情報量と頻度に基づく非同期かつ有用な系列パターンの高速抽出

Mining Asynchronous Interesting Sequential Patterns based on Frequency and Self-Information

村田 順平

Murata Junpei

山梨大学大学院 医学工学総合教育部 コンピュータ・メディア工学専攻

Computer Science and Media Engineering, Interdisciplinary Graduate School of Medical and Engineering, University of Yamanashi
g08mk028@yamanashi.ac.jp

岩沼 宏治

Iwanuma Koji

山梨大学大学院 医学工学総合研究部

Interdisciplinary Graduate School of Medical and Engineering, University of Yamanashi
iwanuma@yamanashi.ac.jp

大塚 尚貴

Ohtsuka Naoki

山梨大学大学院 医学工学総合教育部 コンピュータ・メディア工学専攻^{*1}

Computer Science and Media Engineering, Interdisciplinary Graduate School of Medical and Engineering, University of Yamanashi

keywords: pattern mining, self-information, frequency, sequential data mining,

Summary

In this paper, we propose new methods and gave a system, called IFMAP, for extracting interesting patterns from a long sequential data based on frequency and self-information, and experimentally evaluate the proposed methods in the application of handling a newspaper article corpus. Sequential data mining methods based on frequency have intensively been studied so far. These methods, however, are not effective nor valuable for some applications where almost all high-frequent patterns should be regarded just as meaningless noisy patterns. An information-gain concept is quite important in order to restrain these noisy patterns, and was already studied for integrating it with a frequency criteria. Yang et. al. gave a sequential mining system InfoMiner which can find periodic synchronous patterns being interesting and well-balanced from the both view-points of frequency and self-information. In this paper, we refine and extend the InfoMiner technologies in the following points: firstly, our method can handle ordinary, i.e., asynchronous and non-periodic patterns by using a sliding window mechanism, whereas InfoMiner cannot; secondly we give several combination measures for choosing valuable patterns based on frequency and self-information, while InfoMiner has just one measure which, we show in this paper, is not appropriate nor effective for handling newspaper article corpora; thirdly, we proposed a new unified method for pruning the search space of sequential data mining, which can uniformly be applied to any combination measures proposed here. We conduct experiments for evaluating the effectiveness and efficiency of the proposed method with respect to the runtime and the amount of excluding noisy patterns.

1. ま え が き

本論文では、長大な単一系列データから非同期・非周期型の有用系列パターンを抽出することを目的として、出現頻度と自己情報量を組み合わせた系列パターンの評価基準を幾つか提案する。また、それに基づいた有用系列パターンの高速抽出アルゴリズムを提案し、新聞記事コーパスを用いて実験の評価を行う。

系列データマイニングとは、巨大な系列データから有用な系列パターンを高速抽出する技術である。従来より出現頻度の高いパターンを有用系列として、高速に抽出する技術が盛んに研究 [Agrawal 95, Han 98, Pei 04, Iwanuma 05, Srikant 96] されている。しかし応用によっては、高

頻なパターンはしばしば意味のないノイズとして扱われる。例えば、新聞記事コーパスからの有用系列パターンの抽出問題 [市川 08, 多田 09] では、「訃報」や「おくやみ」などを構成要素とするパターンは極めて高い出現頻度を持つ。これらのパターンはほぼ毎日恒常的に出現しており、あまり有用な情報は保有しておらず、ノイズに近いパターンと考えられる。一方で、出現頻度が非常に小さい系列パターンは偶発的なものが多く、これもあまり有用なパターンとは考えられない。有用な系列パターンは中程度の出現頻度を持つものに多いと考えられる。

出現頻度が低い系列を重要視する最も代表的な基準は情報量である。トランザクションデータベース中の相関規則 (association rule) に対しては、情報量に基づく有用性基準についても多くの研究がある [Blanchard 05, Geng 06].

^{*1} 現職：三洋電機株式会社

本研究で考察の対象とする単一の長大な系列データに対する先行研究としては [Yang 01, Yang 04] の InfoMiner があり、情報量と出現頻度の両方を考慮した有用性基準と利用技術について考察している。ただそこでは周期的・同期的な系列パターンの抽出を考察しており、通常の意味^{*2}での有用系列パターン、即ち非同期・非周期的なパターンの抽出が行えない。またアイテム集合の系列も取扱えず、例えば、我々が研究してきた新聞記事コーパスのマイニングには枠組みが不十分である。更に [Yang 01, Yang 04] では、系列有用性の尺度として、最も素朴な出現頻度尺度と系列要素の自己情報量の和（総和情報量）の組合せを考えているが、通常、出現頻度の値はかなり大きな値となるために、このままで情報量とのバランスが悪く、有用な系列パターンを抽出することは実際には困難である。

本論文では上記の諸問題の解決に取り組む。まず長大な単一系列データ中に出現する通常の（即ち非同期・非周期的な）有用パターンを抽出するために、[Mannila 95]などで広く利用されているスライド窓機構とパターンの正規化を導入する。また新たに、対数出現頻度や平均情報量など幾つかの評価尺度を導入し、その組合せ尺度の特性や有用性について実験の評価を行う。更に系列パターンの抽出計算の高速化手法について考察する。まず探索空間を再構築して効率化し、次に InfoMiner で導入された枝刈手法を拡張・一般化して、種々の組合せ尺度に統一的に適用できる高速化技法を与える。以上の成果を、系列パターン抽出システム IFMAP として実装し、新聞記事コーパスを用いた包括的な評価実験を行う。

関連研究であるが、情報量に基礎を置く有用性尺度は相関規則に対しては多数の研究 [Blanchard 05, Geng 06] があるが、情報量と出現頻度を合わせた評価尺度は、筆者らの知る限りにおいては、他には研究例がないと思われる。一方で、単一の長大な系列データ中のパターン抽出に関しては [Yang 03b] があるが、そこでは単一系列データベース全体ではなく局所的に有用な準周期的なパターン抽出問題を考察している。また [He 08] は情報利得ではなく相関性に関する局所的有用パターンの抽出発見を考察している。一方で、単一系列データ中の非同期パターンの抽出に関しては、[Yang 03a, Yeh 09] の研究があるが、あくまで周期性パターンの抽出問題を考察している。非周期的かつ非同期出現する通常のパターンの抽出には対応していない。また情報利得基準の直接的な改良に取り組んだ研究は、筆者の知る限りにおいては無く、本研究が初めての試みと考えられる。最後に、相関

規則などの抽出データの評価基準は数多くの提案がなされている [Geng 06, Tan 02]。本論文で考察対象とする単一系列データマイニングに関するものとしては、例えば文献 [市川 08, Mannila 95] で抽出系列データに対する種々の確信度 (confidence) が考察されている。また文献 [Lee 08] におけるファジー概念を導入した評価尺度がある。また [Cao 07] では時系列上の空間情報に関する周期性パターンの抽出に適した評価基準について考察している。[Lonardi 07] では「普通でない (unusual) 系列パターン」の抽出を、既存の正常なデータをマルコフモデルとして学習し、その上で統計学的な外れ値問題として捉えている。意図する応用がかなり異なっており、我々が目的としている新聞記事コーパス中の有用系列パターンの抽出問題へ適用できる否かは全くの未知数である。

本論文の構成は以下の通りである。2 章は準備である。3 章は非同期・非周期性パターンの抽出、4 章は各種の情報利得基準について述べる。5 章は有用パターンの高速抽出技術、6 章は実装システム IFMAP のアルゴリズムの概要を述べ、7 章で新聞記事コーパスを用いた包括的な実験の評価を行う。8 章はまとめである。

2. 準備

以下では文献 [Pei 04, Yang 04] に従って諸定義を行う。

【定義 1】全てのアイテムの集合を $E = \{e_1, e_2, \dots, e_n\}$ とする。 $|E|$ で集合 E の濃度を表す。単一系列データベースとはアイテム集合の順序付きリスト $S = \langle s_1, s_2, \dots, s_m \rangle$ である。各 s_i は $s_i \subset E$ である。 S の系列長 m を $|S|$ 、 s_i の濃度を $|s_i|$ で表す。 S のアイテムの総数 S_{all} を、 S に出現するアイテムの延数、即ち $S_{all} = \sum |s_i|$ とする。

先行研究の InfoMiner では、系列 S はアイテムの系列であったが、本研究ではアイテム集合の系列に拡張している点に注意して頂きたい。以下で系列を表記するときは、要素間のカンマを省略し、単一要素からなる系列要素は括弧を省略する。即ち、系列 $\{\{a\}, \{a, b\}, \{c\}, \{d\}, \{a, b, e\}, \{a, c\}\}$ を、 $\langle a\{ab\}cd\{abe\}\{ac\} \rangle$ のように略記する。

【定義 2】* を仮想アイテムと呼ぶ。このときパターン P とは、アイテム集合 $E \cup \{*\}$ 中のアイテムの順序付きリスト $P = \langle p_1, p_2, \dots, p_m \rangle$ である。 P のパターン長を $|P|$ で表す。また、空列パターンを ϵ と表記する。

* はワイルドカードに相当するアイテムである。

【定義 3】パターン $P = \langle p_1, p_2, \dots, p_m \rangle$ とアイテム $e \in E \cup \{*\}$ に対して、 $\langle p_1, p_2, \dots, p_m, e \rangle$ を e による P の単純右拡張と呼び、 $P \circ e$ で表す。また、 P に有限回の単純右拡張を施したものを P の右拡張パターンと呼ぶ。

3. 非同期・非周期的パターンの抽出

先行研究 [Yang 01, Yang 04] では、単一系列データが与えられたとき、系列をある一定の周期で区切り、その

*2 比較的短い複数のデータ系列 s_i から構成されるデータベース S を対象とした従来の系列データマイニングの研究 [Agrawal 95, Han 98, Pei 04, Srikant 96] では、ある部分系列 α が個々の（比較的短い）データ系列 s_i の中に存在するか否かだけが問題とされ、 s_i 中に複数回出現するか否か、あるいはその出現回数などは問題とされない。そのため、部分系列 α の“周期的・同期的な出現”という概念そのものが存在しないことに注意して頂きたい。

区間中に出現する同期パターンの抽出を考察している．

【例 1】単一系列データベース $S = \langle a\{ab\}b\{abe\}\{ac\}\{ba\}bc\{ab\} \rangle$ を，周期 3 で区切るとき，パターン $P = \langle a, *, b \rangle$ は，以下のように周期に同期して 2 回出現している．縦線は区間の区切りを表している．

$$\langle \underline{a\{ab\}b} \mid \{\underline{abe}\}\{\underline{ac}\}\{\underline{ba}\} \mid bc\{ab\} \rangle$$

この種の同期パターンは，曜日などの周期に沿ったイベント間の関係を調べることに適しているが，地震や台風など，特定の周期を持たずに非同期的に発生するとイベントを扱うのには適していない．

【例 2】例 1 の系列 S でパターン $P = \langle a, *, b \rangle$ は，以下のように非同期的なものも含めると 4 回出現している．

$$\langle \underline{a\{ab\}b} \mid \{\underline{abe}\}\{\underline{ac}\}\{\underline{ba}\} \mid bc\{ab\} \rangle$$

本論文では，このように通常普通に考慮される非同期パターンを抽出するために，[Mannila 95, Iwanuma 05] で用いられたスライド窓機構を導入する．

【定義 4】([Iwanuma 05]) 単一系列データベース $S = \langle s_1, \dots, s_m \rangle$ と整数 i ($1 \leq i \leq m$) に対して， S の i 番目の要素を開始位置とする長さ k ($1 \leq k$) のウィンドウ系列 $\text{win}(S, i, k)$ を以下のように定義する．

$$\text{win}(S, i, k) = \begin{cases} \langle s_i \cdots s_{i+(k-1)} \rangle & \text{if } i + (k-1) \leq m, \\ \langle s_i \cdots s_m \rangle & \text{otherwise.} \end{cases}$$

【定義 5】 E をアイテム集合とするととき， $(E \cup \{*\}) \times 2^E$ 上の二項関係 $\triangleleft$ を以下のように定める．

$$e \triangleleft s = \begin{cases} \text{true} & \text{if } e = * \text{ もしくは } e \in s \\ \text{false} & \text{otherwise.} \end{cases}$$

【定義 6】単一系列データベース S 中でのパターン P の基本出現頻度 $\text{freq}(S, P)$ を以下で定める．

$$\text{freq}(S, P) = \sum_{i=1}^{|S|} \delta(\text{win}(S, i, |P|), P)$$

ここで δ は以下の関数と定める．

$$\delta(\langle s_1 \cdots s_l \rangle, \langle p_1 \cdots p_m \rangle) = \begin{cases} 1 & \text{if } l = m \text{ かつ } p_1 \triangleleft s_1, \dots, p_m \triangleleft s_m \\ 0 & \text{otherwise} \end{cases}$$

P の S 中の出現位置とは， $\delta(\text{win}(S, i, |P|), P) = 1$ となる i である． P の S 中の出現位置の集合を $Oc(S, P)$ と表記する．例 1 では， $Oc(S, P) = \{1, 2, 4, 5\}$ である．

【定義 7】パターンに対して非負整数を返す関数 M が右逆単調であるとは，パターン P_1 の右拡張パターン P_2 に対して必ず $M(P_1) \geq M(P_2)$ となる場合を言う．

【補題 1】基本出現頻度 freq は右逆単調である．

《証明》定義より明らかである． $\square$

$P_1 = \langle *, a, b, *, c, * \rangle$ と $P_2 = \langle *, *, a, b, *, c \rangle$ はシフト同値な関係にある．例 1 で示したようなデータ中の周期的な区切りを前提として出現頻度を考える枠組みでは，この 2 つのパターンは本質的に異なるものとして，区別されて取り扱う必要がある．しかし，スライド窓を導入した場合には事情が異なってくる [Iwanuma 05]．即ち，この 2 つのパターンは本質的に同じものと考えられ，別個のものと考えすることは合理的ではない． P_1 と P_2 の両端の $*$ は冗長であると判断される．そこでパターンを正規化し，この冗長性を除去することを考える．即ち，パターン P のスケルトン $Sk(P)$ を， P の部分系列で， P で最初と最後に出現する E の元の間の部分系列を指すものと定める．上記の P_1 と P_2 に対して， $Sk(P_1) = Sk(P_2) = \langle a, b, *, c \rangle$ であり，スケルトンは同一である．スケルトンだけを考えることにより，パターンを正規化でき，結果として抽出パターンの数を押さえ込める．

4. 情報利得尺度

本章では頻度と情報量を組み合わせた情報利得尺度を導入する．先行研究である InfoMiner [Yang 01, Yang 04] におけるパターンの評価尺度 Information Gain にならない，本研究で導入するパターンの評価尺度を情報利得尺度と呼ぶ．情報理論における情報利得とは異なる尺度であることに注意されたい．本論文では [Yang 01, Yang 04] の結果を拡張し，種々の頻度尺度と情報量尺度と組み合わせを考え，より合理的で有用な尺度を考察する．まず基本出現頻度 freq を基にして，3 つの重み付き頻度関数を導入する．

4.1 頻度尺度

【定義 8】 S を単一系列データベース， P をパターンとする． MS を最小出現頻度と呼ぶ非負整数とするととき， S 中の P の単純頻度 $F_s(S, P)$ を以下のように定める．

$$F_s(S, P) = \begin{cases} \text{freq}(S, P) - (MS - 1) & \text{if } \text{freq}(S, P) \geq MS \\ 0 & \text{otherwise.} \end{cases}$$

MS は出現頻度の閾値である．基本出現頻度が MS 未滿ならば， $F_s(S, P) = 0$ となる．単純頻度は非常に大きくなるので，後ほど導入する情報量尺度の値とのバランスが問題となる．そこで以下の 2 つの頻度尺度を導入する．

【定義 9】 S 中の P の対数頻度 $F_l(S, P)$ と対数化相対頻度 $F_r(S, P)$ を以下のように定める．

$$F_l(S, P) = \log(F_s(S, P) + 1)$$

$$F_r(S, P) = - \left(\log \frac{F_s(S, P)}{|S|} \right)^{-1}$$

対数化相対頻度は，相対頻度 $\frac{F_s(S, P)}{|S|}$ *3 を対数化し，尺度の大小関係が逆転を防止するために逆数化したもの

*3 頻度尺度を後述の情報量尺度と組み合わせると利得尺度として使用する場合，系列パターンに対する順位付けに関して，相対

である．明らかに各重み付き頻度は P の基本出現頻度 $\text{freq}(S, P)$ に比例して増減する．補題 1 より freq は右逆単調であるので，各重み付き頻度も右逆単調である．

〔補題 2〕 重み付き頻度関数 $F_s(S, P), F_l(S, P), F_r(S, P)$ は全て右逆単調である．

4.2 情報量尺度

まず自己情報量を [Yang 04] に習って導入する．

【定義 10】 単一系列データベース S 上のアイテム $e \in E \cup \{*\}$ の自己情報量 $\text{info}(S, e)$ を以下のように定義する．

$$\text{info}(S, e) = -\log_{\frac{|S||E|}{S_{all}}} \left(\frac{1}{|S|} \text{freq}(S, \langle e \rangle) \right)$$

出現確率が平均値をとるアイテムの情報量を 1 とするため，対数の底は平均出現確率 $\frac{S_{all}}{|S||E|}$ の逆数としている．仮想アイテム $*$ は $\text{freq}(S, *) = |S|$ なので， $\text{info}(S, *) = 0$ となる．以上を用いて，次に 3 つの情報量関数 $I_s(S, P), I_a(S, P), I_m(S, P)$ を定義する．

【定義 11】 S を単一系列データベース， $P = \langle p_1, \dots, p_m \rangle$ をパターンとする． S 上の P の総和情報量 $I_s(S, P)$ ，平均情報量 $I_a(S, P)$ ，最小情報量 $I_m(S, P)$ をそれぞれ以下のように定める．

$$I_s(S, P) = \sum_{i=1}^{|P|} \text{info}(S, p_i)$$

$$I_a(S, P) = \frac{1}{\text{skel}(P)} \sum_{i=1}^{|P|} \text{info}(S, p_i)$$

$$I_m(S, P) = \min_{i=1}^{|P|} (\text{info}(S, p_i))$$

但し， $\text{skel}(P)$ は，パターン $P = \langle p_1, \dots, p_m \rangle$ に含まれる $e \in E$ の数である．

各情報量関数 $I_j(S, P)$ ($j = s, a, m$) は， $\text{freq}(S, P)$ が小さいほど高い値をとり，各頻度尺度 $F_i(S, P)$ ($i = s, l, r$) と相反する尺度である．

4.3 情報利得尺度

【定義 12】 S を単一系列データベース， P をパターン， $F_i(S, P)$ を重み付き頻度関数， $I_j(S, P)$ を情報量関数とすると，情報利得関数 $G_j^i(S, P)$ を以下で定義する．

$$G_j^i(S, P) = F_i(S, P) \times I_j(S, P)$$

本論文では情報利得を系列パターンの価値と考える．情報量と出現頻度は一般に相反するので，双方のバランスの取れたパターンの利得が高くなる．与えられる最小情報利得 MG と最小出現頻度 MS に対して， MS と MG

以上の利得と出現頻度を持つスケルトンパターンを抽出する．抽出されたパターンを有用パターンと呼ぶ．

先行研究の InfoMiner では，単純頻度 F_s と総和情報量 I_s からなる情報利得だけを取り扱っている．また最小出現頻度 MS も 2 に固定している．しかし頻度尺度と情報量にはこれ以外にも様々なものが考えられ，その有用性は実際の応用に依存して定まると考えられる．本論文では前述の出現頻度と情報量尺度それぞれ 3 種類を組み合わせた 9 種類の情報利得関数を新たに導入し，その性質と有用性について実験的評価を行う．

【定義 13】 情報利得関数 $G_j^i(S, P) = F_i(S, P) \times I_j(S, P)$ が半右逆単調であるとは，頻度関数 F_i が右逆単調である場合を言う．

補題 2 より以下は明らかである．

〔補題 3〕 本論文で導入した 9 個の情報利得関数は全て半右逆単調である．

5. パターン高速抽出技術

5.1 探索空間の再構成

メモリ使用量を削減するため，InfoMiner の探索空間を再構成する．本論文で新たに提案するシステム IFMAP の探索空間は以下の条件を満たす木である． E をアイテム集合， W を抽出パターンの最大長と仮定するとき，

- (1) 各節点は長さ W 以下のパターンである．
- (2) 根は空列パターン ϵ である．
- (3) 葉でない各節点 P は，各 $e \in E \cup \{*\}$ に対する右拡張パターン $P \circ e$ を子節点として持つ．

深さ k の節点は長さ k のパターンであり，深さ W の節点は葉となる．この探索木は， $E \cup \{*\}$ 上の W 以下の長さのスケルトンパターン全てを網羅している．この探索木上で深さ優先探索を行い，葉のパターンの情報利得が閾値以上ならば有用パターンとして抽出する．InfoMiner の探索木は，上の条件 (3) を以下の (3') に置き換えた木であり，明らかに本手法の木よりも分岐数が多い．

- (3') 葉でない各節点 P は，各 $e \in E$ に対して，長さ n 以下の右拡張パターン $P_e = P \circ (*, \dots, *, e)$ の全てを子節点としてもつ．

深さ優先探索時の InfoMiner と IFMAP の最大メモリ使用量の差異は以下の通りである．IFMAP では各節点を持つ子節点の数は $|E| + 1$ である．深さ W になるまで子節点を作成するので，探索時に保持しなければならない節点数の最大数は $W \times (|E| + 1)$ となる．これに対して InfoMiner の場合，まず根の子節点の数は $|E| \times W$ となる．深さ優先探索において保持する節点の最大数を考えているので，最も多くの孫節点を作る節点を次の探索点として選択する．条件 (3') より，短いパターンの節点の子節点が多くなるので，最初は長さ 1 のパターンを持つ子節点を選ぶ．その節点を作る孫節点の数は $|E| \times (W - 1)$ となる．同様にして次に選んだ節点を持つ子節点数は

頻度と単純頻度に基づくそれぞれの利得尺度は全く同じ順位付け効果を持ち，差異は生じない．冗長な議論を避けるために，本論文では相対頻度は考察しない．

$|E| \times (W - 2)$ となる．このようにして探索が深さ W に到達したときの保持節点数は，以下ようになる．

$$\sum_{i=1}^W (|E| \times (W - i + 1)) = \frac{|E| \times (W^2 + W)}{2}$$

以上から InfoMiner と IFMAP が保持すべき節点の最大数はそれぞれ $O(|E| \times W^2)$, $O(|E| \times W)$ となる． W が大きい場合，本提案法の方が格段に効率が良い．

5.2 情報利得に基づく枝刈り手法

有用パターンの探索は，深さ優先戦略に従って探索木を順次拡張しながら行う．効率化のために以下の二つの枝刈り手法を導入する．

パターン拡張判定： 節点（パターン）を拡張して子節点を生成するかどうか判定する．

アイテム選別判定： パターン拡張をする際に用いるアイテムを選別判定する．

InfoMiner ではアイテム選別のみを行うが，IFMAP ではパターン拡張判定を新たに導入し併用する．またアイテム選別も精密化し，より効果的なものに改良する．

§1 パターン拡張判定

パターン拡張判定では「どのように右拡張しても，将来的に最小情報利得値を超えることは不可能なパターンであるか否か」を判断する．評価尺度に逆単調性があれば，アプリアリ型の効果的な枝刈り [Agrawal 95, Pei 04, Iwanuma 05] が適用できるが，残念ながら，殆んどの情報利得は逆単調性を持たない^{*4}．逆単調性を用いずに探索空間を絞り込むために，[Yang 04] では最小必要頻度という概念を導入している．これは，パターン P をうまく右拡張したとき，最小情報利得 MG を超えるために必要な P の出現回数である．本論文では，9 個の情報利得尺度を統一に取り扱うために，Yang らの手法を一般化する．

【定義 14】 S を単一系列データベース， P をパターン， W を最大パターン長， M を S と P に対して非負数を返す関数とする． $M_max(S, P, W)$ を以下の式で定める．

$$M_max(S, P, W) = \max_{P' \in \text{r-ext}(S, P, W)} (M(S, P'))$$

但し $\text{r-ext}(S, P, W)$ は， S に出現している P の長さ W 以下の右拡張パターン全ての集合である．

【例 3】 S を系列 $S = \langle a b c b f i b a c b f c d i a \rangle$ とし， $P = \langle c \rangle$ ， $W = 3$ に対する総和情報量 I_s に対する $I_{s_max}(S, P, W)$ を算出する．まず各アイテムの出現頻度と自己情報量を求める (表 1 参照)．下線で示される S 中の系列 $\langle c \rangle$ に対して，長さ 3 の右拡張パターンは 2 重下線で示される $\langle cbf \rangle$ と $\langle cdi \rangle$ の 2 つとなる．

$$\langle a b \underline{c} \underline{b f i} b a \underline{c} \underline{b f c} \underline{d i} a \rangle$$

*4 最小情報量 I_m は定義より明らかに逆単調である．そのため I_m を用いた 3 つの情報利得も全て右逆単調となり，強力なアプリアリ型の探索空間の枝刈りが行える．

表 1 各アイテムの頻度と自己情報量

アイテム e	$\text{freq}(S, \langle e \rangle)$	$\text{info}(S, e)$
a	3	0.893
b	4	0.738
c	3	0.898
d	1	1.511
i	2	1.125
f	2	1.125

よって $I_{s_max}(S, \langle c \rangle, 3)$ は以下ようになる．

$$\begin{aligned} I_{s_max}(S, \langle c \rangle, 3) \\ &= \max\{I_s(S, \langle cbf \rangle), I_s(S, \langle cdi \rangle)\} \\ &= \max\{2.86, 3.53\} = 3.53 \end{aligned}$$

なお I_{s_max} は，[Yang 01, Yang 04] の I_Max の精密化になっており，その詳細は 5.3 節で述べる．

【定理 1】 S を単一系列データベース， P をパターン， W を最大ウィンドウ幅， MG を最小情報利得とする．また情報利得 $G_j^i(S, P) = F_i(S, P) \times I_j(S, P)$ は半右逆単調であると仮定する．このとき頻度値 $F_i(S, P)$ が

$$\frac{MG}{I_{j_max}(S, P, W)} > F_i(S, P) \quad (1)$$

ならば， P の長さ W 以下のいかなる右拡張パターン P' に対しても， $G_j^i(S, P') < MG$ である．

《証明》 P の右拡張パターンの中で，最大の情報量 I_j を持つパターンを P_{max} とする．このとき $MG \leq G_j^i(S, P_{max})$ ならば，定義より $MG \leq F_i(S, P_{max}) \times I_j(S, P_{max})$ となる．これを整理すれば以下を得る．

$$\frac{MG}{I_j(S, P_{max})} \leq F_i(S, P_{max})$$

仮定より G_j^i が半右逆単調であるから， F_i は右逆単調であり， $F_i(S, P_{max}) \leq F_i(S, P)$ が成立つ．よって

$$\frac{MG}{I_j(S, P_{max})} \leq F_i(S, P) \quad (2)$$

が成立つ．よって式 (2) は， P のある右拡張が MG を満たすための必要条件となっている．また，定義 14 より， $I_j(S, P_{max}) = I_{j_max}(S, P, W)$ である．

以上の対偶より， P が式 (1) を満たすならば，いかなる P の右拡張パターン P' においても， $G_j^i(S, P')$ は MG を超えることは無いことが証明できた．□

【定義 15】 単一系列データベース S ，パターン P ，最大パターン長 W ，最小情報利得 MG ，情報利得 $G_j^i(S, P) = F_i(S, P) \times I_j(S, P)$ に対して， P の最小必要頻度 $\text{min-rep}(S, P, W, MG)$ を次の式で定義する．

$$\text{min-rep}(S, P, W, MG, G_j^i) = \frac{MG}{I_{j_max}(S, P, W)}$$

以下では S, W, MG, G_j^i が明らかな場合には，適宜省略して $\text{min-rep}(P)$ のように略記する．

情報利得 G_j^i が半右逆単調である場合、定理 1 と定義 15 より、 $\min\text{-rep}(P) \leq F(S, P)$ を満たす P のみを拡張しても、探索の完全性は保証される。

〔例 4〕例 3 の S 、および単純頻度 F_s と総和情報量 I_s を組み合わせた情報利得 G_s^i に対して、 $MS = 2, MG = 5$ としたときの $\min\text{-rep}(\langle c \rangle)$ を求める。例 3 より、 $I_{s\text{-max}}(S, \langle c \rangle) \doteq 3.53$ である。

$$\min\text{-rep}(\langle c \rangle) = \frac{MG}{I_{s\text{-max}}(S, \langle c \rangle, 3)} = \frac{5.0}{3.53} \doteq 1.42$$

$F_s(S, \langle c \rangle) = 2 > \min\text{-rep}(\langle c \rangle)$ であるから、候補パターン $\langle c \rangle$ の右拡張の情報利得は閾値を越す可能性があり、探索木において節点 $\langle c \rangle$ の拡張が行われることになる。

§2 アイテム選別判定

探索の効率化のために、拡張パターンが最小情報利得を超える可能性がある拡張アイテムの選別を考える。

〔定理 2〕 S を単一系列データベース、 P をパターン、 W を最大ウィンドウ幅、 MG を最小情報利得とする。また情報利得 $G_j^i(S, P) = F_i(S, P) \times I_j(S, P)$ は半右逆単調であると仮定する。このときアイテム $e \in E$ に対して、 $F_i(P \circ e)$ が

$$\frac{MG}{I_{j\text{-max}}(S, P, W)} > F_i(P \circ e) \quad (3)$$

を満たすならば、 $P \circ e$ のいかなる右拡張パターン $(P \circ e)'$ に対しても $G_j^i(S, (P \circ e)') < MG$ である。

《証明》 P の右拡張パターン中で最大の情報量 I_j を持つパターンを P_{max} とし、 $P \circ e$ の右拡張パターンで最大の情報量 I_j を持つパターンを $(P \circ e)_{max}$ とする。 $P \circ e$ の右拡張パターンの集合は、 P の右拡張パターンの集合の部分集合であるから、明らかに $I_j(S, P_{max}) \geq I_j(S, (P \circ e)_{max})$ である。よって、

$$\frac{MG}{I_{j\text{-max}}(S, P \circ e, W)} \geq \frac{MG}{I_{j\text{-max}}(S, P, W)}$$

となる。従って、 G_j^i は半右逆単調であるので、式 (3) が成り立つ場合は定理 1 が適用でき、 $P \circ e$ のいかなる右拡張パターン $(P \circ e)'$ に対しても、 $G_j^i((P \circ e)')$ は MG を超えることはないことが証明できる。□

本論文で提案した 9 個の情報利得 G_j^i は全て半右逆単調である。よってその全ての G_j^i の探索木の拡張判定（枝刈）には同一の手法が適用できる。 $\min\text{-rep}(P) \leq F_i(S, P)$ を満たすパターン P だけ（パターン選別判定）を、 $\min\text{-rep}(P) \leq F(S, P \circ e)$ を満たすアイテム $e \in E \cup \{*\}$ だけ（アイテム選別判定）を用いて拡張すればよい。枝刈の効果は情報利得ごとに違うと考えられる。特に使用する重み付き頻度尺度の影響は大きいと推測され、次章で実験的に評価・考察を行う。

〔例 5〕例 4 と同じ条件を考える。 $\min\text{-rep}(\langle c \rangle) = 1.42$ であった。一方 $F_s(S, \langle cb \rangle) = 2$ かつ $F_s(S, \langle cd \rangle) = 1$ であるので、 $\min\text{-rep}$ 以上の頻度尺度値を持つアイテムは b

のみである。よって探索木の節点 $\langle c \rangle$ における子節点はパターン $\langle cb \rangle$ と $\langle c, * \rangle$ の 2 つとなる。

5.3 $I_{s\text{-max}}$ の精密化

InfoMiner では情報量尺度として総和情報量を用いて、 $I_{s\text{-max}}$ を算出している。本論文では探索空間をより狭めるために、 $I_{s\text{-max}}$ の見積もり精度を改良している。本論文の $I_{s\text{-max}}(S, P, W)$ は S に実在する P の拡張パターンの最大情報量であり、以下で計算する。

$$I_{s\text{-max}}(S, P, W) =$$

$$I_s(S, P) + \max_{i \in Oc(S, P)} \left\{ \sum_{j=i+1}^{W+i-1} \max_{e \in s_j} (\text{info}(S, e)) \right\}$$

一方、 $I_{s\text{-max}}(S, P, W)$ では実際に出現しない仮想のパターンの情報量も用いるため、本手法より大きく算出される。具体的には以下の式で算出できる。

$$I_{s\text{-max}}(S, P, W) = I_s(S, P) + \sum_{i=1}^{W-|P|} \max_{e \in E(S, P, i)} (\text{info}(S, e))$$

但し $E(S, P, i)$ は、 S 上でパターン P の i だけ後ろに出現するアイテム $e \in E$ の集合である。例えば、単一系列データベース $S = \langle \underline{a} \underline{b} \underline{c} \underline{e} \underline{a} \underline{b} \{ \underline{d} \underline{e} \} \underline{a} \underline{b} \underline{c} \{ \underline{a} \underline{d} \} \rangle$ とパターン $\langle \underline{ab} \rangle$ があつたとき、 $E(S, \langle \underline{ab} \rangle, 1) = \{ \underline{c}, \underline{d}, \underline{e} \}$ および $E(S, \langle \underline{ab} \rangle, 2) = \{ \underline{a}, \underline{d}, \underline{e} \}$ である。一般に $I_{s\text{-max}}(S, P, W)$ は S 上に実在するパターンの情報量より大きくなる。 $I_{s\text{-max}}$ が大きいと $\min\text{-rep}$ の値が小さくなるので、枝刈りが起き難くなる。なお、上の 2 つの $I_{s\text{-max}}$ はどちらも、 S を一回スキャンするだけで求まるので計算コストには本質的な差は無い。

6. IFMAP: 有用パターン高速抽出システム

以上の考察に基づいて開発した系列パターン抽出システム IFMAP のアルゴリズム概要を図 1 に示す。

深さ優先で探索木を構築し、有用パターンを抽出する。メイン関数は $IFMAP(P, S, Oc)$ であり、まず各節点のパターン P と S 上の出現位置 Oc に対して、情報利得 $G_j^i(S, P)$ の計算と、閾値 MS 以上の有用パターンの出力を行う。次に $\min\text{-rep}$ 値を基に、 P の拡張の是非の判定*5、右拡張用のアイテムの選別を行い、節点 P の子節点が追加される。 Oc' は右拡張した $P \circ e$ の出現位置の集合である（関数 δ は定義 6 参照）。 Oc を適宜更新しながら再起呼出し（子節点の探索）時に引き渡すことにより、PrefixSpan[Pei 04] で提案された擬似射影 (pseudo-projection) による系列データベースの縮小を行っており、各種の演算を高速化している。 $F_i(S, P, Oc)$ や $E(S, P, 1, Oc)$ は、頻度関数 $F_i(S, P)$ や $E(S, P, 1)$ と基本的に同じ働きをする関数であるが、 Oc や Oc' を利用して高速計算を行うものであ

*5 最小情報量を用いる場合には、情報利得は右逆単調となるので、 $\min\text{-rep}$ を用いずに、 P の利得が MG を超えているか否かで拡張の是非を判断することもできる。

Input: 系列データベース S , アイテム集合 E ,
最大抽出パターン長 W , 最小情報利得 MG ,
最小出現頻度 MS ,
情報利得関数 $G_j^i(S, P) = F_i(S, P) \times I_j(S, P)$
Output: S 中の, MG 以上の情報利得を持つ長さ W 以下の
パターンすべて

```

1: Main
2: forall  $e \in E$ 
3:   freq( $S, \langle e \rangle$ ) および info( $S, e$ ) を計算し, 保持する;
4: end
5:  $Oc := \{m \mid 1 \leq m \leq |S|\}$ ;
6: forall  $e \in E$ 
7:    $P := \langle e \rangle$ ;
8:    $Oc' := \{m \in Oc \mid \delta(\text{win}(S, m, |P|), P) = 1\}$ ;
9:   if min-rep( $P$ )  $\leq F_i(S, \langle e \rangle, Oc')$  then
10:    IFMAP( $P, S, Oc'$ );
11: end
12: end
13:
14: IFMAP( $P, S, Oc$ )
15: begin
16: if tail( $P$ )  $\neq *$  and  $G_j^i(S, P) \geq MG$  then
17:    $P$  を出力
18: if  $|P| < W$  and min-rep( $P$ )  $\leq F_i(S, P, Oc)$  then
19:   forall  $e \in E(S, P, 1, Oc) \cup \{*\}$ 
20:      $Oc' := \{m \in Oc \mid \delta(\text{win}(S, m, |P \circ e|), P \circ e) = 1\}$ ;
21:     if min-rep( $P$ )  $\leq F_i(S, P \circ e, Oc')$  then
22:       IFMAP( $P \circ e, S, Oc'$ );
23:     end;
24: end;
25: end

```

図 1 IFMAP のアルゴリズム

る．16 行目の tail(P) は P の末尾アイテムを表している．
17 行目ではスケルトンパターンのみを抽出する．

7. 評価実験

本章では提案手法の実験的評価を行う．IFMAP は C 言語で実装して使用している．実験マシンは Pentium3GHz, メモリ 2GB である．実験データは毎日新聞 2006 年社会面記事データを使用し, 各記事に出現する単語から, TFI DayF 値 [多田 09] がいずれかの日で上位 50 位以内となったものをアイテムとして抜き出している．アイテムの種類は 16,895 である．単一系列データベースは, 日毎に出現したアイテム (単語) を集めた集合を系列の要素としており, 系列の長さは 364, 系列に出現するアイテムの出現総数は 38,435 である．

7.1 頻度基準に基づくマイニングとの比較

情報利得基準と頻度基準によるそれぞれの抽出パターン数を比較し, 情報利得基準が無駄なパターンの抽出をどれだけ抑制できるかを調べる．最大パターン長を 7, 最小出現頻度を 2 に固定し, 情報利得基準で上位 1 個, 500

表 2 情報利得法と頻度尺度法との比較

	情報利得法		頻度尺度法	
	MG	抽出数	MS	抽出数
単純頻度と 総和情報量 G_s^s	45.0	2	20	2,434
	33.8	511	15	8,330
	25.5	5,010	8	442,925
対数頻度と 総和情報量 G_s^l	11.90	1	9	160,395
	10.63	500	7	940,971
	9.91	5,067	7	940,971
対数化相対頻度 と総和情報量 G_s^r	0.750	4	2	TO
	0.721	505	2	TO
	0.703	5,182	2	TO
単純頻度と 平均情報量 G_a^s	8.49	1	39	21
	4.60	506	11	42,680
	3.43	5,006	8	442,925
対数頻度と 平均情報量 G_a^l	1.56	1	11	42,680
	1.37	510	8	442,925
	1.27	5,015	7	940,971
対数化相対頻度 と平均情報量 G_a^r	0.0940	1	2	TO
	0.0902	528	2	TO
	TO	5,000	2	TO
単純頻度と 最小情報量 G_m^s	13.00	1	34	56
	6.78	501	13	17,780
	4.39	5,016	8	442,925
対数頻度と 最小情報量 G_m^l	2.10	1	13	17,780
	1.61	502	8	442,925
	1.36	5,054	5	8,314,722
対数化相対頻度 と最小情報量 G_m^r	0.1190	1	13	17,780
	0.1106	632	2	TO
	0.1019	5,168	2	TO

個, 5,000 個のパターンを抽出するように最小情報利得 MG を調節し系列マイニングを行う．次に, 抽出されたパターンを全てを含むように最小出現頻度 MS を調整し, 頻度尺度のみによってパターンを抽出する．実験結果を表 2 に示す．表中の TO とは, 制限時間 12 時間以内に, 指定の個数のパターンの抽出または閾値を超える全パターンの抽出が終わらなかったことを示す．情報利得上位のパターンを頻度尺度だけを用いて見つけ出すことは困難であることが確認できる．特に相対化対数頻度を用いた場合には極めて難しいことが確認できた．単純頻度と対数頻度を比較すると, 単純頻度は情報利得における重みがかかなり大きいことが分かる．単純頻度を用いた場合には, 情報利得が高い系列パターンであっても, その出現頻度はかなり高くなっている．7.2 節で示すように, これらの高頻度パターンは, 少数の非常に出現頻度の高いアイテム (単語) から構成されていることが多い．対数頻度では半分以下の出現頻度をもつパターンが利得最上位にランク付けされており, この意味でより望ましい性質を持っていると考えられる．

表3 情報利得上位 100 個の抽出パターンの特徴

	パターン長	情報量	出現頻度
G_s^s	4~7 (6.24)	1.5~2.5 (2.04)	15~25 (20.64)
G_s^l	7 (7)	1.0~2.0 (1.32)	5~15 (8.58)
G_s^r	7 (7)	6~7 (6.17)	2 (2)
G_a^s	2~7 (3.57)	0.1~0.3 (0.24)	20~60 (27.85)
G_a^l	3~7 (5.55)	0.3~0.5 (0.46)	5~25 (10.12)
G_a^r	5~7 (6.24)	0.5~0.8 (0.78)	2~10 (2.08)
G_m^s	2~5 (2.59)	0.2~0.4 (0.33)	20~60 (32.09)
G_m^l	2~6 (2.67)	0.4~0.6 (0.49)	10~40 (17.32)
G_m^r	2~6 (2.36)	0.4~1.0 (0.82)	2~35 (4.88)

7.2 情報利得の特徴比較

各情報利得 G_j^i ($i = s, l, r$ 及び $j = s, a, m$) を用いて抽出したパターンの特徴を比較することで、情報量尺度と頻度尺度を組み合わせた 9 種の情報利得を調べる。比較する特徴量は以下の 4 つである。

- (1) パターン長
- (2) パターンの情報量値
- (3) パターンの出現頻度値
- (4) パターンに含まれるアイテム

抽出パターン全てを比較するのはあまり意味が無いため、情報利得の上位 100 パターン（同率のパターンを含む）に対して比較を行った。最大パターン長を 7、最小出現頻度を 2 とし、各抽出パターンの特徴 (1), (2), (3) を表 3 に、特徴 (4) を表 4 に示す。

表 3 は、各特徴値の範囲と平均値を表記している。パターン長を比較すると、平均情報量や最小情報量を用いた場合に短いことが分かる。特に最小情報量の場合にこの傾向が著しい。これは平均情報量や最小値情報量を用いた場合、系列の長さに関係なく情報量が計算されるためと思われる。また出現頻度は、平均情報量や最小情報量を用いた場合に高くなっている。長さが短いパターンほど出現頻度が高くなる傾向があるためと思われる。

表 4 は、上位 100 個のパターンに含まれるアイテム（単語）とその出現数を示している。「種類数」は上位 100 パターンに含まれるアイテムの種類の数である。例えば、 G_s^s のアイテム種類数は 3 であるので、上位 100 個の抽出パターンは「サッカー」「ドイツW杯」「全国高校野球」および仮想アイテム「*」の計 4 個の単語の組み合わせだけで構成されることを示している。

InfoMiner の利得基準 G_s^s （単純尺度と総和情報量の組合せ）は、高頻度な単語からなるパターンに高い値を出しやすく、単語種類（語彙）が極端に小さくなるので、新聞記事コーパスの処理は不向きと判断される。逆に、基準 G_m^r （対数化相対頻度と最小情報量）は、あまり出現しない単語からなる短いパターンに大きな評価値を与えるために単語種類（語彙）はかなり大きい。229 種類もの単語が出現しており、個々の単語に有用性を認めることは困難と思われる。これ以外の利得基準はどれも特徴的な単語からなるパターンを抜き出していると判断され

る。実際、「ドイツW杯」「テレクラ放火殺人事件」「偽装」「発注工事談合」「教委贈収賄」「高校履修不足」「ライブドア強制捜査」「姉齒事件」etc ... など 2006 年度に特徴的な単語を含むパターンが抜き出されている。

G_s^s と G_m^r を除いた残りの 7 種類の利得基準のうち、それが新聞記事コーパス中の有用系列パターンの抽出に適しているかを考えてみる。表 4 より、対数化頻度を用いる G_s^l と G_a^l は抽出パターンに出現する単語の種類が 7 と 8 でかなり小さく、パターンを表現する語彙の多様性の点から疑問があると判断される。同様に G_m^s も単語種類は 12 と比較的少なく、また抽出パターンの最大長が 5、平均長が 2.59（表 3 参照）となっている。非常に短いパターンを抽出しており、あまり有用な利得基準ではないと考えられる。残りの 4 つの基準 G_s^r , G_a^s , G_a^r , G_m^l の中で、対数化相対頻度を用いる G_s^r と G_a^r はパターンに出現する単語も似ており、パターンの長さや出現頻度もかなり類似している。この 2 つは同種のものとするのが妥当である。単純頻度と平均情報量を組み合わせた G_a^s は、 G_s^r と G_a^r とは明らかに異なった特性を示している。また G_m^l も抽出パターンの平均長が 2.67 と短いことを除けば、抽出単語も比較的豊富であり、有効性があると考えられる。次節で述べるように、これら 4 つの利得基準の計算の難しさ（計算時間）には大きな差があり、扱うデータによってはこれが大きな要因・障害になる。以上から、新聞記事コーパスからの有用系列パターンの抽出には G_a^r （あるいは G_s^r ）と G_a^s 、あるいは抽出パターンが短くてもよい場合には G_m^l をその目的と対象データ、状況に応じて使い分けることが妥当と考えられる。

7.3 IFMAP の枝刈り性能の評価

アイテム選別とパターン拡張判定をそれぞれ単独で適用する場合、両方を同時に適用する 3 つの場合の実行時間を比較する。最大パターン長 $W = 7$ 、最小出現頻度 $MS = 2$ とし、最小情報利得 MG を変化させ計算し、実行時間を計測する。結果を図 2 から図 8 に示す。縦軸が処理時間、横軸が最小情報利得 MG の値である。また、各グラフの MG は、右端でパターンを 1 つ、左端でパターンを約 30,000 ほど抽出する値である。

全ての場合において、2 つの枝刈り技法を併用する場合が最も速いことが確認された。特に計算時間が掛かる場合において高速化の効果が大きく有用である。計算時間の観点から眺めると、単純頻度に基づく基準が最も高速で、次に対数頻度となり、対数化相対頻度は極端に遅くなっている。対数化相対頻度を用いる G_s^r と G_a^r 基準を用いた実験の結果は、2 つの枝刈り技法を併用しても抽出パターンが 1 個の場合においてさえ計算時間が 1 万 CPU 秒を越えるので、表は作成していない。本論文で提案した最小必要頻度による枝刈りは頻度値の絶対値が大きいほど効果的に働く。対数化相対頻度の絶対値はかなり小さいため、あまり高速化できないことが確認された。

表 4 情報利得上位 100 個の抽出パターンに出現するアイテムの例

	情報利得上位 100 パターンに含まれるアイテムとその出現回数	種類数
G_s^s	サッカー (71), ドイツW杯 (45), 全国高校野球 (41)	3
G_s^l	発注工事談合 (57), いじめ (48), 手 (45), 全国高校野球 (41), 宇和島 (34), どこ (26), いじめ自殺予告 (17), 救い (14)	8
G_s^r	教育 (15), センバツ (14), 理事長 (14), テレクラ放火殺人事件 (13), 教委贈収賄 (12), ..., ライブドア強制捜査 (4), 姉齒事件 (4), ..., 伊福部昭 (2), 台湾補償 (2), リスニングトラブル (2)	33
G_a^s	北海道 (121), 拉致 (118), 滋賀 (96), 東京地裁 (84), 偽装 (74), サッカー (71), ..., 発注工事談合 (57), ..., 全国高校野球 (41), 飛鳥会事件 (35), 高校履修不足 (27), 小1 男児殺害 (25)	17
G_a^l	いじめ (48), 手 (45), 全国高校野球 (41), どこ (26), 都市対抗野球 (18), いじめ自殺予告 (17), 救い (14)	7
G_a^r	海辺 (27), 救い (14), テレクラ放火殺人事件 (13), 教委贈収賄 (12), ..., ライブドア強制捜査 (4), 姉齒事件 (4), ..., 白馬岳遭難 (2), 伊福部昭 (2), 台湾補償 (2), リスニングトラブル (2)	39
G_m^s	北海道 (121), 拉致 (118), 偽装 (74), サッカー (71), 長崎 (71), 愛媛 (69), 知事 (62), 発注工事談合 (57), 談合 (55), 藤里 (48), ドイツ W 杯 (45), 全国高校野球 (41)	12
G_m^l	トリノ冬季五輪 (60), ドイツ W 杯 (45), 全国高校野球 41, 宇和島 (34) パロマ湯沸かし器事故 (31), 高校履修不足 (27), ..., 未来通信 (17), 錬金術 (16), 救い (14), GW(13), 春センバツ (13)	23
G_m^r	全国高校野球 (41), 都道府県大会 (18), 救い (14), 春センバツ (13), 人々(8), アオギリ紀行 (8), 病巣 (8), ..., 大阪 C 型肝炎訴訟 (5), ..., 宇和島誤認逮捕事件 (3), 福島知事辞職 (3), 教育再生会議 (3),...	229

図 2 単純頻度と総和情報量 G_s^s における枝刈効果

情報量を軸に考えると, 最小情報量, 総和情報量, 平均情報量の順に高速である. 最小情報量に基づく利得基準は, 短い系列パターンを有用として抽出する傾向が強く, そのため他と比べて探索木の深さが浅くなると考えられる. 前節の結果と合わせてみると, 大規模データに対して迅速なパターン抽出処理を望む場合には, G_a^s や G_m^l などの利得基準を用いることが適切と判断される.

8. む す び

本論文では, 大規模な系列データベースから頻度と情報量のバランスの取れた系列パターンを抽出することを目的とし, 新たに幾つかの情報利得基準と, それに基づく系列パターンを高速化する技法をしめした. また包括的な評価実験を行い, 個々の利得基準の特性や枝刈手法の性能特性を示した.

我々は長年新聞記事コーパスから有用イベント系列を自動抽出する研究を続けているが, これに適した情報利得基準も, 本研究における実験の評価の結果からある程

図 3 対数頻度と総和情報量 G_s^l における枝刈効果 図 4 単純頻度と平均情報量 G_a^s における枝刈効果

図5 対数頻度と平均情報量 G_a^l における枝刈り効果 図8 相対化対数頻度と最小情報量 G_m^r における枝刈り効果 図6 単純頻度と最小情報量 G_m^s における枝刈り効果 図7 対数頻度と最小情報量 G_m^l における枝刈り効果

度明確になった。即ち、先行研究の InfoMiner で用いられた単純頻度と総情報量を用いる利得基準は、少なくとも新聞記事コーパス中の有用系列パターン発見には適切ではなく、より適した複数の利得基準があることが分かった。日本語固有の性質は殆どの何も使ってはいないので、英語記事コーパスや他のデータでも同様な性質と特徴を持つと予想している。この実証実験は、今後の課題としたい。

一方で新聞記事コーパスからの有用系列パターンの抽出という観点からは、本提案手法はまだ十分ではなく、よ

り有効な新しい利得基準の開発や、本手法と併用する別種のフィルタリング法の開発などが今後必要である。また出現頻度の重みが小さい情報利得においては、半右逆単調性にもとづく枝刈り効果が減少している。これを補うための新たな技術の開発も今後の課題である。

本論文で提案開発した手法は、その幾つかは他の枠組みにも直接適用でき、また有効であると予想している。例えば種々の情報量利得尺度はアイテム集合抽出問題の枠組みでも定義可能であり、実際にフィルタリングに有効であると思われる。一方で、高速化のための枝刈り手法は単一系列データマイニング特有の問題に対処するために最適化されており、これをそのままアイテム集合抽出問題に適用することは難しく、また意味も薄いと思われる。

謝 辞

本研究の一部は、JST 平成 21 年度シーズ発掘試験 (06-002) 及び科研費基盤研究 A(No.20240016) の援助を受けている。実験の補助をしてくれた本学卒研究生の宇野祥吾君 (現在:(株) ニューメディア総研) に感謝する。また日頃から議論を頂く本学の鍋島英知准教授、山本泰生助教に深く感謝する。

◇ 参 考 文 献 ◇

- [Agrawal 95] R. Agrawal and R. Srikant: Mining Sequential Patterns. *Proc. 1995 Inter. Conf. on Data Engineering (ICDE'95)*, pp.3-14 (1995)
- [Blanchard 05] J. Blanchard, F. Guillet, R. Gras and H. Briand: Using Information-theoretic Measures to Assess Association Rule Interestingness. *Proc. of IEEE Inter. Conf. on Data Mining (ICDM 2005)*, pp.66-73 (2005)
- [Cao 07] H. Cao, N. Mamoulis and D. W. Cheung: Discovery of Periodic Patterns in Spatiotemporal Sequences. *IEEE Transactions on Knowledge and Data Engineering*, Vol.19 No.4,p.453-467 (2007)
- [Geng 06] L. Geng and H. J. Hamilton: Interestingness Measures for Data Mining: A Survey. *ACM Computing Surveys*, Vol.38, No.3 (2006)
- [Han 98] J. Han, W. Gong and Y. Yin: Mining Segment-Wise Periodic Patterns in Time-Related Databases. *Proc. Inter. Conf. on Knowledge Discovery and Data Mining*, pp.214-218 (1998)
- [He 08] Z. He, X. S. Wang, B. S. Lee and A. C. H. Ling: Mining Partial Periodic Correlations in Time Series, *Knowledge and Inform-*

mation Systems, Vol.15, No.1, pp.31-54 (2008)

- [市川 08] 市川博規, 岩沼宏治, 鍋島英知: 因果関係に注目した
コンフィデンスに基づく高速系列データマイニング. 第 68 回 人
工知能基本問題研究会 (SIG-FPAI), 2008
- [Iwanuma 05] K. Iwanuma, R. Ishihara, Y. Takano and
H. Nabeshima: Extracting Frequent Subsequences from a Sin-
gle Long Data Sequence: A Novel Anti-Monotonic Measure and
a Simple On-line Algorithm. *Proc. of IEEE Inter. Conf. on Data
Mining (ICDM 2005)*, pp.186-193 (2005)
- [Lee 08] W.-J. Lee, J.-Y. Jiang and S.-J. Lee: Mining Fuzzy Periodic
Association Rules. *Data & Knowledge Engineering*, Vol.65 No.3,
pp.442-462, (2008)
- [Lonardi 07] S. Lonardi, J. Lin, E. Keogh and B. Chiu: Efficient Dis-
covery of Unusual Patterns in Time Series. *New Generation Comput-
ing*, Vol.25 No.1, pp.61-93 (2007)
- [Mannila 95] H. Mannila H. Toivonen and A. I. Verkamo: Discov-
ering frequent episodes in sequences. *Proc. the First Inter. Conf.
on Knowledge Discovery and Data Mining (KDD'95)*, pp.210-215
(1995)
- [Pei 04] J. Pei, J. Han, B. Mortazavi-Asl, J. Wang, H. Pinto, Q. Chen,
U. Dayal and M.-C. Hsu: Mining Sequential Patterns by Pattern-
Growth: The PrefixSpan Approach. *IEEE Trans. on Knowledge and
Data Engineering*, Vol.16, No.10 pp.1-17, (2004)
- [Srikant 96] R. Srikant and R. Agrawal: Mining Sequential Patterns:
Generalizations and Performance Improvements. *Proc. the Fifth Int'l
Conf. on Extending Database Technology (EDBT)*, pp.3-17 (1996)
- [多田 09] 多田知道, 岩沼宏治, 鍋島英知: イベント系列マイニン
グを目的とする新聞記事からの時間情報に基づく単語抽出. 人
工知能学会論文誌 Vol.24, No.6, pp.418-493 (2009)
- [Tan 02] P. K. Tan and J. Srivastava: Selecting the right interest-
ingness measure for association patterns. *Proc. the 8th Inter. Conf.
on Knowledge Discovery and Data Mining (KDD 2002)*, pp.32-41
(2002)
- [Yang 01] J. Yang, W. Wang and P. S. Yu: InfoMiner: Mining Surpris-
ing Periodic Patterns. *Proc. of the 7th ACM Inter. Conf. on Knowledge
Discovery and Data Mining (KDD'01)*, pp.395-400 (2001)
- [Yang 03a] J. Yang, W. Wang and P. S. Yu: Mining asynchronous pe-
riodic patterns in time series data. *IEEE Trans. Knowledge and Data
Engineering* Vol.15, pp.613-628, (2003)
- [Yang 03b] J. Yang, W. Wang and P. S. Yu: STAMP: on Discovery
of Statistically Important Pattern Repeats in Long Sequential Data.
Proc. of the Third SIAM Inter. Conf. on Data Mining (SDM'03),
pp.224-238 (2003)
- [Yang 04] J. Yang, W. Wang and P. S. Yu: Mining Surprising Peri-
odic Patterns. *Data Mining and Knowledge Discovery*, Vol.9, No.2,
pp.189-216 (2004)
- [Yeh 09] J.-S. Yeh and S.-C. Lin: A New Data Structure for Asyn-
chronous Periodic Pattern Mining. *Proc. of the 3rd Inter. Conf. on
Ubiquitous Information Management and Communication*, pp.426-
431 (2009)

[担当委員: 浅井 達哉]

2009 年 12 月 24 日 受理

著者紹介

村田 順平

2008 年山梨大学工学部コンピュータ・メディア工学科卒
業。現在、同大学大学院修士課程コンピュータ・メディア
工学専攻在学。系列データマイニングにおける情報量と出
現頻度の融合と、オンライン型データマイニングに関する
研究と開発に従事。

岩沼 宏治 (正会員)

1985 年東北大学大学院・電子及通信工学専攻修士課程修
了。同年山形大学情報工学科助手。現在、山梨大学大学院・
医学工学総合研究部教授。工学博士。人工知能の基礎、特
に非単調論理、定理自動証明、SMT 技術、WEB 知的処
理、系列データマイニング等の研究に従事。1987, 89, 90,
91 年人工知能学会全国大会優秀論文賞受賞。2004 年 FIT
優秀論文賞受賞

大塚 尚貴

2007 年山梨大学大学院修士課程コンピュータ・メディア
工学専攻修了。データマイニング、特に系列データマイ
ニングにおける情報量と出現頻度の融合に関する研究と開発
に従事。現在、三洋電機株式会社勤務。