2G3-2

単一系列データマイニングにおける情報量基準とその補完尺度

An Information Measure and its Complementing ones on a Sequential Data Mining

大柴亮[*1] 岩沼宏治[*2] 山本泰生[*2]
Ryo OSHIBA Koji IWANUMA Yoshitaka YAMAMOTO

[*1]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 Medicine and Engineering, University of Yamanashi

[*2]山梨大学大学院医学工学総合研究部
Interdisciplinary Graduate School of Medicine and Engineering, University of Yamanashi

In this research, we consider new methods for extracting interesting patterns from a long sequential data based on both an information measure and new complementing ones for evaluating pattern interestingness. We implemented a mining system, and report results of evaluation experiment using a newspaper article corpus for verifying the usefulness of the proposed measures.

1. はじめに

本研究では，単一で長大な系列データベースから有用系列パターンを抽出マイニングすることを目的として，情報量基準とそれを補完する新たな系列評価尺度について考察を行う．また，提案尺度の有用性を検証するために，抽出マイニングシステムを実装し，新聞記事コーパスを用いて評価実験を行ったので，その結果を報告する．

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．予ねてより，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研究 [Agrawal 95, Han 98, Pei 04, Iwanuma 05] が盛んに行われている．しかし，応用によっては高頻出の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（社会面) 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」は極めて高い出現頻度を持つ．これらの要素を持つパターンは頻繁に出現しているが，有用な情報ではなく，むしろノイズに近いと考えられる．一方で出現頻度が非常に低い系列パターンは偶発的なものが多く，こちら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．

中頻度の系列を重視するものとして情報量と出現頻度を考慮する系列評価尺度があり，IFMAP[村田 10] などで用いられている．IFMAP では 3 種類の頻度尺度と 3 種類の情報量尺度を組み合わせた 9 つの評価基準を設け，有用な系列パターンの抽出を試みた．しかしながら，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まだ十分ではない．例えば同一のアイテムが連続したパターン（< 全国高校野球 全国高校野球 > など）を多く抽出する傾向がある．同一のアイテムからなるパターンはある期間に話題が集中した現象系列を表すと考えられるが，ある意味で面白みのないパターンである．抽出パターン中に複数種類のアイテムが存在することで，アイテム間の隠れた関係を読み取ることができる可能性があるからだ．

本論文では上記の問題の解決に取り組む．IFMAP によりランク付けされた抽出パターン上位 1000 位に対し，別の補完的な評価尺度を導入して系列の抽出を行う．系列パターン抽出システム IFMAP を拡張実装し，新聞記事コーパスを用いた評価実験を行う．

本論文の構成は以下の通りである．2 章は準備である．3 章で提案する補完尺度について説明し，4 章で実験結果と考察を行う．5 章はまとめである．

連絡先: 大柴亮
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
〒 400-8511 山梨県甲府市武田 4-3-11
g10mk008@yamanashi.ac.jp

2. 準備

定義 1 全てのアイテムの集合を $E = \{e_1, e_2, \cdots, e_n\}$ 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 $S = \langle s_1, s_2, \cdots s_m \rangle$ 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．

以降，単一要素からなる系列要素は括弧を省略する．即ち，系列 $\langle \{a\}, \{a,b\}, \{c\}, \{d\}, \{a,b,e\}, \{a,c\} \rangle$ 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び，$E^+ = E \cup \{*\}$ を**拡張アイテム集合**と呼ぶ．このときパターン P とは，拡張アイテム集合 E^+ 中のアイテムの順序付きリスト $P = \langle p_1, p_2, \cdots, p_m \rangle$ 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．
IFMAP ではスライド窓機構を用いてパターン抽出を行う．

定義 3 ([Iwanuma 05]) 単一系列データベース $S = \langle s_1, \ldots, s_m \rangle$ 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** $\mathrm{win}(S,i,k)$ を以下のように定義する．

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

定義 4 E をアイテム集合とするとき，$E^+ \times 2^E$ 上の二項関係 $\lhd$ を以下のように定める．

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

定義 5 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** $\mathrm{freq}(S,P)$ を以下で定める．

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

ここで δ は以下の関数と定める．

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

P の S 中の**出現位置**とは，$\delta(\mathrm{win}(S,i,|P|),P) = 1$ 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．

2.1 頻度尺度

IFMAP では基本出現頻度を基にして，3 つの重み付き頻度関数を用いる．

定義 6 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き，S 中の P の**単純頻度** $F_s(S,P)$，**対数頻度** $F_l(S,P)$，**対数化相対頻度** $F_r(S,P)$ をそれぞれ以下のように定める．

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

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

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

MS は出現頻度の閾値である．基本出現頻度が MS 未満ならば，$F_s(S,P) = 0$ 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．そこで対数頻度と対数化相対頻度の２つの頻度尺度を導入している．対数化相対頻度は，相対頻度 $\frac{F_s(S,P)}{|S|}$ 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である．明らかに各重み付き頻度は P の基本出現頻度 $\mathrm{freq}(S,P)$ に比例して増減する．

2.2 情報量利得尺度

定義 7 単一系列データベース S 上のアイテム $e \in E^+$ の**自己情報量** $\mathrm{info}(S,e)$ を以下のように定義する．

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

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 $\mathrm{freq}(S,*) = |S|$ なので，$\mathrm{info}(S,*) = 0$ 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 8 S を単一系列データベース，$P = \langle p_1, \cdots, p_m \rangle$ 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．

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

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

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

但し，$skel(P)$ は，パターン $P = (p_1, \cdots, p_m)$ 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，$\mathrm{freq}(S,P)$ 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．

定義 9 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き，**情報利得関数** $G_j^i(S,P)$ を以下で定義する．

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

IFMAP 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．

3. 補完尺度

IFMAP により抽出された有用パターンを補完する尺度を提案する．

3.1 単語異なり尺度

定義 10 S を単一系列データベース，P をパターンとするとき，**単語異なり尺度** $W_{diff}(P)$ を以下のように定義する．

$$W_{diff}(P) = \frac{p_{kind}}{|P|}$$

但し，p_{kind} は，パターン $P = (p_1, \cdots, p_m)$ に含まれる $e \in E$ の種類数である．

単語異なり尺度はパターン P 中に含まれるアイテム（単語）の豊富さを調べる単純な尺度である．

表 1: 情報量利得と単語異なり尺度，負出尺度における上位 10 位の系列

	$G_a^s(S,P)$	W_{diff}	$Negative$
1	< 発注工事談合 発注工事談合 >	< ドイツW杯 サッカー >	< サッカー 小１男児殺害 >
2	< 全国高校野球 全国高校野球 >	< サッカー ドイツW杯 >	< 小１男児殺害 サッカー >
3	< サッカー サッカー >	< サッカー 藤里 >	< サッカー サッカー 小１男児殺害 >
4	< 全国高校野球 全国高校野球 全国高校野球 >	< 北海道 愛媛 >	< ドイツW杯 サッカー サッカー 小１男児殺害 >
5	< 拉致 拉致 >	< ドイツW杯 * サッカー >	< ドイツW杯 サッカー ドイツW杯 小１男児殺害 >
6	< 全国高校野球 全国高校野球 全国高校野球 全国高校野球 >	< 発注工事談合 知事 >	< サッカー 小１男児殺害 サッカー >
7	< ドイツW杯 サッカー >	< 知事 発注工事談合 >	< サッカー サッカー サッカー 小１男児殺害 >
8	< サッカー ドイツW杯 >	< サッカー 飛鳥会事件 >	< ドイツW杯 サッカー 小１男児殺害 >
9	< 偽装 偽装 >	< 藤里 サッカー >	< ドイツW杯 小１男児殺害 サッカー >
10	< サッカー サッカー サッカー >	< 東京地裁 北海道 >	< サッカー ドイツW杯 小１男児殺害 >

表 2: 先頭コンフィデンスと系列全体コンフィデンス，系列コサインにおける上位 10 位の系列

	H-$conf$	SA-$conf$	S-cos
1	< 救い 救い >	< 救い 救い >	< 救い 救い >
2	< 救い 救い 救い >	< 救い 救い 救い >	< 救い 救い 救い >
3	< 救い 救い どこ >	< 全国高校野球 * 全国高校野球 >	< 全国高校野球 全国高校野球 >
4	< 救い どこ 救い >	< 全国高校野球 全国高校野球 >	< 救い 救い 救い 救い >
5	< 全国高校野球 * 全国高校野球 >	< 全国高校野球 * * 全国高校野球 >	< 都道府県大会 都道府県大会 >
6	< 全国高校野球 全国高校野球 >	< 救い 救い 救い 救い >	< 都市対抗野球 都市対抗野球 >
7	< 全国高校野球 * * 全国高校野球 >	< 都道府県大会 都道府県大会 >	< 全国高校野球 全国高校野球 全国高校野球 >
8	< ドイツW杯 サッカー >	< 全国高校野球 全国高校野球 * 全国高校野球 >	< 高校履修不足 高校履修不足 >
9	< 小１男児殺害 サッカー >	< 都市対抗野球 都市対抗野球 >	< 救い 救い どこ >
10	< 救い 救い 救い 救い >	< 全国高校野球 全国高校野球 全国高校野球 >	< 救い どこ 救い >

3.2 負出尺度

定義 11 S を単一系列データベース，$P = (p_1, \cdots, p_m)$ をパターンとするとき，**負出尺度** $Negative(S,P)$ を以下のように定義する．

$$Negative(S,P) = \max_{p_i \in P} info(S,p_i) - \min_{p_j \in P} info(S,p_j)$$

負出尺度$Negative(S,P)$ は，パターン P 中のアイテムの出現割合の差が大きい有用パターンに重みを置いた尺度である．(一般に負パターンと呼ばれている [Tan 06])

3.3 コンフィデンスと相関性尺度

市川ら [市川 08] が提案した単一長大な系列データから有用系列の抽出を目的とした系列評価尺度をいくつか導入する．頻度計算法が IFMAP と異なるため，IFMAP に対応した評価尺度を示すが，本質的な部分に変わりはない．

定義 12 S を単一系列データベース，P をパターン，p_1 をパターンの先頭要素としたとき，**先頭コンフィデンス** H-$conf(S,P)$，**系列全体コンフィデンス** SA-$conf(S,P)$，**系列コサイン** S-$cos(S,P)$ をそれぞれ以下のように定義する．

$$H\text{-}conf(S,P) = \frac{freq(S,P)}{freq(S,p_1)}$$

$$SA\text{-}conf(S,P) = \frac{freq(S,P)}{\max_{p_i \in P} freq(S,p_i)}$$

$$S\text{-}cos(S,P) = \frac{freq(S,P)}{\sqrt[|P|]{\prod_{p_i \in P}(freq(S,p_i)}}$$

先頭コンフィデンスは順序を考慮したコンフィデンスであり，系列全体コンフィデンスはアイテムの相関性を考慮したコンフィデンスである．また系列コサインは余弦尺度を系列に対応させたものであり，最大の出現頻度を持つアイテムの影響が軽減され，系列全体の相関性を計ることができる尺度である．

4. 実験結果と考察

4.1 実験環境

本論文で提案した補完尺度の有用性を検証するため実験を行った．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，TFIDayF 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．

有用パターン抽出システム IFMAP を拡張実装し，実験を行った．系列の出現幅となるウィンドウ幅は 7 日間，最小サポート値 MS は 2 に固定し，情報量利得基準で上位 1000 位の系列を抽出する．次に，各補完尺度により上位 1000 位の系列を再ランク付けし，各尺度における上位 10 個のパターンの違いについて比較する．

4.2 実験結果

今回はスペースの都合上, 単純頻度と平均情報量の組み合わせ $G_a^s(S,P)$ における結果のみ載せる．表 1 と表 2 が抽出結果及び補完結果である．

情報量利得値 $G_a^s(S,P)$ による抽出結果では同一のアイテムから構成されているパターンが多く存在する．単語異なり尺度 W_{diff} はパターン中のアイテム種類数により評価するため，異なるアイテムからなるパターンが上位に出現する．W_{diff} はローカルフィルタであり，単一系列データベース S 中の情報を利用していない．それに対し，他の補完尺度は S の情報を使うグローバルフィルタである．負出尺度 $Negative$ は情報量の差を利用し，滅多

表 3: 補完尺度の相関性

	W_{diff}	*Negative*	*H-conf*	*SA-conf*	*S-cos*
W_{diff}	1				
Negative	0.09507	1			
H-conf	0.17154	-0.27785	1		
SA-conf	0.26045	-0.58356	0.51294	1	
S-cos	-0.12700	0.18721	0.14930	0.07028	1

に起きないであろうと考えられるパターンに重みが高くつく.「サッカー」と「小 1 男児殺害」に直接的な因果関係はないが，滅多に起きにくい事象が抽出されたという点で，何かしらの有用性があるのではないかと考えられる．一方で，パターン中で最大の情報量を持つアイテムと最小の情報量を持つアイテムが複数のパターンで共通である場合，補完結果が偏ってしまうため，尺度の改善余地がまだあると考えられる．先頭コンフィデンス *H-conf* では，パターンの先頭要素が与える起因率の高さで重みづけしており，一般的なコンフィデンス [Tan 06] と似た意味を持っている．系列全体コンフィデンス *SA-conf* では，パターン中で最大の出現頻度を持つアイテムが与える起因率を見ており，バースト出現をしているパターンを抽出している傾向がある．毎日新聞 2006 年社会面のコーパスでは「救い」という単語が 14 日連続で出現している．また,「全国高校野球」という単語は一定期間に集中して出現することは容易に理解していただけるであろう．系列コサイン *S-cos* は系列全体コンフィデンスと比べ，最大の出現頻度を持つアイテムの影響を軽減しているため，系列全体の相関性を計っている．2006 年に毎日新聞は「死なないで いじめ救いの手はどこに」という特集を組んでおり,「救い」と「どこ」の間に正の相関があったため，< 救い 救い どこ > といったパターンが補完結果として上位に抽出されたと考えられる．

4.3 尺度間の相関性解析

相関分析により各尺度間の相関性の有無を確認した．表 3 は各尺度間のスピアマンの順位相関係数を示したものである．負出尺度と系列全体コンフィデンスの間で負の相関があることが読み取れる．負出尺度はパターン中のアイテムの情報量差が大きいほどスコア値が高く，逆に系列全体コンフィデンスではパターン中のアイテムが同一で，かつバースト出現しているものほどスコア値が高くなる．よって両者に負の相関があると考えられる．また，先頭コンフィデンスと系列全体コンフィデンスの間には正の相関があることが読み取れる．以上から，系列全体コンフィデンスについては，先頭コンフィデンスと負出尺度によってある程度補うことができると考えられる．

新聞記事コーパスからの有用系列抽出を考えた場合，IFMAP を単独利用せず，負出尺度や先頭コンフィデンス，系列コサインのいずれかと併用することは意味があると考えられる．また，単語異なり尺度は単一系列データベース S 中の情報を利用しないため，他の尺度と異質であるが，有用系列抽出においては十分に効果を発揮している．

5. まとめ

本論文では，単一で長大な系列データベースから有用系列パターンを抽出マイニングすることを目的として、情報量基準とそれを補完する新たな系列評価尺度について提案した．新聞記事コーパスを用いた評価実験により，情報量基準に加えて提案補完尺度を併用することの有用性を示した．しかしながら，補完結果は十分と言えず，まだまだ改善していく余地があると考えられる．

我々は 2006 年新聞記事社会面に有用系列が存在するという仮定のもとで研究を行ってきた．どのような有用系列が実際に存在するかの詳細はまだ未知である．ターゲットとする有用系列を明確にすることで，新たな補完尺度を機械学習により考案する手掛かりになる可能性が高いと思われる．

謝辞

本研究は一部，文科省科学研究費補助金（基盤 C：No.22500127）の援助を受けている．

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)

[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)

[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 Single 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)

[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)

[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)

[村田 10] 村田順平, 岩沼宏治, 大塚尚貴: 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. 人工知能学会論文誌 Vol.25, No.3, pp.464–474 (2010)

[Tan 06] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar: Introduction to Data Mining, Section 6.1,7.6 (2006)