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

櫻井　茂明† 　植野　研† 　酢山　明弘† 　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,ken.ueno,akihiro.suyama,ryohei.orihara}@toshiba.co.jp

あらまし　時間情報を持ったテキストデータからイベントを抽出し，特徴的なイベントの並びを注目時系列イベントパターンとして発見する方法を提案した．本手法におけるイベントには，イベント間に時間間隔が存在しており，連続するイベントであったとしても，そのイベントの並びには必ずしも意味があるとはいえない．このため，従来法では意味のない注目時系列イベントパターンを発見する危険性があった．本論文では，新たに 7 種類の時間制約を導入することにより，時間間隔を柔軟に考慮した注目時系列イベントパターンを発見する方法を提案する．また，その効果を SFA システムから得られる営業日報の分析タスクに適用し検証する．

キーワード　テキストマイニング，時系列イベントパターン，時間制約

Introduction of Time Constraints to Time Series Event Pattern Mining Method

Shigeaki SAKURAI†, Ken UENO†, Akihiro SUYAMA†, and Ryohei ORIHARA†

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

Abstract　We have proposed a method that extracts events from textual data with time information and discovers characteristic time series event patterns. The method has a problem that might even extract non-characteristic rows of events because the time constraints are only available for the intervals between contiguous events. In this paper, we introduce seven flexible time interval constraints that enable to find meaningful time series event patterns, and verify the effectiveness on discovering the patterns from textual data of daily business reports collected by our SFA system.

Key words　Text Mining, Time Series Event Patterns, Time Constraints

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．このような時系列データを扱う方法として，論文 [3] [11] はセンサーデータを対象とした数値的な時系列データを扱う方法を提案している．一方，テキスト的な時系列データを扱う方法として，論文 [5] 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抽出し，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] 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 [10] はテキストに含まれる名詞句及び固有名詞に基づいてテキストを特徴付け，指定した時間に含まれるテキストとそうで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う方法を提案している．

これに対して，我々は新たな観点での時系列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特徴的なイベントの並びを注目時系列イベントパターンとして発見する方法を提案した [7] [8] [9]．提案法では，頻出するイベントの並びから特徴的なイベントの並びへと絞り込む方法として，分析者が特定のイベントの並びを部分的に指定する枠組を用意している．このため，分析者は自身の分析意図に合った注目時系列イベントパターンを比較的容易に発見することができる．しかしながら，従来法では，イベントの連続性のみに着目しているため，時間的には掛け離れたイベントであり，実際には意味のない連続するイベント間の関係をも発見する危険性があった．

そこで，本論文では，イベント間に柔軟に時間制約を導入し，時間的に掛け離れて意味のないイベント間の関係を削除することにより，注目時系列イベントパターンをより容易に発見する

方法を提案する．また，その効果を SFA システムから得られた営業日報の分析タスクに適用し検証する．

2. 時系列イベントパターンマイニング

時系列イベントパターンマイニング法では，図 1 に示すようなひとつの時間情報，複数の属性情報，ひとつのテキスト情報からなる時間情報を持ったテキストデータを入力データとする．図 1 においては，日報を記述した日時が時間情報であり，日報を記述した担当者，日報で対象としている顧客名や案件名が属性情報に対応する．また，テキストで記述された報告内容がテキスト情報に対応する．

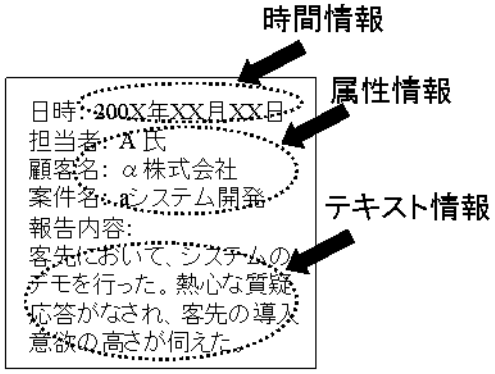


図 1　時間情報を持ったテキストデータ

この時間情報を持ったテキストデータの集合から，時系列イベントパターンマイニング法は，特徴的なイベントの並びを時系列イベントパターンとして発見する．すなわち,「引き合い」,「デモ実施」,「導入意欲」,「内定/受注」といったイベントからなる時系列イベントパターンを考えた場合，図 2 に示すような時系列イベントパターンを発見する．図 2 の例においては，矢印で区切られたイベントあるいはイベントの集合が 3 つの異なる時間で，左から順に発生することを表している．すなわち,「引き合い」の後に,「デモ実施」と「導入意欲」が同時に発生し，その後で,「内定/受注」に至ったことを表している．ここで，以下においては，時系列イベントパターンを構成する異なる時間に発生するイベントあるいはイベントの集合を要素と呼ぶ．図 2 の例の場合，要素のサイズは 3 と与えられている．

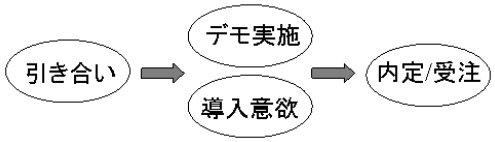


図 2　時系列イベントパターン

当該時系列イベントパターンの発見は，図 3 に示す処理の流れに従って行われる．以下においては各処理を簡単に説明する．

形態素解析及び引き続くキー概念抽出では，各テキストデータのテキスト情報からテキストデータを特徴付けるイベントの集合を抽出する．このとき，キー概念抽出では，概念クラス，キー概念，表層表現と呼ばれる 3 層構造からなるキー概念辞書 [2] を参照して，イベントの抽出を行う．ただし，表層表現はテキストデータ中に記述されている表現を正規表現を用いて記述しており，同一の意味を有する表層表現をキー概念としてまとめている．また，関連するキー概念を概念クラスとしてまとめている．この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テキストデータから抽出される．

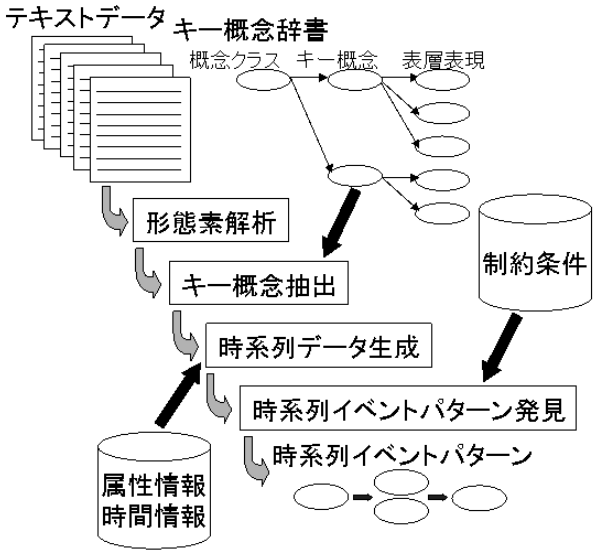


図 3　時系列イベントパターン発見の流れ

時系列データ生成では，属性情報の値が一致するテキストデータをまとめることによりグループを生成する．次に，グループごとに各テキストデータからイベントの集合を取り出して，時間情報とイベント集合とを組とする要素を生成する．このとき，同一の時間情報を持ったテキストデータが複数存在する場合には，複数のテキストデータに対応するイベントの集合を併せたイベントの集合を時間情報に対応するイベントの集合とする．また，ひとつの時間情報に対して，複数回同一のイベントが抽出される場合には，ひとつのイベントだけを残して，残りのイベントを削除する．この要素を時間情報に基づいて時間順に並べ替えることにより，グループごとにひとつの時系列データを生成する．

時系列イベントパターンの発見では，論文 [1] に提案されている AprioriAll をベースとした方法に基づいて，時系列イベントパターンを発見する．すなわち，最初に，式 (1) に定義される支持度が指定した最小支持度 (MinSup) 以上となるイベント及びイベントの組合せを取り出して，1 次頻出時系列イベントパターン (L_1) を生成する．

$$Sup(s) = \frac{\text{時系列イベントパターン } s \text{ を含む時系列データ数}}{\text{時系列データ数}} \quad (1)$$

次に，L_1 のふたつの頻出時系列イベントパターンを基にして，サイズが 2 となる候補時系列イベントパターンを生成し，最小支持度以上となる候補時系列イベントパターンを 2 次頻出時系列イベントパターン L_2 として生成する．最終的には，L_2 の他の頻出時系列イベントパターンに含まれている頻出時系列イベントパターン及び制約条件として指定した部分時系列イベントパターンを含んでいない頻出時系列イベントパターンを取り除

いて，注目時系列イベントパターンとして出力する．この頻出時系列イベントパターンの取り出しから注目時系列イベントパターンの出力までの一連の処理を頻出時系列イベントパターンのサイズをひとつずつ大きくしながら，$(k-1)$ 次頻出時系列イベントパターン (L_{k-1}) が空になるまで繰り返し，すべての注目時系列イベントパターンを発見する．ここで，注意しなければならないのは，指定した部分時系列イベントパターンを含む時系列イベントパターンが，部分時系列イベントパターンの一部しか含まない，ふたつの $(k-1)$ 次頻出時系列イベントパターンから生成される可能性がある点である．このため，指定した部分時系列イベントパターンを含まない $(k-1)$ 次頻出時系列イベントパターンを保持する必要がある．従って，時系列イベントパターン発見アルゴリズムの擬似コードは図 4 に示すように与えられる．図 4 においては，SeqDB が時系列データの集合，check_freq_seq() が最小支持度以上かどうかを判定する関数，subseq() が時系列イベントパターンの最後の要素を取り除いた部分時系列イベントパターンを抽出する関数，$\bowtie_{seq}$ がふたつの時系列イベントパターンからサイズがひとつ大きな候補時系列イベントパターンを生成する演算，check_inclusion() が L_k の他の時系列イベントパターンに含まれているかどうかを判定する関数，check_subsequence() が制約条件として与えられる部分時系列イベントパターンを含んでいるかどうかを判定する関数，CondList が制約条件として指定する部分時系列イベントパターンのリストを表している．また，本発見法における候補時系列イベントパターンの生成に対応する，subseq() を利用した条件判定から $\bowtie_{seq}$ の実施までのイメージを図 5 に示す．図 5 においては，各丸印がひとつのイベントを表している．

```
MinSup = UserDefinedValue;
L_1 = φ;
For each element el ∈ es, es ∈ SeqDB
   If check_freq_seq(el, SeqDB, MinSup, 1)
   Then add el to L_1;
L_k = φ;
For(k=2; L_{k-1} != φ; k++)
   For each sequence es1 ∈ L_{k-1}
      For each sequence es2 ∈ L_{k-1}
         If subseq(es1, k-2)==subseq(es2, k-2)
         Then cs = es1 ⋈_seq es2;
         add cs to L_k;
   For each sequence cs ∈ L_k
      If check_freq_seq(cs, SeqDB, MinSup, k)
      Else delete cs from L_k;
   For each sequence cs ∈ L_k
      If !check_inclusion(cs, L_k)
         If check_subsequence(cs, CondList)
            output cs;
```

図 4　注目時系列イベントパターン発見法

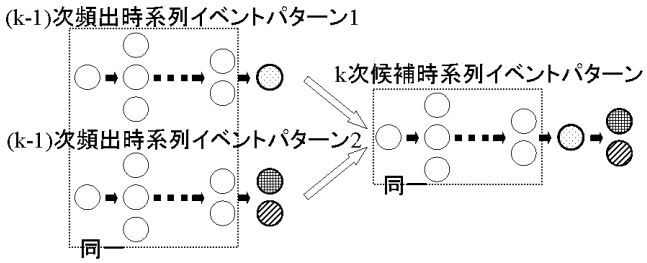


図 5　候補時系列イベントパターンの生成

3. 時間制約の導入

従来の時系列イベントパターンマイニング法では，分析者の意図に合った注目時系列イベントパターンを抽出するため，指定した部分時系列イベントパターンを含んだ時系列イベントパターンを抽出する枠組を提供していた．このため，小さな支持度を設定したとしても，多くの頻出時系列イベントパターンの抽出を回避することができ，注目時系列イベントパターンを容易に発見することができる．しかしながら，頻出時系列イベントパターンの発見においては，出現するイベントの順序関係だけを考慮しているため，時間的に掛け離れた，実際には意味のないイベントの並びも時系列イベントパターンの頻度としてカウントすることになる．このため，意味のないイベントの並びを含んだ時系列イベントパターンを注目時系列イベントパターンとして抽出する危険性があった．この問題に対して，論文 [6] では，min-gap 及び max-gap を導入することにより，隣接するイベントが指定された時間間隔に収まることを目的とした時間制約を導入している．本時間制約により，関連のないイベントの組合せを含んだ時系列イベントパターンを削減できると期待できるものの，イベント間の時間的関係は必ずしも一意的に指定できるとは限らない．すなわち，イベントの組み合わせごとに時間間隔が満たすべき条件は異なっている可能性がある．また，隣接していないイベント間に時間制約が存在することも考えられる．そこで，以下に示す 7 つの時間制約を導入することにより，イベント間の時間間隔をより柔軟に考慮できるようにする．また，本時間制約を時系列イベントパターンマイニング法に組み込むことにより，分析者の意図に合った注目時系列イベントパターンをより容易に発見できるようにする．

(1) 始端イベント，終端イベント間の時間制約:

本制約は時系列イベントパターン全体に関する時間間隔の最小値及び最大値を指定し，最小値と最大値の間に含まれる時系列イベントパターンを抽出する．すなわち，時系列データが対象とする時系列イベントパターンを含んでいる場合に，当該時系列データにおける時系列イベントパターン全体の時間間隔を計算する．また，本制約を満たすかどうかを検査し，本制約が満たされる場合に，対象とする時系列イベントパターンの頻度を 1 積算する．本制約のイメージを図示すると図 6 のように与えられる．

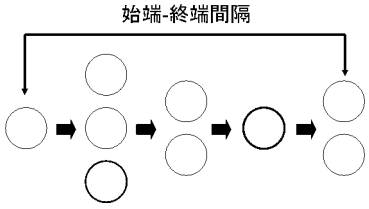


図 6　始端イベント，終端イベント間の時間制約

(2) 始端イベント，特定イベント間の時間制約:

本制約は時系列イベントパターンの最初のイベントから分析者が指定する特定のイベントまでの時間間隔の最小値及び最大値を指定し，最小値と最大値の間に含まれる時系列イベントパターンを抽出する．本制約のイメージを図示すると図 7 のように与えられる．図 7 においては，網掛けされた丸印が分析者の指定する特定のイベントを表している．

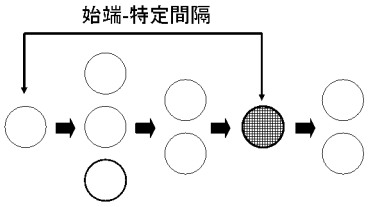


図 7　始端イベント，特定イベント間の時間制約

(3) 特定イベント，終端イベント間の時間制約:

本制約は分析者が指定する特定のイベントから時系列イベントパターンの最後のイベントまでの時間間隔の最小値及び最大値を指定し，最小値と最大値の間に含まれる時系列イベントパターンを抽出する．本制約のイメージを図示すると図 8 のように与えられる．

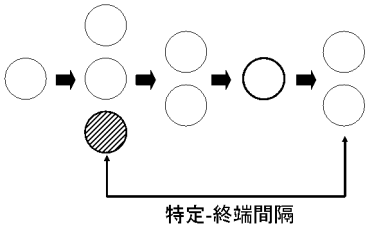


図 8　特定イベント，終端イベント間の時間制約

(4) 隣接イベント，隣接イベント間の時間制約:

本制約は，論文 [6] に提案されている min-gap と max-gap に対応した時間制約であり，任意の隣接するイベント間に対してその時間間隔の最小値及び最大値を指定する．本制約では，隣接するイベントが指定した最小値と最大値の間に含まれる時系列イベントパターンを抽出する．本制約のイメージを図示すると図 9 のように与えられる．

(5) 特定イベント，特定イベント間の時間制約:

本制約は分析者が指定するふたつの特定のイベント間に対してその時間間隔の最小値及び最大値を指定し，最小値と最大値

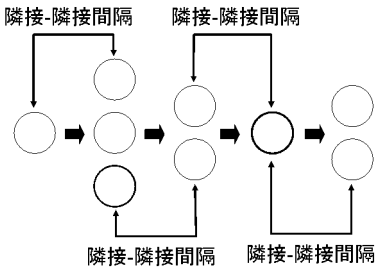


図 9　隣接イベント，隣接イベント間の時間制約

の間に含まれる時系列イベントパターンを抽出する．だだし，ひとつの時系列イベントパターンの中に特定のイベントの組が複数存在する場合には，その出現順序に従ってイベントの組を考えることにし，他の組合せに関しては本制約を考えないことにする．このように制約を実施することにより，時系列イベントパターンの中で循環するような部分時系列イベントパターンを含んだ時系列イベントパターンを抽出することができる．本制約のイメージを図示すると図 10 のように与えられる．図 10 においては，網掛けされた 2 種類の丸印が分析者の指定する 2 種類の特定のイベントを表しており，バツ印が付されたイベント間に関しては，制約が考慮されていないことを表している．

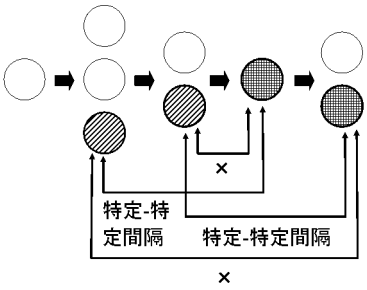


図 10　特定イベント，特定イベント間の時間制約

(6) 隣接イベント，特定イベント間の時間制約:

本制約は分析者が指定した特定のイベントとその前側に隣接しているイベント間に対してその時間間隔の最小値及び最大値を指定し，最小値と最大値の間に含まれる時系列イベントパターンを抽出する．本制約のイメージを図示すると図 11 のように与えられる．

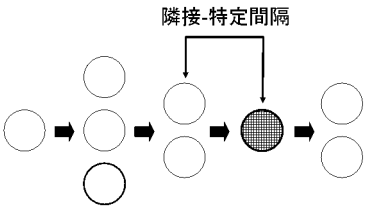


図 11　隣接イベント，特定イベント間の時間制約

(7) 特定イベント，隣接イベント間の時間制約:

本制約は分析者が指定した特定のイベントとその後側に隣接しているイベント間に対してその時間間隔の最小値及び最大

値を指定し，最小値と最大値の間に含まれる時系列イベントパターンを抽出する．本制約のイメージを図示すると図 12 のように与えられる．

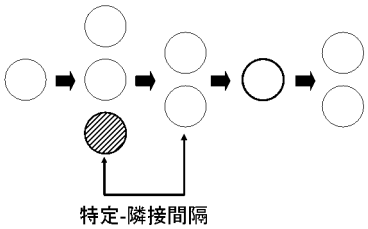


図 12　特定イベント，隣接イベント間の時間制約

上記に導入する 7 種類の時間制約を判定する処理を加えた注目時系列イベントパターン発見アルゴリズムの擬似コードは，図 4 に示す擬似コードの check_freq_seq() を check_freq_seq_time() に変更することにより構成することができる．ただし，check_freq_seq_time() は 7 種類の時間制約を考慮して，最小支持度以上かどうかを判定する関数とする．このとき，指定した時間制約が成立する時系列イベントパターンの部分時系列イベントパターンは，必ずしも指定した時間制約を満たしていない点に注意する必要がある．このため，候補時系列イベントパターンに対して，系列データベースに立ち返ってその頻度を再計算する必要がある．

本時間制約を導入した時系列イベントパターンマイニング法は，時間的に意味のある範囲にあるイベントに関して，その並びに従った時系列イベントパターンを発見することができる．このため，従来の時系列イベントパターンマイニング法よりも注目時系列イベントパターンを容易に発見することが期待できる．また，7 種類の時間制約を用意することにより，特別の知見のないイベントに関しては時間制約を導入しなくてもよいため，分析者は柔軟に時間制約を指定でき，時間制約を導入したとしても注目時系列イベントパターンの取りこぼしを少なくすることが期待できる．

4. 数値実験

4.1 実験データ

(株) 東芝社内の 5 つの営業部門 (A Dept. ～ E Dept.) 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，活動日が時間情報，活動内容がテキスト情報，その他の項目が属性情報に対応している．このような SFA データに対して，顧客名，案件名が一致するテキストデータごとにグループ化を行い，時系列データを生成する．本実験データの場合，表 1 にその内訳を示すような 6,434 件の時系列データを生成することができる．表 1 においては，系列，要素，概念，日報は，それぞれ，時系列データの数，時系列データを構成する要素の数，テキストマイニングの結果抽出されるキー概念の数 (イベントの数)，SFA データに含まれる日報の数を表している．

表 1　時系列データ関連情報

部門名	系列	要素	概念	日報
A Dept.	416	818	1,101	849
B Dept.	1,302	3,612	4,601	4,116
C Dept.	334	951	2,072	1,018
D Dept.	3,984	17,056	44,778	19,737
E Dept.	398	1,812	4,581	2,011

ただし，キー概念抽出で利用するキー概念辞書は，概念クラス数が 3，キー概念数が 61，表層表現数が 835 からなる規模の辞書を利用する．このため，時系列データにおいては，61 種類のイベントが存在する．

4.2 実験方法

各部門の時系列データに対して，最小支持度を 2.0%，1.0%，0.75%，0.5%と順次変えるとともに，概念クラスが「不評」であるイベントが発生した後で，イベント「内定/受注」が発生する 23 種類の部分時系列イベントパターンのいずれかを含む時系列イベントパターンを抽出する．また，これらの条件に加えて，表 2 に示す 6 種類の時間制約のいずれかを導入し，注目時系列イベントパターンを抽出する．本時間制約により，時系列イベントパターン全体の時間間隔を考慮することができるとともに，概念クラスが「不評」であるイベントとイベント「内定/受注」との間の時間間隔を考慮することができる．このため，実際には関係のない「不評」と「内定/受注」との間の関係に対応した注目時系列イベントパターンを削除することが期待できる．また，tc4～tc6 の実験結果を比較することにより，時間間隔を変更したことによる注目時系列イベントパターンの傾向の変化を観測することも期待できる．表 2 においては,「始端」は始端イベント,「終端」は終端イベント,「不評.*」は概念クラスが「不評」となる特定イベント,「内定/受注」は「内定/受注」となる特定イベントを示しており，時間間隔は日を単位としている．

表 2　時間制約

ID	制約内容
tc1	始端-終端間隔 ∈ [0,180]
tc2	tc1 ∩ 不評.*-内定/受注間隔 ∈ [0,90]
tc3	tc1 ∩ 不評.*-内定/受注間隔 ∈ [0,60]
tc4	tc1 ∩ 不評.*-内定/受注間隔 ∈ [0,30]
tc5	tc1 ∩ 不評.*-内定/受注間隔 ∈ [31,60]
tc6	tc1 ∩ 不評.*-内定/受注間隔 ∈ [61,90]

4.3 実験結果

図 13 に実験結果の一部を示す．図 13 においては，各部門のデータ及び最小支持度に対して，発見される時系列イベントパターンの個数が，系列サイズに対して推移する様子を示している．図 13 においては，縦軸が展開ルール数を表しており，横軸が系列のサイズを表している．ただし，展開ルール数とは，発見された時系列イベントパターンに含まれるすべての部分集合をひとつとして計算した場合の時系列イベントパターンの個数である．例えば，イベント集合 ab，イベント集合 cd が連続した (ab)(cd) といったサイズが 2 となる時系列イベントパターンの展

開ルール数を考えた場合，ab の部分集合として a，b が得られ，cd の部分集合として c，d が得られるので，その数は 9(=3× 3) と与えられる．また，条件無し，0∼ ∞，0∼90，0∼60，0∼30 が，それぞれ時間制約無し，時間制約 tc1，tc2，tc3，tc4 に対応した結果を示している．時間制約 tc5，tc6 に関連した結果は紙面の都合上掲載していない．

発見される時系列イベントパターンの妥当性をより正確に評価するには，発見された時系列イベントパターンに関連するテキストを，テキスト間の時間的な発生順序も考慮した上で評価する必要がある．しかしながら，時系列イベントパターンと関連テキストの組み合わせは膨大なものとなるため，すべての組み合わせを評価することは到底できない．そこで，本論文では展開ルール数の削減率による労力の低下を中心に評価し，内容的な妥当性に関しては，A. Dept，最小支持度 1.0%，サイズ 2 となる場合の時系列イベントパターンをサンプルとして取り上げて評価する．

4.4 考　　察

(1) 時系列イベントパターン:

概念クラスが「不評」となるイベントの後に,「内定/受注」に関するイベントが発生する時系列イベントパターンでは，当該イベント間に時間制約を導入することにより，抽出する時系列イベントパターンを大幅に削減することができる．また，実験結果としては示していないものの，時系列イベントパターンの基になった関連テキストの数も大幅に削減することができる．このため，分析者はより少ない労力で，発見された時系列イベントパターンの妥当性を確認することができる．一方，時間制約を導入したとしても，論文 [8] で発見した 3 種類の有用な関連テキストの系列を含む時系列イベントパターンが抽出されており，妥当な時間制約が導入されたといえる．このため，妥当な時間制約を導入することにより，分析者にとって有用な時系列イベントパターンを取りこぼしなく，効率よく発見することができたといえる．

(2) 時間間隔の扱い:

従来の時系列イベントパターンマイニング法 [9] では，イベントとイベントの間に時間間隔を表現するイベントを挿入することにより，時間間隔を考慮した時系列イベントパターンを抽出しており，時系列データにおいて隣接するイベントに対してしか時間間隔を考慮することができなかった．また，論文 [6] に導入されている時間制約を導入したとしても，時系列イベントパターンにおいて隣接するイベントの時間間隔しか考慮することができなかった．一方，時間制約を導入したい概念クラス「不評」に対応するイベントとイベント「内定/受注」との間には，複数のイベントが挿入される可能性がある．また，そのような複数のイベントを挟んだ時系列イベントパターンを抽出したい，他のイベント間に関しては特に時間間隔を考慮したくないとの要望があった．このため，論文 [6] の時間制約だけでは，隣接するイベント間にかなり緩い時間制約しか導入できず，発見される時系列イベントパターンをあまり削減することは期待できなかった．これに対して，今回導入した時間制約では，特定のイベント間に対して，時間間隔を指定することができる．このため，より厳しい時間制約を導入することができ，発見される時系列イベントパターンを削減することができた．本時間制約の導入に一定の効果があったといえる．

このような効果に加えて，時間制約 tc4，tc5，tc6 の条件に示すように，各条件に対応する時系列イベントパターンを比較することにより，時間間隔の違いに応じて，抽出される時系列イベントパターンが推移する様子を観測することができる．今回の実験では，残念ながら，時間制約 tc5，tc6 に関連した時系列イベントパターンの中に，それ程有用と思われる時系列イベントパターンは含まれておらず，時間間隔の違いによる時系列イベントパターンの推移の重要性を十分には検証できなかった．しかしながら，タスクによってはこのような時間間隔による違いを分析することも重要になると考えている．

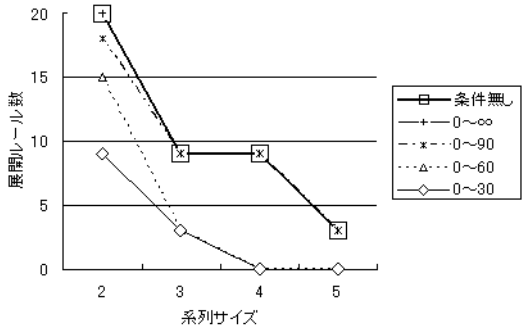
(3) 系列サイズ:

A Dept. の場合を除いて，時間制約の導入の有無に係わらず，展開ルールの数は系列サイズが大きくなるにつれて一旦増加し，その後に減少するといった傾向を示している．この傾向は，系列サイズが長くなるに従って時系列イベントパターンに含まれるイベントの組合せが多くなる一方，長い系列サイズを持つ時系列イベントパターンを含む時系列データの数が少なくなる結果と考えることができる．現在の時系列イベントパターン発見法の実装では，一度の延伸で生成される時系列イベントパターンの数が全時系列イベントパターンを発見するためのボトルネックとなっている．このため，一度の延伸で生成できる時系列イベントパターンの個数を少なくすることができれば，より小さな支持度が与えられる場合にも，全時系列イベントパターンを発見することが可能になると考えている．

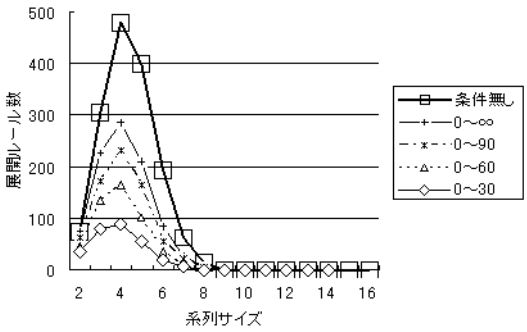
(4) 時間制約の妥当性:

本論文で導入した 7 つの時間制約は実験に利用した SFA データ及び現在分析作業を進めている健診データへの適用を考慮して導入されたものである．SFA データの場合，最終的には「内定/受注」か「失注」のいずれかのイベントに到達するため，終端イベントとその前に発生するイベントとの間に時間制約を導入する必要がある．また，ひとつの時系列データで売り込みから内定/受注あるいは失注までの系列が構成されるのが理想ではあるが，複数の系列がひとつの時系列データとしてグループ化されることがある．このため，あまりにも長い期間に渡る時系列データは複数の時系列データから構成されている可能性が高い．従って，時系列イベントパターンの始端から終端までの時間制約を導入する必要がある．さらには,「不評」といった評価が与えられた場合には，その評価を覆すために，適切な時期に，適切な対策を実施することが担当者にとっては重要である．このため，特定のイベントとその直後のイベントとの間に時間制約を導入する必要がある．

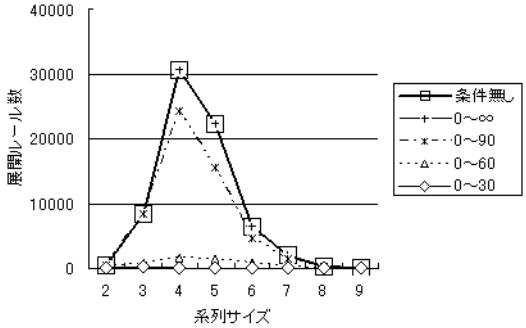
これに加えて，健診データの場合には,「薬の服用」→「血圧の低下」等のイベントの組が周期的に系列データに発生することがある．このため，イベントの対応関係を明らかにして時間



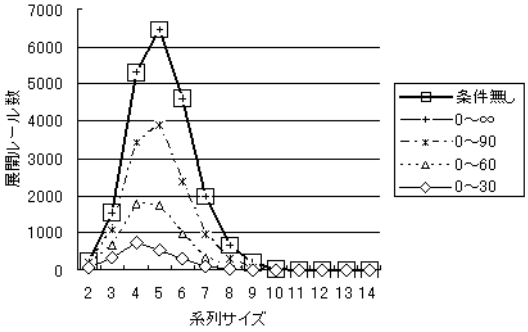
(a) A Dept., support=0.5%



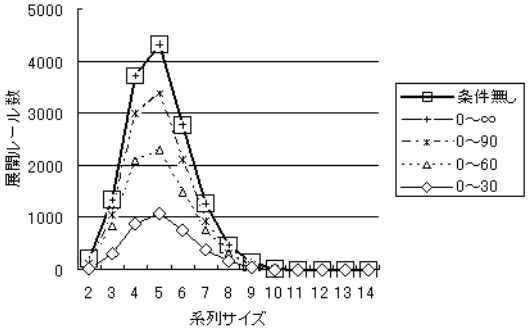
(b) B Dept., support=0.5%



(c) C Dept., support=0.75%



(d) D Dept., support=1.0%



(e) E Dept., support=2.0%

図 13　展開ルール数の変化 (時間間隔幅限定)

制約を導入する必要があり，特定イベント間に時間制約を導入する必要がある．このように，提案する時間制約は，検討したデータにとって必要な時間制約を網羅したものになっている．しかしながら，他のデータを対象とした場合には，現行の時間制約では上手く扱えない状況が発生する可能性は残されている．今後，他のデータへの適用を通して必要な時間制約が取りこぼされていないか見極めていく必要がある．

(5) 時間制約の導入方針:

時間制約を導入する方法は対象とするデータに依存する部分も多く，一般的に記述することはかなり難しい．分析データ及びその分析意図に従って時間制約を導入する必要がある．SFA データの分析タスクの場合，グループ化されたデータの中に複数の時系列データが含まれている状況が確認されており，はじめに始端イベントと終端イベントの間に時間制約を導入している．また,「不評」という判断がなされたにも係わらず，最終的に内定/受注に至った時系列イベントパターンに分析者の興味があった．このため，時間的に関連がある「不評 → 内定/受注」を含む時系列イベントパターンに絞り込むため，不評と内定/受注に時間制約を導入している．このような分析法方針は，営業活動が付随するデータを分析する分析タスクに関しては，活用することができると考えている．しかしながら，一般化を行うには分析例が不足しているため，他のデータへの適用を通して今後一般化を試みていきたい．

(6) 特定イベントの指定:

改良された注目時系列イベントパターン発見アルゴリズムでは，分析者が特定イベントを手動で指定する必要がある．今のところ，分析者は何らかの目的を持って分析していることを前提としているため，その目的にあった特定イベントを指定することはそれ程困難ではないと考えている．しかしがら，全く予想もしなかった関係を発見する上では特定イベントの発見を支援する機能も必要と考えられる．現在，発見された時系列イベントパターンを視覚化するとともに，GUI を通して制約条件を変えたり，関連するテキストを閲覧したりする支援システムを構築している．本システムのひとつの機能として，何らかの基準に基づいて特定イベントの候補を提示する機能を今後検討していきたい．

5. まとめと今後の課題

今回の論文では，より特徴的な注目時系列イベントパターンを発見しやすくするため，従来の時系列イベントパターンマイニング法に，新たに時間制約により時系列イベントパターンを絞り込む枠組を導入した．この時間制約により，分析者は自身の意図に合った注目時系列イベントパターンをより少ない労力で発見することが期待できる．

今後の課題としては，制約といった分析者の知見を参考とした注目時系列イベントパターンの発見に加えて，支持度とは異なる観点で時系列イベントパターンが特徴的かどうかを判断する指標を検討する予定である．また，従来の離散的なイベントに加えて，連続的なイベントを扱えるように時系列イベントパターン発見法を拡張していく予定である．この他，今回適用した SFA データとは異なる分野から得られる時系列データに適用し，その効果を検証していく予定である．

文　献

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

[2] 市村 由美 他,「営業日報を対象としたテキストマイニング -成功事例及び機会損失情報の抽出-」, 第 14 回人工知能学会全国大会, 532-534 (2000).

[3] 倉橋 節也, 寺野 隆雄,「学習分類子システムを用いたプロセス時系列からのデータマイニング」, 第 56 回知識ベースシステム研究会 (SIG-KBS) 合同研究会 (2002).

[4] V. Lavrenko et al., “Mining of Concurrent Text and Time-Series”, Proc. of the KDD-2000 Workshop on Text Mining (2000).

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

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

[7] 櫻井 茂明, 植野 研,「テキストデータからの時系列パターンの発見」, 第 21 回ファジィワークショップ, 4-2 (2003).

[8] S. Sakurai and K. Ueno, “Analysis of Daily Business Reports Based on Sequential Text Mining Method”, Proc. of the SMC2004, 3279-3284 (2004).

[9] 植野 研 他,「時間間隔を考慮した日報からの系列パターン抽出」, 第 3 回情報科学技術フォーラム (FIT2004) 一般講演論文集第 2 分冊, 339-340 (2004).

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

[11] 山田 悠 他,「時系列決定木による分類学習」, 第 17 回人工知能学会全国大会論文集, 1F5-06 (2003).