知能と複雑系　132－5
(2003. 3. 13)

エピソードルールの近似的冗長性を考慮した効率的な時系列データセットマイニング

藤田　雄介†　　原口　　誠†

† 北海道大学大学院工学研究科電子情報工学専攻　〒060–8628 札幌市北区北13条西8丁目
E-mail: †{fujita,makoto}@db-ei.eng.hokudai.ac.jp

あらまし　時系列データセットからの相関ルールマイニングは，重要なデータマイニングのタスクとして考えられる．一般に，大規模データベースから抽出されるルールは膨大な数に及び，ユーザが全てのルールを検証，評価することは現実的に不可能な作業と言っても過言ではない．本稿では，時系列データセットより生成されるエピソードルールの大部分が冗長であることに着目し，ルールの近似的冗長性の概念に基づき，非冗長なエピソードルールのみを抽出する近似情報基生成手法を導入する．全ての冗長ルールが近似情報基から再構築可能である点で，生成ルール抑制による情報損失は無いと言える．計算機実験の結果により，本手法による抽出したエピソードルールの数が，既存の手法よりも大幅に減少したことを示す．

キーワード　相関ルール発見，時系列データセット，イベント列，エピソードルール，近似的冗長性

An Efficient Mining Method for Episode Rules using Approximate Informative Basis

Yusuke FUJITA† and Makoto HARAGUCHI†

† Division of Electronics and Information Engineering Hokkaido University
N13 W8, Sapporo, Hokkaido, 060–8628 Japan
E-mail: †{fujita,makoto}@db-ei.eng.hokudai.ac.jp

Abstract　Discovery of association rules from time-series datasets is an important data mining task. Generally, the number of potential rules grows rapidly as the size of database increases. It is therefor hard for a user to analyze the rules and realize useful ones among them. To avoid such a difficulty, we make some rules invisible to users, provided they are redundant and approximately reconstructed from another non-redundant ones. In another words, only non-redundant rules are presented to users and will be checked for their interestingness. For this purpose, we first define a notion of *approximate informative basis* consisting of only non-redundant rules, and then present an efficient method to construct it. The degree of approximate reconstruction is associated with the basis as a real parameter adjustable by users. Our experimental results show that the number of non-redundant rules in the approximate informative basis is much reduced.

Key words　association rule mining, time-series dataset, event sequence, episode rule, approximate redundancy

1. はじめに

相関ルールの発見は，データマイニングにおける重要な課題の一つであるが，一般に大規模データベースからは膨大な数の相関ルールが抽出されるため，ユーザがそれら全てを解釈・評価することは，現実的に困難な作業となり，価値あるルールを見落とすことにもなりかねない．

この問題を解消する一つの方法として，冗長性の概念を導入し，非冗長なルールのみを絞り込んで生成する手法が提案されている [1]．ここでの冗長なルールとは，非冗長なルールから簡単な操作により，正確に復元可能であるルールのことを意味する．非冗長なルールのみ対象として出力するため，従来よりも大幅に生成ルール数を減少させることが可能となり，それらを評価するユーザの負担は大幅に軽減される．

しかし神田は，生成ルールの絞り込みがまだ十分になされていないという立場をとり，冗長性を拡張した近似的冗長性を考

慮して，さらなる相関ルールの減少を試みた[2]．この手法では，任意の相関ルールは，非冗長ルールから許容しうる誤差の範囲の頻出度，確信度を持って再現される．この手法により，従来では非冗長であったルールが，近似的に冗長であると見なされるため，さらなる生成ルールの絞込みが実現できる．

本稿では，時系列データセットを対象に，近似的冗長性を考慮した非冗長エピソードルール生成手法の研究と考察を行う．この非冗長なエピソードルール集合を近似情報基と呼び，全ての基準を満たすルールを再現することを保証する．本稿の最後に，実験を行い本手法の有効性を確認する．

2. 準　　備

本研究では，時系列データセットを時間軸に沿ったイベントの列として表す．時間上の制約としてウィンドウを定義し，イベント列上でウィンドウをスライドさせることで，頻出なイベントの組み合わせを計算する[3]．イベントタイプの集合 E が与えられた時の，イベント列 S は，トリプル $(t_B, t_F, \mathcal{S})$ と記述する．ここで，t_B，t_F はそれぞれイベント列の開始時間，終了時間を表す．$\mathcal{S}$ は，生起時間順に従う有限のイベントの列であり，$\mathcal{S} = \{et_1, et_2, \cdots, et_m\}$ と記述する．イベント et_i は，イベントタイプとそのイベントの生起時間の組 (et_i, t_i) で表される．全てのイベントタイプ et_i に対して，$et_i \in E (i = 1, 2, \cdots, m)$ が成り立つ．また，$t_i \leq t_{i+1} (i = 1, 2, \cdots, m-1)$ と $t_B \leq t_i \leq t_F (i = 1, 2, \cdots, m)$ が成り立つ．

イベント列 S 上のウィンドウは，S のイベント部分列として $S_w = (t_b, t_f, w)$ と定義する．ここで，$t_b < t_F$，$t_f > t_B$ である．w は，$t_b \leq t < t_f$ となるイベント (et, t) より構成される．ウィンドウの幅は，$width(S_w) = t_f - b_b$ で定義される．イベント列 S 上で，$width(S_w) = win$ である時の全てのウィンドウ S_w の集合は，$\mathcal{W}(S, win)$ で与えられる．

イベント列上で特定の順序を持つイベントの組み合わせをエピソードと呼ぶ．エピソードがパラレルであるのは，イベント間の順序を考慮しない時である．また，シリアルであるのは，エピソードのイベントが特定の順序を持つ時である．

[定義 1]　順序データマイニングコンテキストは，$\mathcal{D}_S = (\mathcal{W}(S, win), \mathcal{E}, \mathcal{R})$ で与えられる．$\mathcal{W}(S, win)$ は，イベント列 S 上で $width(S_w) = win$ とした時の，全てのウィンドウの集合である．$\mathcal{E}$ は，S 上でのエピソードの集合を表す．$\mathcal{R}$ は，ウィンドウとエピソードの二項関係で，$\mathcal{R} \subseteq \mathcal{W}(S, win) \times \mathcal{E}$ である．

[定義 2]　$\mathcal{D}_S = (\mathcal{W}(S, win), \mathcal{E}, \mathcal{R})$ は，順序データマイニングコンテキストである．$X \subseteq \mathcal{W}(S, win)$，$Y \subseteq \mathcal{E}$ に対して，写像 ϕ, ψ は以下に定義する:

$\phi : 2^{\mathcal{E}} \to 2^{\mathcal{W}(S,win)}$:

$$\phi(Y) = \{w \in \mathcal{W}(S, win) \mid \forall e \in Y, (w, e) \in \mathcal{R}\}$$

$\psi : 2^{\mathcal{W}(S,win)} \to 2^{\mathcal{E}}$:

$$\psi(X) = \{e \in \mathcal{E} \mid \forall w \in X, (w, e) \in \mathcal{R}\}$$

写像 $\phi(Y)$ は，Y の全てのエピソードに関係しているウィンドウの集合を，Y に関連付ける．一方，写像 $\psi(X)$ は，X の全てのウィンドウに共通しているエピソードの集合を，X に関連付ける．

直感的に言えば，$\phi(Y)$ は，Y の全てのエピソードを持つウィンドウの最大の集合である．また $\psi(X)$ は，X の全てのウィンドウで共有しているエピソードの最大集合である．ここで写像 $\sigma : 2^{\mathcal{E}} \to 2^{\mathcal{E}}$ を，$\sigma(Y) = \psi(\phi(Y))$ として定義する．エピソード集合 $Y \subseteq \mathcal{E}$ が，$\sigma(Y) = Y$ を満足する時，Y をエピソードの閉じた集合と呼ぶ．Y に対して $maximal(Y)$ は，集合 Y の極大元集合を求め．その要素を，閉包エピソードと呼ぶ．

次に完全閉包元の定義を行う．Y をエピソードの閉じた集合とし，エピソード f を Y の閉包とする．エピソード g に対して，$f \in maximal(\sigma\{g\})$ が成り立つ時，g は f の完全閉包元である．f の完全閉包元集合 $\mathcal{EG}(f)$ は，$\mathcal{EG}(f) = \{g \in \mathcal{E} \mid f \in maximal(\sigma\{g\})\}$ で定義される，$\mathcal{EG}(f)$ の極小元集合は，$\mathcal{MEG}(f) = \{g \in \mathcal{EG}(f) \mid \nexists g' \in \mathcal{EG}(f), g' \prec g\}$ で与える．

閉包エピソード f とその完全閉包元 $g \in MEG(f)$ によって構成されるタプル (g, f) は，EGC タプルと呼ばれる．任意の EGC タプル (g, f) が与えられた時，$g \preceq e \preceq f$ なるエピソード e に対して，$fr(g) = fr(e) = fr(f)$ となる，コンテキスト $\mathcal{D}_S$ に関する EGC タプル集合は，$EGC(\mathcal{D}_S)$ で与えられる．

イベント列 S，ウィンドウ幅 win，クラス $\mathcal{E}$ が与えられた時，エピソード $e \in \mathcal{E}$ の頻出度は，次式で定義される．

$$fr(e, S, win) = \frac{|\{w \in \mathcal{W}(S, win) \mid e \ occurs \ in \ w\}|}{|\mathcal{W}(S, win)|}$$

頻出度の閾値 $minfr$ が与えられた時，エピソード e が $fr(e, S, win) \geq minfr$ を満たすならば，e は頻出である．

エピソードルールとは，エピソード間の時間的な共起関係を表し，$\alpha \prec \beta$ である二つのエピソードより $r : \alpha \to \beta$ の形で与えられる．エピソードルールは，頻出度および確信度を用いて評価され，それぞれ $fr(r) = fr(\beta), conf(r) = fr(\beta)/fr(\alpha)$ と定義される．有効なエピソードルールとは，ユーザが与える最小頻出度 $minfr$ と最小確信度 $minconf$ を満たすルールのことである．

3. エピソードルールの冗長性

イベント列より，頻出なエピソードを効率的に計算するアルゴリズムに，WINEPI アルゴリズムがある[3]．WINEPI では，最小頻出度，最小確信度を満たす全てのエピソードルールを出力してしまうために，非常に膨大な数のルールを生成してしまう．従って，ユーザにとって出力された全てのルールを検証することは，現実的に困難な問題であると考えられる．この問題を解決するために，生成されたルールの大部分が冗長であることに注目し，エピソードルールの冗長性を定義する．

[定義 3] エピソードルール $r:\alpha \rightarrow \beta$ が，非冗長であるのは，$\alpha' \preceq \alpha, \beta \preceq \beta'$ $fr(r) = fr(r'), conf(r) = conf(r')$ となるような，エピソードルール $r':\alpha' \rightarrow \beta'$ が存在しないことである $(r' \neq r)$.

すなわち，正確に等しい頻出度，確信度を持つエピソードルールの中で，極小の条件部と極大の結論部を持つルールを非冗長であると定義する．この非冗長なエピソードルールは，閉包エピソードとその完全閉包元の極小元集合より構築される．また，冗長性の定義より生成される非冗長のエピソードルール集合を情報基 (IB) と呼ぶ．情報基より全ての有効なエピソードルールは再現され，正確な頻出度と確信度を得ることが保証されている．

本研究でのいくつかの実験結果では，イベント列から生成される非冗長エピソードルール数は，最小頻出度，最小確信度を満たす全てのエピソードルール数よりも大幅に減少されていることが確認された．

4. 近似的冗長性に基く近似情報基生成

生成されるエピソードルールの冗長性に注目することで，さらなるルールの減少が実現できることを前章で述べた．しかし，抽出対象を非冗長なエピソードルールのみに限定したとしても，その数は決して少ないものではなく，それらを検証，評価するには，依然としてかなりの労力が必要とされる．こうした観点より，我々は，抽出対象となるエピソードルールをさらに絞り込む必要があると考える．そこで，本研究では，エピソードルールの冗長性を近似的に捉え直し，近似的に冗長と判断されるルールを抽出対象から除くことで，求めるべきエピソードルールをさらに絞り込むことの実現を試みる．また，こうした近似的冗長性に基く非冗長エピソードルールのもとでも，任意のエピソードルールは簡単な操作で再構築でき，さの頻出度・確信度は，ある一定誤差の範囲内に同定可能であることを示す．

4.1 近似閉包元

近似的冗長性に基く非冗長相関ルールを抽出するために，閉包エピソードに対する完全閉包元の概念を拡張した近似閉包元を新たに導入する．近似閉包元は以下の通り定義される．

[定義 4] e をエピソード，f を閉包エピソードとする．e が f の近似閉包元であるのは，$\exists f' \in maximal(\sigma(\{e\}))$，$f' \preceq f$ かつ $1 \geq fr(f)/fr(f') \geq 1-\varepsilon$ となるとき．ただし，ε はユーザ定義のパラメータであり，$0 \leq \varepsilon < 1$ の範囲の値をとる．

閉包エピソード f の完全閉包元は，f の近似閉包元でもある．上の定義より，以下の性質が導き出される．

[命題 1] g を閉包エピソード f の近似閉包元とする．$g \preceq e \preceq f$ であるエピソード e に対して，以下の性質を有する：

(i) $fr(g) \geq fr(e) \geq (1-\varepsilon) fr(g)$

(ii) $\dfrac{fr(f)}{(1-\varepsilon)} \geq fr(e) \geq fr(f)$

上の命題より，閉包エピソード f とその近似閉包元 g に対して，$g \preceq e \preceq f$ であるエピソード e の頻出度 $fr(e)$ は，$fr(g)$ と $fr(f)$ に対して，それぞれ最大誤差 ε で決定される範囲に収まることが分かる．ここで $\varepsilon = 0$ の時，近似閉包元はすなわち完全閉包元に相当する．

4.2 $EGC(\mathcal{D}_S)$ の近似

任意のエピソード e に対して，e の頻出度は，$g \preceq e \preceq f$ なる EGC タプル (g, f) によって，$fr(g) \preceq fr(e) \preceq fr(f)$ であることより正確に同定できる．

ここで $EGC(\mathcal{D}_S)$ の近似を定義する．この近似の定義により，任意のエピソードの頻出度は，最大誤差 ε で決定される近似的な頻出度として同定することができる．

[定義 5] コンテキスト $\mathcal{D}_S$，最小頻出度 $minfr$ が与えられたとき，頻出閉包エピソード集合を $\mathcal{FC}$ とする．ε はユーザ定義のパラメータである $(0 \leq \varepsilon < 1)$．$\mathcal{FC}$ を以下の条件に従い，複数の集合に分割する $(\mathcal{FC} = \{\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_k\})$ (注1)．各 $\mathcal{P}_i (i = 1, \ldots, k)$ に対して，$\forall f \in \mathcal{P}_i, f \preceq f_i^*$ かつ $fr(f_i^*)/fr(f) \geq 1-\varepsilon$ であるような $f_i^* \in \mathcal{P}_i$ が存在する．さらに $\mathcal{P}_i$ について，$AGC(\mathcal{P}_i) = \{(g, f_i^*) \mid g \in minimal^{(注2)}(\bigcup_{f \in \mathcal{P}_i} \mathcal{MEG}(f))\}$ を得る．以上より，EGC $(\mathcal{D}_S)$ の近似として，$AGC(\mathcal{D}_S, \varepsilon)$ を以下に定義する：

$$AGC(\mathcal{D}_S, \varepsilon) = \bigcup_{i=1}^{k} AGC(\mathcal{P}_i)$$

$AGC(\mathcal{D}_S, \varepsilon)$ の要素を AGC タプルと呼ぶ．また EGC タプルの近似の結果，得られる新たな頻出閉包エピソード集合は，$\mathcal{FC}^*(\varepsilon) = \bigcup_{i=1}^{k} f_i^* \in \mathcal{P}_i$ で与えられ，$\mathcal{FC}^* \subseteq \mathcal{FC}$ であることは明らかである．

4.3 エピソードルールの近似情報基

閉包エピソードとその近似閉包元のタプル集合 $AGC(\mathcal{D}_S, \varepsilon)$ より，我々は，近似情報基と呼ばれる非冗長エピソードルール集合を構築する．近似情報基より，任意の有効なエピソードは容易に再構築可能であり，その頻出度，確信度は近似的に同定される．近似情報基の定義を与える前に，新たに近似エピソードルールの概念を導入する．

[定義 6] コンテキスト $\mathcal{D}_S$，最小頻出度 $minfr$ より，頻出閉包エピソード集合 $\mathcal{FC}$ が与えられる．パラメータ ε より構築される $AGC(\mathcal{D}_S, \varepsilon)$ に基づき，$\mathcal{FC}$ は，$\mathcal{FC} = \{\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_k\}$ に分割される．

$(g, f) \in EGC(\mathcal{D}_S)$ なる EGC タプルに対して，$f \preceq f_i^*$ とな

(注1)：ここで，各 $\mathcal{P}_i$ の集合積が $\mathcal{FC}$ に等しくなることと $(\bigcup_{i=1}^{k} \mathcal{P}_i = \mathcal{FC})$，$\mathcal{FC}$ 内の任意の集合 $\mathcal{P}_i, \mathcal{P}_j (i \neq j)$ は互いに素の関係となる $(\mathcal{P}_i \cap \mathcal{P}_j = \emptyset)$

(注2)：$minimal(\mathcal{P})$ は，エピソード集合である $\mathcal{P}$ に対して，$(\mathcal{P}, \preceq)$ の極小元の集合を求める．

る $\mathcal{P}_i$ を考える．この時，近似エピソードルールとは，$g \to f_i^*$ で定義される．$AGC(\mathcal{P}_i)$ の近似エピソードルールの集合は以下に定義する：

$$AER(\mathcal{P}_i) = \{r : g \to f_i^* \mid (g, f) \in EGC(\mathcal{D}_S), f \preceq f_i^*\}$$

$AGC(\mathcal{D}_S, \varepsilon)$ に関する近似エピソードルール集合は，以下の通り定義される：

$$AER(\mathcal{D}_S, \varepsilon) = \bigcup_{i=1}^{k} AER(\mathcal{P}_i)$$

$r : g \to f^*$ を近似エピソードルールとする．任意のエピソードルール $e_1 \to e_2$ が r から再構築可能であるのは，$g \preceq e_1 \preceq f$ でありかつ，AGC タプル $(g^*, f^*) \in AGC(\mathcal{P})$ に対して，$g^* \preceq e_2 \preceq f^*$ となる時である．

[命題 2] $r : g \to f^*$ を近似エピソードルールとする $((g, f) \in EGC(\mathcal{D}_S), (g^*, f^*) \in AGC(\mathcal{D}_S, \varepsilon))$．$r' : e \to e^*$ は，r から再構築されたエピソードルールである．r' の頻出度，確信度に関して以下の性質を保つ．

$$\text{(i) } fr(r)/(1-\varepsilon) \geq fr(r') \geq fr(r),$$
$$\text{(ii) } conf(r)/(1-\varepsilon) \geq conf(r') \geq conf(r)$$

命題 2 より，r より再構築されたエピソードルール r' の頻出度と確信度は，最大誤差 ε で決定される近似的な頻出度，確信度として得ることができる．

次に最小頻出度，最小確信度を満たす全てのエピソードルールを再構築できる近似情報基の定義を行う．

[定義 7] r をエピソードルールとし，パラメータ ε をユーザより与えられる任意の値とする $(0 \leq \varepsilon < 1)$．もし，$fr(r) \geq minfr, minconf > conf(r) \geq (1-\varepsilon)minconf$ であるとき，ルール r は，近似的に有効である．

[定義 8] $\mathcal{D}_S$ はコンテキスト，ε はユーザ定義のパラメータとする $(0 \leq \varepsilon < 1)$．$\mathcal{D}_S$，ε に関するエピソードルールの近似情報基 AIB は，$(1-\varepsilon)minconf$ を上回る近似エピソード集合として以下の通り定義される：

$$AIB(\mathcal{D}_S, \varepsilon) = \{r \in ASR(\mathcal{D}_S, \varepsilon) \mid conf(r) \geq (1-\varepsilon)minconf\}$$

近似情報基 AIB より再現されるエピソードルールについて，二つの定理が与えられる．

[定理 1]（近似情報基の弱健全性） 近似情報基 AIB より再現されるエピソードルールは有効であるか，あるいは最悪でも近似有効である．

[定理 2]（近似情報基の完全性） 近似情報基 AIB より全ての有効な近似エピソードルールは再現される．

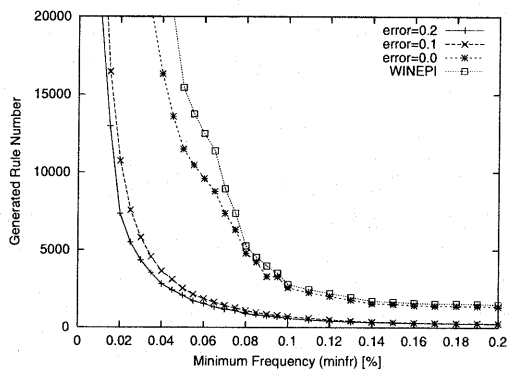

図 1 $minfr$ 値増加による有効エピソードルール数の推移 (パラレル)

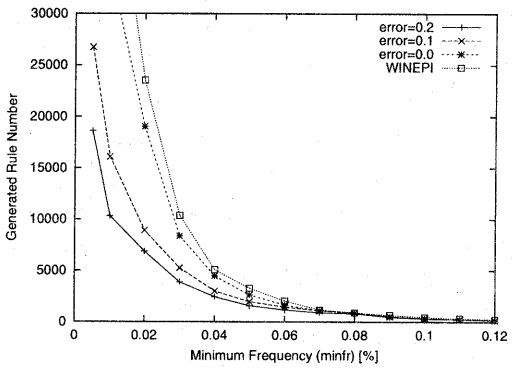

図 2 $minfr$ 値増加による有効エピソードルール数の推移 (シリアル)

この二つの定理によって，エピソードルールの近似情報基は，最小頻出度と最小確信度を満たす全てのエピソードルールを再現することが可能である．すなわち，必要なルールは全て取りこぼすことなくユーザに提示される．また情報基より再現されるルールの確信度は，最悪でも $(1-\varepsilon)minconf$ を満たすことを保証される．

4.4 近似情報基の生成

近似情報基を生成する問題は，大きく三つのサブタスクに分けることができる．

(1) コンテキスト $\mathcal{D}_S$，最小頻出度 $minfr$ より，$EGC(\mathcal{D}_S)$ を計算する．

(2)$EGC(\mathcal{D}_S)$ と，ユーザ定義のパラメータ ε より，$AGC(\mathcal{D}_S, \varepsilon)$ を導出する．

(3)$AGC(\mathcal{D}_S, \varepsilon)$ と最小確信度 $minconf$ より，近似情報基を $AIB(\mathcal{D}_S, \varepsilon)$ を生成する．

5. 実　　験

エピソードルール抽出において近似的冗長性を考慮することの有効性を確認するために，実装実験を行った(注3)．

本研究で用いた実験データは，Internet Traffic Archive のサ

(注3)：システムの実装は C++で行った．

イトに置かれているNASA WWWサーバのHTTPリクエストログを使用した．このデータは，1995年7月1日00:00:00から1995年7月7日23:59:59までの一週間の間に記録されたログデータである．実験では，パラレルとシリアルの二つの場合のエピソードルールの生成数を比較した．パラメータεに与える値は，0.0，0.1，0.2と，WINEPIアルゴリズムの4つのケースを考えた．図1は，パラレルのエピソードルール数の推移をプロットしたグラフである．最小頻出度は，0.005%から0.2%まで，最小確信度は40%に設定した．図2は，シリアルのエピソードルール数の推移をプロットしたグラフである．最小頻出度は，0.005%から0.12%まで，最小確信度は40%に設定した．ここでパラメータεが0.0であるときは，近似的な冗長性を考慮しない場合を示す．

図1より，近似的冗長性を考慮することで，生成されるエピソードルールの効率的な絞込みが行われていることが確認できる．パラメータεが0.0の場合，生成される非冗長なルール数は，WINEPIにより生成されるルール数と比較して，それほど大きな減少は見られないが，εの値を0.1,0.2と増加させることにより，大きく減少していることが分かる．表1より，ε=0.0に対して，εの値が0.1,0.2である場合，生成ルール数は10%から30%程度のルールを削減している．

図2は，シリアルのエピソードルール生成数の推移のグラフである．このグラフよりεの近似によるルール数の減少は，パラレルほど顕著ではないが確認することができる．表2より，最小頻出度が0.04%まで，εが0.0である場合と比較して，$\varepsilon = 0.1, 0.2$に設定しているときは，生成ルール数を30%から60%程度減少させている．これらの結果から，ユーザがルールの取捨選択をする際の負担が大幅に軽減されることがわかる．

近似情報基の近似エピソードルールの具体例として，以下に最小頻出度0.01%，最小確信度40%，$\varepsilon = 0.2$とした時の，出力ルール例を挙げる．

$[conf = 79.5\%]$ $[fr = 0.12\%]$
/history/history.html
/history/apollo/apollo-13/images/ $\Longrightarrow$
　　/history/apollo/apollo.html
　　/history/apollo/apollo-13/apollo-13-info.html
　　/history/apollo/apollo-13/apollo-13.html

$[conf = 87.3\%]$ $[fr = 0.02\%]$
/history/apollo/publications/sp-350/sp-350.txt
/history/apollo/apollo-13/index.html $\Longrightarrow$
　　/history/apollo/new.html
　　/history/apollo/appo-13/apollo-13-info.html

前者は，パラレル，後者はシリアルのエピソードルールである．前者の近似エピソードルールは，冗長とした六つのエピソードルールを包含している．後者も同様に，六つの冗長なエピソードルールを包含している．

6. まとめ

本研究では，神田が提案した近似的冗長性を考慮した相関ルール数抑制手法を，イベント列マイニングを対象に拡張した手法として提案した．また実験の結果より，本手法がルール数削減の観点から有効であることが確認された．しかし，今後の方針として二つの考慮すべき課題が挙げられる．一つ目は，近似的冗長性に基くルール数抑制手法が，真に興味深いルールを獲得するという目的に対して，直接的に有効な手法であるとは必ずしも言い切れない問題がある．その理由は，本手法により絞り込んだルール数が，依然として膨大であり，ユーザがそこから全てを検証，評価するにはまだまだ負担が大きいと考えるためである．そこで，その問題を解消するために，パラメータεを柔軟にカスタマイズさせることで，興味深いルールを同定させるための対話的なシステムを作成する必要がある．

もう一つは，実験データの規模の問題がある．今回の実験では，1つのWWWサーバを対象としてきたが，出力されるルールには筆者の主観的判断にもよるが，それほど興味深いと思われるルールを発見することができなかった．それは，WWWサーバのサイトの規模が比較的小さいため，クライアントのサイト内を巡回するパターンがある程度限られてくることが挙げられる．従って，より規模の大きいWWWサーバや，あるいは複数のWWWサーバの統合したログデータを用いる必要があると考える．また，データの規模が大きくなれば生成されるルールはさらに膨大となり，本研究による有効性がより強く反映することが期待できる．

表1　抽出されるエピソードルール数(パラレル)

$minfr(\%)$	$\varepsilon = 0.2$	$\varepsilon = 0.1$	$\varepsilon = 0.0$
0.005	40,997	63,055	186,071
0.020	7,345	10,738	43,205
0.040	2,805	3,632	16,317
0.060	1,544	1,852	9,607
0.080	890	1,091	4,781
0.100	582	700	2,557

表2　抽出されるエピソードルール数(シリアル)

$minfr(\%)$	$\varepsilon = 0.2$	$\varepsilon = 0.1$	$\varepsilon = 0.0$
0.005	18,626	26,735	45,893
0.010	10,340	16,089	34,004
0.020	6,890	8,928	19,053
0.030	3,884	5,252	8,374
0.040	2,237	2,697	4,469
0.050	1,571	1,961	2,624

文　献

[1] Y. Bastide, N. Pasquier, R. Taouil, G. Stumme, and L. Lakhal. Mining minimal non-redundant association rule using frequent closed itemset. In *CL 2000, LNAI 1861,* pp. 972–986, (2000).

[2] K. Kanda, M. Haraguchi, Y. Okubo. Constructing approximate informative basis of association rules. (2001).

[3] H. Mannila, H. Toivonen, and A. I. Verkamo. Discovery of frequent episodes in event sequences. Technical report, Department of Computer Science, University of Helsinki, Finland, Report C-1997-15. (1997).