# 利用者の好みに基づくニュースストリームマイニング

清水　一宏†　　三浦　孝夫†

† 法政大学 工学研究科 電気工学専攻　〒 184–8584 東京都小金井市梶野町 3-7-2
E-mail: †shimizu@gs-eng.hosei.ac.jp, ††miurat@k.hosei.ac.jp

**あらまし**　頻出な選言パターン抽出は，ニュースストリーム環境下において，損失数え挙げ (Lossy Counting) に基づくオンライン型単一パスアルゴリズムを考えることができ，精緻なデータストリームマイニング手法として知られている．本稿では上記のアルゴリズムを拡張し，頻度だけでなく利用者の好み情報を用いることで，対象となる選言パターンに対して重み付けを行い，それらのパターンを含むデータ構造構築のための動的なパラメタを推定する手法を提案する．また実験によりその有用性を検証する．

**キーワード**　ニュースストリーム，選言パターン，興味パターン，オンライン型単一パスアルゴリズム

# News Stream Mining Based on User Preference

Kazuhiro SHIMIZU† and Takao MIURA†

† Dept.of Elect.& Elect. Engr., HOSEI University　3-7-2, KajinoCho, Koganei, Tokyo, 184–8584 Japan
E-mail: †shimizu@gs-eng.hosei.ac.jp, ††miurat@k.hosei.ac.jp

**Abstract**　Extraction of frequent *disjunctive patterns* is known to be a sophisticated method of data stream mining by which we can discuss an *online and single-pass algorithm* based on *Lossy Counting* under *news stream* environment. In this work, we propose a new weighting method for disjunctive patterns based on *user preference*. Also, we discuss dynamic parameter estimation for constructing a data structure which contains weighted disjunctive patterns. And we discuss some experimental results.

**Key words**　News Stream, Disjunctive Sequential Pattern, Interesting Pattern, Online and Single-Pass Algorithm

## 1. 前　書　き

これまで，文書データから重要な語句や規則を抽出するために，統計的推定や，データマイニング手法などの定量的な分析を適用することは容易ではないと言われてきた [4], [7]．しかし，テキストを対象とした研究が注目されるにつれ，この手法を系列データ (sequence data)，特に文字列テキストに適用するテキストマイニングに関する研究が開始されている [13].

一方，近年のネットワークや Web 技術の発達により，ネットワーク上を流れる膨大な量の動的なデータであるデータストリーム (data stream) に対するデータマイニング研究が注目されている [5]．一般にデータストリームは，(1) 大量，(2) 高速，(3) データ分布の動的変化，(4) 連続的という特徴を持っている．これを対象とする手法には，限られた計算資源で高速かつ長期間の解析が要求され，任意の時点において正しい解が出力されるような手法であることが望ましい．しかし，過去のデータマイニング手法を適用するのは困難であると言われている [5], [8].

データストリームの中でも特に，時系列順のニュース記事や放送内容のアノテーション・トランスクリプションの配信を，ニュースストリーム (news stream) と呼ぶ．本稿ではこのニュースストリームを対象とする．一般に，ニュース記事の発行間隔は不定であるが，発行される記事は速報性が高いため，本稿では記事の発行時間と内容時間は一致すると仮定する．

ニュースストリーム環境下では，絶えず記事が発生しており，新しい記事と過去の記事が混在している．利用者は，新しい記事の内容に興味を持っていることが多く，興味のあるテーマに関してニュース記事が集中的かつ爆発的に生じる．ニュースストリームの特性として両データを同等の重みで扱うことはできず，新しい記事に生じる語句ほど重要である．

APRIORI など静的なデータを対象としたデータマイニングの主要な研究では，探索空間の削減にパターン頻度の逆単調性が利用されるが，系列パターンでは成り立たない [2], [3]．著者らは**選言** (disjunctive) パターンを導入し，この上で逆単調性を満たす出現尺度を提案した [9].

利用者が興味あるパターンを発見する場合，検出結果を見ながら設定条件を変化させるというプロセスを繰り返すことが多い．各ステップは効率よく処理できても，全体としてデータを繰り返し走査するため，膨大な読込みが生じ，最終結果を得るまでに多くの時間を要する．また，各処理が独立であるため，設定条件を変更するたびに再計算が必要となる．

これらの問題を回避するため，オンライン分析手法が提案されている [1]．特にデータストリーム環境下では，過去に遡った再計算が困難であるので，情報を集約した予備計算結果を利用するオンライン分析の考え方が重要となる．これまで筆者らは，オンライン分析に基づき，静的なデータ環境下で効率的な選言パターン抽出法を提案し [14]，またデータストリーム環境でこの考え方を一般化した手法を提案している [10].

オンライン分析を用いた手法では、効率面での問題は解決できるが，パラメタを与えて計算した予備計算結果中に、利用者が興味を持つパターンが含まれているか否かは，実際の抽出を行うまでわからない．APRIORI に基づく手法の多くでは，頻度が重要度を表す．しかし，高頻度なパターンの多くは利用者にとって自明であることが多く，利用者が興味を持つパターンであると評価するのは困難であることから別の興味基準が検討されている [6].

本稿では，利用者の好みをパターンの抽出結果に反映し，頻度だけに依らない選言パターン抽出方法を提案する．具体的には，利用者が興味を持つパターンの分布と抽出パターンの分布の類似度に基づき重み付けを行い，パラメタの動的推定や，利用者の好みを反映したパターンを含むデータ構造を構築する．

本論の構成は以下の通りである．2 章で関連研究を紹介し，3 章では選言パターンとその逆単調性を満たす出現尺度を示す．4 章で利用者の好みを反映したデータの構築法やパラメタの推定法，及びそのデータからの選言パターン抽出法の提案を行う．5 章で実験結果を示し，6 章で結びとする．

## 2. 関 連 研 究

自明な高頻度パターンを結果から取り除くため，安部ら [11] は最適パターン発見手法を提案している．ここでは，利用者が興味を持つ文書集合 (目的群) と，平均的な文書集合 (対照群) をデータマイニングシステムへの入力とし，目的群で高頻度かつ対照群で低頻度に出現するようなパターンを興味あるものとする．この枠組みでは，中頻度であっても対象内の頻度が低ければ興味あるパターンとして抽出できる．一方で，静的なデータ環境下を対象としているため，利用者の興味や頻出となるパターンの変化等，時間経過を考慮する問題は想定されていない．

また，利用者の興味の評価方法として，サブパターンとスーパーパターンそれぞれの推定頻度を用いる sub+super 法が提案されている [15]．これは，対象となるパターンが包含するサブパターンと，包含されるスーパーパターンの両方から対象頻度を推定し，対象パターンの実際の頻度が推定頻度より大きくなる程，興味深いと判断する．この手法は，共起性の高いアイテム群を含むデータベースにおいて特に有効であるが，サブパターンとスーパーパターンの推定頻度も含めた計算に時間を要するため，高速計算が要求される動的データを扱う環境への適用は難しい．

## 3. 選言パターン (Disjunctive Pattern)

与えられたアルファベット (あるいはアイテムとも言う) の集合 $I = \{i_1, .., i_K\}, K > 0$ に対し，**系列データ** (記事)$S$ とはアルファベットの順序リスト $\langle s_1 ... s_m \rangle, \infty > m > 0$ をいう．$S$ には同じアルファベットが複数回生じてよい．

**選択**とは，アルファベットの集合 $\{a_1 a_2 ... a_m\}, m > 1$ であり，$[a_1 a_2 ... a_m]$ と表される．各 $a_j$ は相異なるアルファベットで，その順序は考慮しない．**選言パターン** (あるいは単にパターン) $p$ とは順序列 $t_1 t_2 ... t_n$ で表され各 $t_i$ はアルファベット $a$ であり，$t_i ... t_{i+m}$ は対応するアルファベット $a_1 a_2 ... a_m$，もしくは選択 $[a_1 a_2 ... a_m]$ である．パターン $p = t_1 t_2 ... t_n$，$q = v_1 v_2 ... v_m, m \leqq n$ があるとき，$q$ が $p$ の**部分パターン** である ($q \sqsubseteq p$ と記す) とは，$1 \leqq j \leqq k < ... < m \leqq l \leqq n$ を満たす整数 $j, k, l$ があり，各 $v_k$ は $t_j$ に対応して次を満たす：$v_k$ がアルファベット $a$ のとき， $t_j = a$ もしくは選択中のアルファベット $a$ ．$v_k ... v_{k+m}$ が 選択 $[a_1 a_2 ... a_m]$ のとき，$t_j ... t_{j+l} = [b_1 b_2 ... b_l]$ でかつ $\{a_1, .., a_m\} \subseteq \{b_1, .., b_l\}$.

［例 1］ “ac” は “abcd” の部分パターンである．同様に，“[ac]” は “[abcd]” の，“bd” は “[ab]b[cd]de” の，“b” は “[ab]” の，そして “ac”は “[ab][cd]” の部分パターンである．しかし，“ab” は “[ab]” の部分パターンではない．また “[ab]” は “ab” の部分パターンではない．

系列データ $S$ に関してパターンから非負整数への関数 $\mathcal{M}$ が**逆単調性** (Anti Monotonicity, AM) を満たすとは，パターン $p, q$ に対して $q \sqsubseteq p$ のとき $\mathcal{M}_S(q) \geqq \mathcal{M}_S(p)$ が成り立つときを言う．この性質を用いれば，部分パターンの探索範囲を減少させることができる．本稿では逆単調性を満たすパターン出現尺度を用いる [9].

系列データ $S = s_1 s_2 ... s_r$，パターン $p$ を $t_1 t_2 ... t_n$ とすると系列先頭頻度は以下のように表される．

$$H(S, p) = \sum_{i=1}^{r} Val(S, i, p) \tag{1}$$

ただし $Val(S, i, p)$ は次を満たすとき 1, さもなければ 0 であるとする: $S(i)$ を $S$ の $i$ 番目からの接尾辞 $s_i ... s_r$ とする．$t_1$ がアルファベット $a$ のとき，$s_i = a$ かつ $t_2 t_3 ... t_n$ が $S(i+1)$ に適合する．$t_1 ... t_m$ が選択 $[a_1 a_2 ... a_m]$ のとき，$s_i = a_j$ となる整数 $j \leqq m$ があり (例えば $j = 1$)，$[a_2 a_3 ... a_m] t_{m+1} ... t_n$ が $S(i+1)$ に適合する．

$H(S, p)$ は $S$ またはその接尾辞において $p$ が先頭から適合する回数を表しているが，逆単調性を満たさない．そこで， $p$ の全部分パターン $q$ を調べ，その $H(S, q)$ のうちの最小の値を全体出現頻度とする．実際には，$p$ の接尾辞 ($n$ 個存在) のうち，長さの小さいものから順にその最小値を決定すればよい．

$$D(S, p) = \mathrm{MIN}\{H(S, p), D(S, p(2))\} \tag{2}$$

ただし，$p(2)$ は $p$ より 1 つ長さの短い部分パターンである．このとき，$D(S, p)$ は逆単調性を示すことが知られている [9].

ここで，系列先頭頻度について $S$ の長さを考慮した重みを与える．ここでは，各頻出語の出現頻度として索引語頻度 (term frequency) を用いる [12]．すなわち，系列データ $S$ における，アイテム数 $n$ のパターン $p$ に対する，重みを考慮した系列先頭頻度 $H_w(S, p)$ は，次で定義される．

$$H_w(S,p) = \frac{H(S,p)}{|S| - (n-1)} \quad (3)$$

ただし，$|S|$ は $S$ 中の総単語数を表す．従って，全体出現頻度も以下のように書き直すことができる．

$$D_w(S,p) = \mathrm{MIN}\{H_w(S,p), D_w(S,p(2))\} \quad (4)$$

本稿では以降，この重み付けられた系列先頭頻度 $H_w(S,p)$ および全体出現頻度 $D_w(S,p)$ を用いる．

［例 2］ $S$ を cabcbba, $p = ab$ とすれば $H(S,p) = D(S,p) = 1$, $H_w(S,p) = D_w(S,p) = 0.17$ である．また $p = [ac]$ とすれば $H(S,p) = 3$, $D(S,p) = 2$, $H_w(S,p) = 0.5$, $D_w(S,p) = 0.29$ である．

## 4. 提案手法

### 4.1 問題定義

本稿では，ニューススストリームからの選言パターン抽出問題を次のように定義する．

> **入力**: ニューススストリーム $NS$, 最小支持度 $\sigma$, 許容誤差 $\epsilon$
> **問題**: 現在時刻 $time_{now}$ において，出現頻度が $\sigma$ 以上となるような全ての選言パターンを出力せよ．

ニューススストリーム $NS$ とは，系列データを要素とする無限長順序リスト $\langle S_1 ... S_i ... S_j ... \rangle$ であり，$S_i$ とは，$S_1$ を基準として $NS$ 中の第 $i$ 番目の系列データである．また，各 $S_i$ はそれぞれ発行時刻 $time_i$ を持ち，$i \leqq j$ ならば $time_i \leqq time_j$ である．本研究では，発行時刻 $time_i$ が等しい複数の系列データ $S_{i_1}, ..., S_{i_{wsize}}$ は，直列に結合する単一の系列データ $S_i$ とみなす．ここで，$wsize$ をウインドウサイズと呼ぶ．$wsize$ は上限値を有すると仮定し，一度の計算で同時に処理できる系列データ数と考える．許容誤差 $\epsilon$ とは，頻度誤差の許容値を示しており，本稿ではデータ構造を構築する際のパラメタとして利用する．現在時刻 $time_{now}$ は問い合わせが行われた時刻を表す．本稿では最小支持度 $\sigma$ 以上の頻度を持つパターンを**頻出パターン** (frequent pattern) と呼ぶ．

### 4.2 手法の概要

本稿では，5 つのフェーズからなるオンライン型単一パスアルゴリズムを提案する．

（1） 系列データからの候補パターン集合生成
（2） 候補パターンへの好みに基づく重み付け
（3） 最適許容誤差の推定
（4） 候補パターン集合からの選言パターン束構築
（5） 選言パターン束からの頻出パターン集合抽出

このアルゴリズムは，ニューススストリームから系列データが入力されるごとに (1) から (4) を繰り返し実行する．利用者からの任意の問合せに対しては (5) を実行し，その時点での結果を出力する．

(1) では，系列データから，アイテム数 $m$ で，許容誤差の初期値として与えられる初期許容誤差 $\epsilon_0$ 以上の頻度を持つ $m$ -**候補パターン** (candidate pattern) の生成を行う．

(2) では，利用者の好み情報に基づくパターンへの重み付けを処理する．利用者は正負それぞれの**興味パターン** (のリスト) を提示する．リスト中の各パターンと，候補パターンの分布を比較し，正もしくは負の重みを与えることで，利用者の好みを反映させる．正負どちらの興味パターンにも類似しないものは意外性があり，利用者は興味を持つ可能性があるとみなして大きな重みを与える．

また，(3) ではデータ構造を構築するときのパラメタである許容誤差を推定し，この値を最適許容誤差 $\epsilon_{opt}$ として (4) で利用する．

(4) では，アイテム集合 $I$，系列データ $S$ に対して，図 1 のような，あるしきい値を持つアイテムのためのべき集合 $2^I$ 上の**束** (lattice) を構築する．この束を**選言パターン束**（Disjunctive Pattern Lattice, 以下 DPL）と呼ぶ．DPL は，閉路の無いラベル付き根付き有向グラフ $(V, E)$ である．$V$ は節点の有限集合, $E$ は有向辺の集合 ($E \subseteqq V \times V$) を表す．DPL にはただ一つの **根節** (root) が含まれ，この節はラベルを有さない．根からの距離が $n$ の**節点** $v_p \in V$ は $n$ -候補パターン $p$ を表し，3 項組 (推定頻度:$f$, 最大誤差:$\Delta$, 節点最終更新時刻:$time_{last}$) を**ラベル**として持つ．ここで $time_{last}$ は，DPL 中の各節点の最終更新時刻である．$f$ は $time_{last}$ の時点までに計算されたパターンの推定頻度値である．また $\Delta$ は，節点が DPL に最初に追加された時点までに処理された全系列データ数を表しており，節点の更新時には変更されない．$(n+1)$-候補パターン $v'$ が $v$ を部分パターンに持つとき，**有向辺** $e(v, v') \in E$ が存在する．

［例 3］ $S \in NS$ を cabcbba, $time_{now} = 1$, $\epsilon_{opt} = 0.1$ とした時，DPL は図 1 のように構築される．$v_{[a,\ b]}$ は，ラベルとして $(f : 0.29, \Delta : 0, time_{last} : 1)$ を持つ．また，$e(v_a, v_{[a,\ b]})$, $e(v_b, v_{[a,\ b]})$ から，$v_a$, $v_b$ をそれぞれ部分パターンとして持つ．

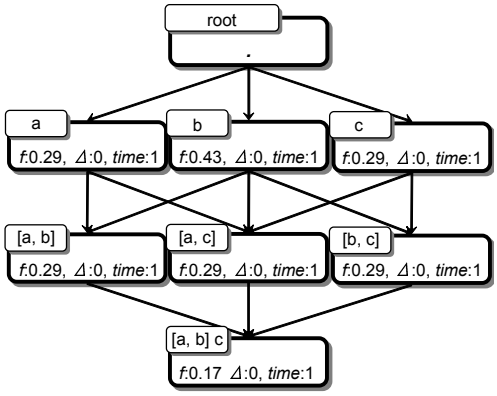


図 1 $\epsilon_{opt} = 0.1$ で構築された DPL

また (4) と (5) において，発行時刻に基づく重み $\omega$ を導入し，発行時刻が新しい系列データ上に生じる頻度ほど，より大きな重みを与える．すなわち，DPL 中の各節点へのアクセス時に**忘却関数** (decay function) による重みを用い，発行時刻の古い系列データ上で生じる頻度を指数的に減衰させる．

$$\omega = u\lambda^{(time_{now} - time_{last})} \quad (5)$$

$\lambda$ は忘却定数 $0 \leqq \lambda \leqq 1$ を表す．この値が小さいほど $\omega$ も小さく，頻度の忘却の度合いが大きい．また $u$ は，

$$u = \begin{cases} 1\ ,(if\ time_{now} - time_{last} \leqq \tau) \\ 0\ ,(if\ time_{now} - time_{last} > \tau) \end{cases} \tag{6}$$

で表される単位ステップ関数 (unit step function) であり，$\tau$ は有効期間を表す．従って，節点の最終更新時刻からの経過期間が有効期間外であれば，$\lambda$ がどのような値でも $\omega = 0$ となる．

### 4.3 系列データからの候補パターン集合生成

$NS$ 中の系列データ $S$ に対し，候補パターンの生成を行う．生成には，候補パターン $p$ の $S$ 中での出現開始位置 $head$ と終了位置 $tail$ を保持した**出現位置リスト**を用いる．すなわち，$p$ の $S$ 中での出現位置を 2 項組 $(head, tail)$ とするとき，$p$ に関する出現位置リスト $list_p$ を次のように定義する．

$$list_p = \langle (head_{p_1}, tail_{p_1}), ..., (head_{p_h}, tail_{p_h}) \rangle \tag{7}$$

$h$ は系列先頭頻度 $H(S,p)$ を表し，各 $head_{p_i} \leqq tail_{p_i}, 1 \leqq i \leqq h$ は同じリスト中では重複しない．

ここでは出現位置リストを用いる．系列データ $S$ からパターン中のアイテム数が $n$ で，かつ頻度が初期許容誤差 $\epsilon_0$ 以上を満たすような $n$ -候補パターン集合 $C_n$ を生成する手順を以下に示す．

---

**入力**: 系列データ $S$, 初期許容誤差 $\epsilon_0$
**出力**: $n$ -候補パターン集合 $C_n$
**手順**:
$n = 1$ **の時**: $S$ を 1 回走査して 1-候補パターンとその出現位置リストを生成し，頻度が $\epsilon_0$ 以上を満たすものを $C_1$ に追加する．
$n > 1$ **の時**: $(n-1)$-候補パターンの出現位置リストから，順次 $n$ -候補パターンとその出現位置リストを取得し，頻度が $\epsilon_0$ 以上を満たすものを $C_n$ に追加する．

---

上述の手順で，$(n-1)$-候補パターン $p$,$q$ の出現位置リスト $list_p$, $list_q$ から $n$-候補パターン $pq$ の出現位置リスト $list_{pq}$ を得る手順を示す．$list_p$ 中の $tail_{p_i}$ が $list_q$ 中の $head_{q_j}$ より小さいとき，$pq$ は $head_{p_i}$ から $tail_{q_j}$ までの間に出現する．すなわち，$pq$ の出現位置 $head_{pq_k}, tail_{pq_k}$ は，$head_{p_i}, tail_{q_j}$ と一致する．

**入力**: 出現位置リスト $list_p$, $list_q$
**出力**: 出現位置リスト $list_{pq}$
**手順**:
（1） $list_p$，$list_q$ を突き合わせ，$tail_{p_i} < head_{q_j}$ を満たす $head_{p_i}$，$tail_{q_j}$ をすべて見つけ，それぞれを $head_{pq_k}, tail_{pq_k}$ として $list_{pq}$ に追加する．
（2） すべての出現位置を追加し終わったら $list_{pq}$ を出力する．

［例 4］ $S$ を `cabcbba` とすると，パターン $a$,$b$,$c$ の出現位置リスト $list_a$, $list_b$, $list_c$ はそれぞれ，

$$\begin{aligned} list_a &= \langle (2,2),(7,7) \rangle \\ list_b &= \langle (3,3),(5,5),(6,6) \rangle \\ list_c &= \langle (1,1),(4,4) \rangle \end{aligned}$$

となり，パターン $ab$ と $[ac]$ の出現位置リスト $list_{ab}$, $list_{[ac]}$ は，

$$\begin{aligned} list_{ab} &= \langle (2,3) \rangle \\ list_{[ac]} &= \langle (1,2),(2,4),(4,7) \rangle \end{aligned}$$

となる．

### 4.4 候補パターンへの好みに基づく重み付け

候補パターン集合 $C_n$ 中の各パターン $p$ に対し，好みに基づく重み付けを行う．与えられた正負の興味パターンの有限集合 $Pos = \{p_{pos_1}, ..., p_{pos_l}\}, Neg = \{p_{neg_1}, ..., p_{neg_m}\} (l, m \geqq 0)$ に対し，**興味パターンリスト**とは $L = \langle Pos, Neg \rangle$ をいう．

以下に，興味パターンリストを用いた候補パターンへの好みに基づく重み付けを行う手順を示す．

---

**入力**: 候補パターン集合 $C_n$, 興味パターンリスト $L$, 更新判定しきい値 $th_{new}$
**出力**: 興味重み集合 $W_{C_n}$, 興味パターンリスト $L$
**手順**:
$n = 1$ **の時**: (1) $C_1$ 中の各候補パターンと $L$ 中の各興味パターンとの類似度を計算し，$Pos$, $Neg$, $L$ との類似度をそれぞれ求める．
(2) 興味重み集合 $W_{C_1}$ を計算し，同時に $L$ との類似度が $th_{new}$ 以下の候補パターンを $L$ に追加する．
$n > 1$ **の時**: (1) $C_n$ 中の各候補パターンの全部分パターンの興味重みを用い，その合計の平均値を候補パターンの興味重みとし，$W_{C_n}$ を求める．

---

$n = 1$ の時の手順を詳細に示す．まず $C_1$ 中の各 $p$ に対して，$Pos$ および $Neg$ 中のパターン $x$ との類似度 $sim(x,p)$ を計算する．本稿では類似度として，(8) 式の Kullback-Leibler(KL) 情報量を用いる．$Prob(p)$ および $Prob(x)$ はそれぞれ $p$ ,$x$ の確率分布を表す．

次に (9) 式で $p$ と $Pos$, $Neg$ との類似度 $sim(Pos,p)$, $sim(Neg,p)$ および，$p$ と $L$ との類似度 $sim(L,p)$ をそれぞれを求め，(10) 式で $p$ の興味重み $w_p$ を求める．$th_{max_{pn}}$ とは，興味重みの最大値を決定する値であり，$th_{max_{ue}}$ は，意外性重みの最大値を決定する値である．また $th_{ue}$ は，意外性を判断するしきい値である．$w_p$ を求める一連の計算は，$p$ が持つ興味の方向性とその大小を決定することに対応する．

$sim(L,p)$ が興味パターンリストの更新しきい値 $th_{new} \leqq th_{ue}$ より小さければ，(11) 式に従い $Pos$ もしくは $Neg$ に $p$ を新しい興味パターンとして追加する．

$$sim(x,p) = \sum Prob(x) log(\frac{Prob(x)}{Prob(p)}) \tag{8}$$

$$sim(L,p) = \mathrm{MIN}\{sim(\alpha,p) | \alpha \in L\} \tag{9}$$

$$w_p = \begin{cases} 1 + th_{max_{ue}}(\frac{sim(Pos,p)+sim(Neg,p)}{2}) \ , \\ \qquad\qquad (if\ sim(L,p) \geqq th_{ue}) \\ 1 + th_{max_{pn}}(1 - sim(Pos,p)) \ , \\ \quad (if\ sim(Pos,p) < sim(Neg,p)) \\ 1 - th_{max_{pn}}(1 - sim(Neg,p)) \ , \\ \quad (if\ sim(Pos,p) > sim(Neg,p)) \end{cases} \tag{10}$$

$$\begin{cases} p \text{ を } Pos \text{ に追加}\ ,(if\ 1 < w_p < 1 + th_{max_{pn}}) \\ p \text{ を } Neg \text{ に追加}\ ,(if\ 1 > w_p > 1 - th_{max_{pn}}) \end{cases} \tag{11}$$

### 4.5 最適許容誤差の推定

候補パターン集合 $C_n$ に対する興味重み集合 $W_{C_n}$ を用いて，最適許容誤差 $\epsilon_{opt}$ の推定を行う．推定には F 値を用い，$C_n$ 中の各候補パターン $p$ の頻度が推定許容誤差 $\epsilon_{est}$ 以上で，$W_{C_n}$ 中の興味重み $w_p$ が 1 以上となるパターンを正解とする．本稿での F 値 $F$ および再現率 ($Recall$) $\beta$，適合率 ($Precision$) $\gamma$ の定義を以下に示す．

$$\begin{cases} \beta = \frac{\text{頻度 } \epsilon_{est} \text{ 以上かつ } w_p > 1 \text{ なパターン数}}{w_p > 1 \text{ なパターン数}} \\ \gamma = \frac{\text{頻度 } \epsilon_{est} \text{ 以上かつ } w_p > 1 \text{ なパターン数}}{\text{頻度 } \epsilon_{est} \text{ 以上なパターン数}} \\ F = \frac{2\beta\gamma}{\beta+\gamma} \end{cases} \tag{12}$$

上記の $F$ を用い，初期許容誤差 $\epsilon_0$ から繰り返し計算を行い，$F$ を最大とするような $\epsilon_{est}$ を $\epsilon_{opt}$ とする．以下に最適許容誤差の推定を行う手順を示す．

---

**入力**: 興味重み集合 $W_{C_n}$, 初期許容誤差 $\epsilon_0$, 許容誤差倍率 $rate$, 繰り返し回数 $loop$
**出力**: 最適許容誤差 $\epsilon_{opt}$
**手順**:
$n = 1$ **の時**: (1) 推定許容誤差 $\epsilon_{est}$ を $\epsilon_0$ から $\frac{\epsilon_0 rate}{loop}$ 刻みで大きくしていき，計 $loop$ 回の F 値計算を行う．
(2) 計算された全ての F 値から，F 値を最大とする $\epsilon_{est}$ を $\epsilon_{opt}$ として出力する．
$n > 1$ **の時**: (1) $n = 1$ で推定した $\epsilon_{opt}$ を出力する．

---

$rate$ は $\epsilon_{opt}$ が取りうる最大値が $\epsilon_0$ の何倍となるかを定める許容誤差倍率を，また $loop$ は最適許容誤差計算の繰り返し回数を定める値を意味する．$rate$ が一定のとき，$loop$ を大きくすると $\epsilon_{opt}$ の推定精度は向上する．

### 4.6 候補パターン集合からの選言パターン束構築

系列データ $S_i$ から生成される $n$ -候補パターン集合中のパターン $p \in C_n$ について，DPL $\mathcal{D}$ への節点としての追加，更新，削除の各操作を次に示す．

---

**入力**: 候補パターン集合 $C_n$, 興味重み集合 $W_{C_n}$, 最適許容誤差 $\epsilon_{opt}$
**出力**: DPL $\mathcal{D}$
**操作**:
**追加**: $\exists v_p \notin \mathcal{D}$ かつ $w_p \sum D_w(S_i,p) \geqq \epsilon_{opt} wsize$ であれば，ラベル ($w_p \sum D_w(S_i,p)$, $N_{last}$, $time_i$) を持つ節点として $\mathcal{D}$ へ $v_p$ を追加する．
**更新**: $\exists v_p \in \mathcal{D}$ かつ ($w_p \sum D_w(S_i,p) + \omega f + \Delta \geqq \epsilon_{opt} N_{all}$) であれば，$v_p$ のラベルを ($w_p \sum D_w(S_i,p) + \omega f$, $\Delta$, $time_i$) として $v_p$ を更新する．
**削除**: $\exists v_p \in \mathcal{D}$ かつ ($w_p \sum D_w(S_i,p) + \omega f + \Delta < \epsilon_{opt} N_{all}$) であれば，$v_p$ およびその下位節点 ($n+1$ 以降) を全て削除する．

---

$N_{all}$ は $S_i$ も含め，これまでに処理された全系列データ数を表す．また，$N_{last}$ は ($S_i$ を除く)，これまでに処理された系列データ数 $N_{all} - wsize$ を表す．

$\sum D_w(S_i,p)$ は $N_{last}+1$ から $N_{all}$ までの系列データの全体出現頻度 $D_w(S,p)$ の合計であり，次の式で表される．

$$\sum D_w(S_i,p) = \sum_{l=N_{last}+1}^{N_{all}} D_w(S_{i_l},p) \tag{13}$$

発行時刻に基づく重み $\omega$ の計算に用いられる現在時刻 $time_{now}$ は，$S_i$ の発行時刻 $time_i$ と一致することに注意したい．

［例 5］ 図 1 の DPL 中の，節点 $v_a$ を更新する．$\epsilon_{opt} = 0.1$, $N_{all} = 3$, $time_{now} = 3$, $wsize = 1$, $\lambda = 0.98$, $\tau = 3$, $\sum D_w(S,a) = 0.2$, $w_a = 1.5$ の時，$\omega = 0.96$, 頻度 $w_a \sum D_w(S,p) + \omega f + \epsilon_{opt}\Delta = 0.58$ となる．$\epsilon_{opt} N_{all} = 0.3$ であるので，$v_a$ の新しいラベルは $(f : 0.58, \Delta : 0, time_{last} : 3)$ と更新される．また，$v_c$ を DPL 中 から削除する場合，その下位節点 $v_{[a,c]}, v_{[b,c]}, v_{[a,b]c}$ も削除される．それぞれの操作を反映した DPL を図 2 に示す．

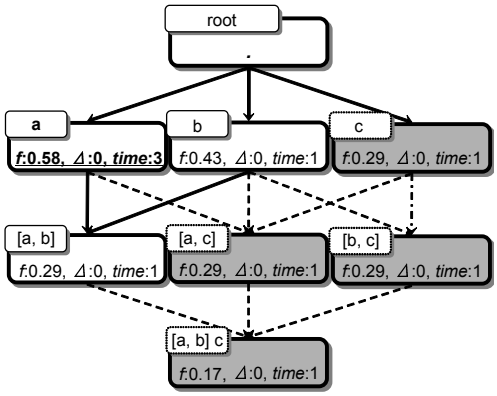


図 2 DPL に対しての更新・削除操作

### 4.7 選言パターン束からの頻出パターン集合抽出

DPL $\mathcal{D}$ からのパターン抽出を行うため，**深さ優先探索** ($Depth\text{-}First\ Search$) を用いる．各節点について，$\sigma N_{all} \leqq \omega f + \epsilon_{opt}\Delta$ を満たす節点は頻出であるとして頻出パターン集合 $\mathcal{F}$ に出力する．

本手法では，最適許容誤差 $\epsilon_{opt}$ を含めた頻度を用いて頻出パターン判定を行う．$\epsilon_{opt}$ は，各繰り返し時に推測され変動するが，$\epsilon_0$ から $\epsilon_0 \cdot rate$ の間に収まる．

［例 6］ 図 1 において，$\sigma = 0.4$, $\epsilon_{opt} = 0.1$, $N_{all} = 1$, $time_{now} = 2$, $\lambda = 0.98$, $\tau = 3$ とするとき，$\sigma N_{all} = 0.4$ 以上の度数を有するすべてのパターンを抽出する．この探索は図 3 のような順路に従い，頻出パターン集合 $\mathcal{F}$ を得る．図中の丸囲い数字は，探索の順序を示す．パターン a については，$\omega f = 0.28$ となり頻出とはならない．しかし，パターン b については，$\omega f = 0.42$ となるので頻出パターンとして出力する．

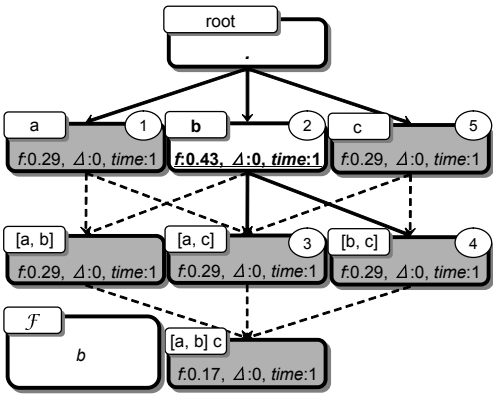


図 3　DPL からの頻出パターン抽出

## 5. 実　　験

### 5.1 実験の方法

本稿では 3 つの実験を行う．実験 1 の目的は，頻出パターンの評価にある．実験 2 では，提案手法の実行効率および DPL 構築負荷の評価を，更に実験 3 は，提案手法が頻出パターンに与える影響と DPL の空間効率を評価することを目的とする．

本稿では，Reuters-21578 text categorization test collection をコーパスとして使用する．これは 1987 年のロイター社の発信記事を日時順に並べたものであり，21,578 件の記事からなるが，2200 件分の記事 (同発行時刻の記事 100 件を 1 組として，時系列順に 22 組) を用いる．日単位で扱い，それぞれの記事の発行時刻は，1 組目の発行時刻を基準とした相対時刻で考える．各記事については，予め 6 つのグループ {“銀行”, “会社経営”, “株式”, “商取引”, “決算報告”, “その他”} に分類しておく．初めの 5 グループに (重複を含め) 属する記事は，実験データの 70.2%にあたる 1544 件，いずれにも属さない “その他”に属する記事は，29.8%にあたる 656 件である．また，記事中の不要語 (*stop word*) の削除および単語のステミング [12] をしておく．表 1 に実験データ中の記事の例，表 2 に実験データの詳細を示す．

実験データ中に現れる全 16336 アイテム (語) の分布情報は，事前知識として利用する．また，実験で生成される候補パターンは，(a, [ab], [ab]c 等のように) 選言の深さを 1 レベルのみに制限する．

全ての実験で利用される興味パターンは，正の興味パターン $Pos = \{$year, compani, bank, share, net$\}$，負の興味パターン $Neg = \{$said, mln, dlr, sale, pct, ct$\}$ とする．$Pos$ には主に bank, share 等の金融関係のパターン，$Neg$ には said のような新聞記事特有のパターンや，dlr, pct などの数値や通貨の単位を表すパターンをそれぞれ選択しておく．

表 1　“株式”グループに属する記事の例

| |
|---|
| **shr** **loss** 51 ct **loss** ct |
| **net** **loss** 7041000 **loss** 467000 |
| rev 138 9 mln 131 4 mln |
| 12 month |
| **shr** **loss** 64 ct profit 46 ct |
| **net** **loss** 8843,000 profit 63060000 |
| rev 558 9 mln 556 7 mln |
| note: **net** **loss** 4th qtr 1986 includ charg |
| restructur 2 6 mln dlr tax 19 ct **share** |
| 1986 **net** **loss** includ tax special |
| charg 2 7 mln dlr 20 ct **share** |

### 5.2 実 験 結 果

実験 1 では，実験データ 1700 件に対して，$\epsilon_0$ を 0.0025, $\lambda$ を 0.98, $\tau = 300$, $wsize$ を 100 件，$th_{new}$ を 0.1，$th_{ue}$ を 0.3 とそれぞれ設定し，頻出パターン抽出を行う．比較手法には，提案手法である興味重み計算と最適許容誤差の推定を行う手法 (KL+F 手法)，どちらの計算も行わない (none 手法) の 2 手法を用いる．予備実験の結果から，KL+F 手法では $th_{max_{pn}}$ を 1，$th_{max_{ue}}$ を 5，$loop$ を 1000 回，$rate$ を 1.2 とそれぞれ設定を行うとともに，$\sigma$ を 0.003 とした時の頻出パターン上位 10 件を表 3 に示す．パターン横の数字は，そのパターンの頻度を表している．

表 3　頻出パターン上位 10 件 (1700 件)

| 順位 | KL+F | none |
|---|---|---|
| 1 | compani(17.1) | said(39.7) |
| 2 | share(16.5) | mln(22.0) |
| 3 | year(15.8) | dlr(18.0) |
| 4 | 1(15.7) | [said, dlr](16.8) |
| 5 | billion(15.5) | ct(15.6) |
| 6 | net(13.9) | [said,mln](14.3) |
| 7 | shr(13.6) | pct(14.1) |
| 8 | bank(13.3) | [mln, dlr](13.2) |
| 9 | loss(11.3) | [said, pct](12.9) |
| 10 | april(11.0) | compani(8.59) |

実験 2 では，実験 1 と同じパラメタを用い，実験データ 2200 件に対しての DPL 構築時間と負荷 (節点の追加, 更新, 削除の各回数) を計測する．比較手法には，KL+F 手法，none 手法に加え，興味重み計算のみを行う手法 (KL 手法) の 3 手法を用いる．実験結果を表 4 に示す．

実験 3 では，実験 1 と同じパラメタを用い，実験データ 1700 件に対し，KL+F 手法，KL 手法，none 手法を用いて DPL 構築を行い，総節点数および $\sigma$ を 0.003 として抽出した際の頻出パターン数を計測する．また結果から，頻度パターン上位 30 件のカバレージ $Cover$ と，総接点数に対して頻出パターン数の

表 2　実験データの詳細

| データ [組] | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 時刻 [day] | 0 | 5 | 7 | 11 | 14 | 15 | 19 | 21 | 25 | 26 | 28 | 32 | 34 | 36 | 40 | 41 | 46 | 95 | 96 | 113 | 235 | 236 |
| 銀行 [件] | 13 | 15 | 11 | 29 | 20 | 14 | 16 | 27 | 33 | 10 | 16 | 10 | 23 | 14 | 10 | 10 | 36 | 17 | 20 | 24 | 19 | 12 |
| 会社経営 [件] | 35 | 37 | 32 | 31 | 36 | 36 | 27 | 33 | 12 | 34 | 34 | 33 | 28 | 31 | 21 | 35 | 28 | 44 | 29 | 38 | 18 | 38 |
| 株式 [件] | 18 | 25 | 22 | 24 | 20 | 17 | 24 | 24 | 17 | 25 | 22 | 24 | 16 | 23 | 11 | 19 | 24 | 20 | 15 | 25 | 31 | 25 |
| 商取引 [件] | 11 | 13 | 36 | 14 | 17 | 15 | 17 | 13 | 5 | 18 | 16 | 13 | 10 | 23 | 10 | 17 | 14 | 13 | 10 | 15 | 14 | 11 |
| 決算報告 [件] | 22 | 29 | 27 | 10 | 25 | 23 | 26 | 28 | 7 | 20 | 21 | 17 | 24 | 15 | 12 | 24 | 12 | 14 | 19 | 14 | 26 | 34 |
| その他 [件] | 32 | 19 | 14 | 28 | 26 | 30 | 29 | 28 | 42 | 32 | 30 | 38 | 22 | 23 | 55 | 27 | 30 | 30 | 36 | 34 | 30 | 21 |

表 4　DPL 構築負荷と実行時間

| 手法 | KL+F | KL | none |
|---|---|---|---|
| 追加 [回] | 3045 | 3092 | 2644 |
| 更新 [回] | 1920 | 1943 | 1781 |
| 削除 [回] | 2729 | 2772 | 2368 |
| 実行時間 [msec] | 192888 | 190609 | 45999 |

占める割合を示す DPL 空間効率をそれぞれ計算する．$Cover$ は (14) 式で定義し，実験結果を表 5 に示す．

$$Cover[\%] = \frac{\text{比較 2 手法で共に抽出されたパターン数}}{30} \times 100 \quad (14)$$

表 5　カバレージと DPL 空間効率

| 手法 | KL+F | KL | none |
|---|---|---|---|
| V.S. | KL | none | KL+F |
| Cover[%] | 100 | 43.3 | 43.3 |
| 総節点数 | 690 | 722 | 417 |
| 頻出パターン数 | 43 | 43 | 30 |
| DPL 空間効率 [%] | 6.23 | 5.96 | 7.19 |

### 5.3 考　　察

実験結果の考察を行う．実験 1 では，頻出パターンを抽出した．頻出パターンと興味重みの有無については，表 3 より，KL+F 手法では `compani`, `share` などの正の興味パターンが高頻出となっており，興味重みの有効性が確認できる．

表 2 より，“会社経営”に属する記事はデータ 1 組あたり平均 31 件と最も多く，両手法で高頻出パターン `compani` が抽出できる．一方で，“銀行”, “株式”に属する記事はそれぞれ平均 18 件，21 件と少ないが，KL+F 手法では，高頻出パターン `bank`, `share` が抽出できる．すなわち，興味重みを考慮することで，関連する記事の多少にかかわらず，興味パターンを高頻出であるとして抽出できる．

実験データ中の記事と KL+F 手法で得られる頻出パターンを比較する．表 1 の “株式”に属する記事に関して，正の興味パターン `net` や `share`(太字) に加え，興味パターンではない `loss` や `shr`(下線) についても抽出できる．後者は，記事の具体的内容を表しており，また前者と類似しているために高頻出パターンとして抽出されると考えられる．このことから，興味重みを考慮して得られる結果は，利用者の好みに合うことがわかる．

実験 2 では提案手法の実行効率と DPL 構築負荷を解析した．実行効率に関しては none 手法が最も良く，実行時間は KL+F,KL 両手法と比べ約 4.2 倍高速である．しかし，KL+F 手法と KL 手法では，実行時間差は約 2 秒と差がなく，このことから，提案手法では興味重み計算が中心となっていることがわかる．また，KL+F 手法と KL 手法を用いたときの構築負荷を比べると，KL+F 手法の方が約 1.5%少なく，最適許容誤差の推測により，DPL 構築時の負荷を軽減できることがわかる．

実験 3 では提案手法が頻出パターン抽出結果に与える影響と，DPL の空間効率を調べている．none 手法と KL+F,KL 両手法との各カバレージは 43.3%となる．つまり，興味重みによる重み付けの影響によって，これまで非頻出とされたパターンが抽出され，抽出結果が大きく変化したことを示している．また KL+F 手法は，KL 手法とのカバレージが 100%となり抽出結果自体は変わらないが，最適許容誤差の推測により，DPL 中の空間効率が 0.27%改善されている．

## 6. 結　　論

本稿では，ニュースストリームから利用者の好みを反映した頻出パターンを抽出するマイニング手法を提案した．提案手法は，オンライン型の単一パスアルゴリズムであり，利用者からの興味パターンリストを用いて候補パターンに重み付けを行うと共に，最適許容誤差の推測を行うことで DPL を効率よく構築可能である．また，実験を通して，現実のデータに対して十分に有効であることを確認した．

## 文　　献


[1] Aggarwal, C.C.. and Yu, P.S.: Online Generation of Association Rules, ICDE, 1998, pp.402-411

[2] Agrawal, R. and Srikant, R.: Fast Algorithm for Mining Association Rules, proc. VLDB, 1994, pp.487-499

[3] Goethals, B.: A Survey on Frequent Pattern Mining, Univ.of Helsinki, 2003

[4] Han,J. and Kamber, M.: Data Mining : Concepts and Techniques, Morgan Kaufmann, 2000

[5] Jian, N. and Gruenwald, L.: Research Issues in Data Stream Association Rule Mining, SIGMOD Record 35-1, 2006, pp.14-19

[6] Mcgarry, K.: A Survey of Interestingness Measures for Knowledge Discovery, The Knowledge Engineering Review Vol.20,Issue 1, 2005, pp.39-61

[7] Hand, D., Mannila, H. and Smyth, P.: Principles of Data Mining, MIT Press, 2001

[8] Manku G. S. and Motwani R.: Approximate frequency counts over data streams, VLDB, 2002, pp.346-357

[9] Shimizu, K. and Miura, T.: Disjunctive Sequential Patterns on Single Data Sequence and its Anti-Monotonicity, International Conference on Machine Learning and Data Mining (MLDM), 2005, pp.376-383

[10] Shimizu, K., Shioya, I. and Miura, T.: Mining Disjunctive Sequential Patterns from News Stream, 8th Intn'l Conf. on Intelligent Data Engineering and Automated Learning (IDEAL), 2007, pp.630-642
[11] 安部, 藤野, 下薗, 有村, 有川: テキストデータからの高速データマイニング：探索的文書ブラウジングとウェブデータへの応用, 人工知能学会誌 Vol.15, No.4, 2000, pp.618-628
[12] 北, 津田, 獅子堀: 情報検索アルゴリズム, 共立出版, 2002
[13] 高野，岩沼，鍋島: 単一の長大なデータ系列上の系列パターンの出現尺度とその逆単調性, 第 3 回情報科学技術フォーラム (FIT), 2004, pp.115-118
[14] 清水, 三浦: 選言パターン抽出のオンライン分析, DBSJ Letters Vol.4 No.3, 2005, pp.9-12
[15] 吉田，太田，小林，湯上: サブパターンとスーパーパターンからの推定頻度に基づくパターンの興味深さの尺度の評価, DBSJ Letters Vol.3 No.1, 2004, pp.57-60