案しており，[Huang and Kao 05] では，量的なデータをファジィ化した系列データから，異なる時間に発生したトランザクションに含まれるアイテム間の関係を示すファジィinter-transaction ルールを発見する方法を提案している．

この他，時間情報によらない方向での拡張として，[Fiot et al. 08] では，特定のアイテムの量的な値を記録した系列から，ファジィ化した傾向データ系列を生成し，ファジィ系列パターンを発見する方法を提案している．また，[Fiot et al. 07] では，欠損値を含んだ不完全な系列データに着目し，欠損値に対して，各々の起こりそうな値の確率を考慮することにより，欠損しているアイテム集合における頻度を算出して，頻出系列パターンを発見する方法を提案している．加

えて，［Chiang et al. 04］は，バイオメディカル分野の文献を対象として，文献内の各単語に対して，品詞，遺伝子名，単語間関係を示すラベルを付与した系列データを生成し，その中から遺伝子間の関係を記述したファジィ系列ルールを発見する方法を提案している．

以上に簡単に紹介した関連研究以外にも精力的な研究開発［Tzvetkov et al. 05］が行われており，本分野における研究開発の広がりを感じることができる．

7．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特徴的な系列パターンを発見するための方法として，時間制約，アイテム間制約，結論部制約を紹介した．

本解説では取り上げなかったが，著者らのグループでは，系列パターン数が爆発することを回避するために，アイテムの多様性を維持しつつ，指定した最大パターン数以下の系列パターンを発見する，上位パターン制約に関する研究開発も進めており，近く発表する予定である．また，系列データのBig Data 解析に向けて，Apache^{TM} Hadoop$^{®}$ を用いた並列化も実現している．

一方，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，系列データと非系列データの組み合わせ分析などに対しても，高い分析ニーズがある．このように，多くの研究課題が山積しており，本分野は今後益々，研究開発が進むものと考えられる．

参 考 文 献

［Agrawal and Srikant 95］ R. Agrawal, R. Srikant: Mining Sequential Patterns, *Proc. of the 11th International Conference on Data Engineering*, pp.3－14（1995）

［Allen 83］ J. F. Allen: Maintaining Knowledge about Temporal Intervals, *Communications of the ACM*, vol.26, no.11, pp.832－843（1983）

［Ayres et al. 02］ J. Ayres, J. E. Gehrke, T. Yiu, J. Flannick: Sequential Pattern Mining using Bitmaps, *Proc. of the 8th International Conference on Knowledge Discovery and Data Mining*, pp.429－435（2002）

［Chiang et al. 04］ J.－H. Chiang, Z.－X. Yin, C.－Y. Chen: Discovering Gene－gene Relations from Fuzzy Sequential Sentence Patterns in Biomedical Literature, *Proc. of the 13th IEEE International Conference on Fuzzy Systems*, vol.2, pp.1165－1168（2004）

［Fiot et al. 07］ C. Fiot, A. Laurent, M. Teisseire: Approximate Sequential Patterns for Incomplete Sequence Database Mining, *Proc. of the 16th IEEE International Conference on Fuzzy Systems*, pp.1－6（2007）

［Fiot et al. 08］ C. Fiot, F. Masseglia, A. Laurent, M. Teisseire: TED and EVA: Expressing Temporal Tendencies among QuantitativeVariables using Fuzzy Sequential Patterns, *Proc. of the 17th IEEE International Conference on Fuzzy Systems*, pp.1861－1868（2008）

［Giannotti et al. 06］ F. Giannotti, M. Nanni, D. Pedreschi: Efficient Mining of Temporally Annotated Sequences, *Proc. of the 2006 SIAM International Conference on Data Mining*, pp.348－359（2006）

［Höppner 01］ F. Höppner: Discovery of Temporal Patterns－Learning Rules about the Qualitative Behaviour of Time Series, *Proc. of the 5th European Conference on Principles of Data Mining and Knowledge Discovery*, pp.192－203（2001）

［Huang and Kao 05］ Y.－P. Huang, L.－J. Kao: A Novel Approach to Mining Inter－transaction Fuzzy Association Rules from Stock Price Variation Data, *Proc. of the 14th IEEE International Conference on Fuzzy Systems*, pp.791－796（2005）

［Ichimura et al. 2001］ Y. Ichimura, Y. Nakayama, M. Miyoshi, T. Akahane,T. Sekiguchi,Y. Fujiwara: Text Mining System for Analysis of a Salesperson's Daily Reports, *Proc. of the Pacific Association for Computational Linguistics 2001*, pp.127－135（2001）

［Jiang et al. 05］ J.－Y. Jiang, W.－J. Lee, S.－J. Lee: Mining Calendar－based Asynchronous Periodical Association Rules with Fuzzy Calendar Constraints, *Proc. of the 14th IEEE International Conference on Fuzzy Systems*, pp.773－778（2005）

［Lin et al. 03］ J. Lin, E. Keogh, S. Lonarrdi, B. Chiu: A Symbolic Representation of Time Series, with Implications for Streaming Algorithms, *Proc. of the 8th ACM SIGMOD Workshop on ResearchIssues in Data Mining and Knowledge Discovery*, pp.2－11（2003）

［Lkhagava et al. 06］ B. Lkhagava,Y. Suzuki, K. Kawagoe: Extended SAX: Extension of Symbolic Aggregate Approximation for Financial Time Series Data Representation, *Proc. of the Data Engineering Workshop 2006*, 4A0－8（2006）

［Malinowski et al. 13］ S. Malinowski, T. Guyet, R. Quiniou, R. Tavenard: 1d－SAX: A Novel Symbolic Representation for Time Series, *Advances in Intelligent Data Analysis XII*, *Lecture Notes in Computer Science*, vol.8207, pp.273－284（2013）

［Pei et al. 01］ J. Pei, J. Han, B. Mortazavi－Asl, H. Pinto, Q. Chen, U. Dayal, M. Hsu: PrefixSpan: Mining Sequential Patterns Efficiently by Prefix－projected Pattern Growth, *Proc. of the 2001 International Conference on Data Engineering*, pp.215－224（2001）

［櫻井12］ 櫻井茂明：多様なデータに対する系列パターンマイニングの適用，人工知能学会誌，vol.27，no.2，pp.128－135（2012）

［櫻井他13］ 櫻井茂明，早川ルミ，岩崎秀樹：時系列パターン分析におけるアイテム間制約の効果，第27回人工知能学会全国大会予稿集，2C1－5（2013）

［櫻井・植野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）

［Sakurai et al. 08］ S. Sakurai, K. Ueno, R. Orihara: Discovery of Time Series Event Patterns based on Time Constraints from Textual Data, *International Journal of Computational*

Intelligence, vol.4, no.2, pp.144－151（2008）
［Srikant and Agrawal 96］ R. Srikant, R. Agrawal: Mining Sequential Patterns: Generalizations and Performance Improvements, *Proc. of the 5th International Conference on Extending Database Technology*, pp.3－17（1996）
［Tzvetkov et al. 05］ P. Tzvetkov, X. Yan, J. Han: TSP: Mining Top k Closed Sequential Patterns, *Knowledge and Information Systems*, vol.7, issue 4, pp.438－457（2005）
［Vautier et al. 05］ A. Vautier, M.－O. Cordier, R. Quiniou: An Inductive Database for Mining Temporal Patterns in Event Sequences, *Proc. of the 2005 ECML/PKDD Workshop on Mining Spatial and Temporal Data*, pp.1640－1641（2005）
［Zaki 01］ M. J. Zaki: SPADE: An Efficient Algorithm for Mining Frequent Sequences, *Machine Learning*, vol.42, no.1, pp.31－60（2001）
［Zalewski et al. 12］ W. Zalewski, F. Silva, H. D. Lee, A. G. Maletzke, F. C. Wu: Time Series Discretization based on the Approximation of the Local Slope Information, *Proc. of the 13th Ibero-American Conference on AI*, *Lecture Notes in Computer Science*, vol.7637, pp.91－100（2012）
（2014年6月12日　受付）

［問い合わせ先］
〒212-8585　神奈川県川崎市幸区堀川町72-34
（株）東芝クラウド＆ソリューション社ビックデータ・クラウドテクノロジーセンター
櫻井　茂明
TEL：044-331-1255
E-mail：shigeaki.sakurai@toshiba.co.jp

著 者 紹 介

櫻井（さくらい）　茂明（しげあき）［非会員］

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5月から約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．2013年東芝ソリューション(株)IT技術研究開発センター転籍．2014年4月(株)東芝クラウド＆ソリューション社ビックデータ・クラウドテクノロジーセンター出向．現在，同センター主査．2009年6月から，約4年間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)．人工知能学会，電子情報通信学会各会員．