プロセスマイニングにおける系列パターン生成と評価指標群に関する考察

Comparing Sequential Pattern Evaluation Indices for Process Mining

阿部秀尚†
Hidenao Abe†
†文教大学　情報学部
† Faculty of Information and Communication, Bunkyo Univ.

要旨
本研究では，情報システムの操作ログデータを対象としたプロセスマイニングにおいて，利用者理解のための特徴的操作系列の抽出を対象に系列パターン生成とその評価指標群の導入について検討を行う．従来，プロセスマイニングは，各操作にあたるイベントの確率的遷移からパターンを見出す方法として提案されてきた．しかし，状態遷移の組み合わせが増大するにつれ，専門家がより特徴的なイベント列のパターンを見出すことが困難となる課題があった．これに対し，系列パターン生成による特徴的な操作系列選定を効率的に行うため，評価指標群による各操作系列データの計量化を行い，分析者が得たい事象と強く関連する特徴的な操作系列を得る必要がある．本稿では，部分操作系列の頻度に基づく系列パターン生成とその評価指標について，共通データにおける系列パターン評価の比較について示す．

1. はじめに

Web 上の各種オンラインシステムをはじめ，情報システムにおけるデータ分析は，ログデータに基づいてソフトウエア単体やシステムの不具合や不正操作を検出する一手法として扱われてきた[1]．しかしながら，多くの情報システムにおけるアクセスやプログラム起動に関するログデータは，一時的に保存することが主な目的であり，問題とする事象が起きなければ一定の期間が過ぎると消去されてしまうことが多い．これは，従来のログデータ分析で提唱されてきたアプローチが設計された範囲内での検証にとどまり，本来，情報システムに関与する利用者や対象タスクに関連した不確実な事象を対象としてこなかったことに起因すると考えられる．

以上のような状況に対し，van der Aalst らは，これまでのログデータ分析にあたる**プロセス検証**だけではなく，外延である**プロセス発見**や可視化による**プロセス理解**を含め，新たなログデータ分析を Process Mining（以下，プロセスマイニング）という枠組みを提唱している[2]．プロセスマイニングでは，ログデータから得られる利用者の操作系列や対象業務のプロセスをベイジアンネットワークとしてイベント間の確率的な遷移モデルとし，可視化インタフェースを通して専門家に提示することで実現されている．しかしながら，ネットワーク上の重要な部分グラフ抽出は専門家に委ねられ，複雑なプロセスを対象とした場合，多くのイベント間での関係が生じることが課題である．

本研究では，従来のプロセスマイニングが確率的なモデルとしてプロセスを提示していたのに対し，“**系列パターン生成手法による部分系列の生成**”と“**分析者が得たい目的事象と特徴的な系列パターンの関係性の抽出**”の２段階でデータマイニング手法を適用する．これは，ログデータに内在するイベント系列パターンを効率的に生成し，各系列パターンの出現について多面的な評価を行い，分析者に了解性の高い分析結果を提示するためである．

本稿では，先行研究で示した系列パターンを利用した特徴的操作系列分析手法[3]に加えて，系列パターンの出現頻度に基づく複数の評価指標を導入し，より分析者に分かりやすい if-then ルールなどのモデルによって目的事象と関連する特徴的な系列パターンとの関連性を指摘する手法の開発に向けた考察を行う．このため，系列パターン出現頻度を基にした評価指標について，代表的な評価指標を示し，共通 Web ページ閲覧ログデータ[4]での系列パターン抽出と系列パターンの並び替えについて述べる．この結果を基に系列パターンの多面的評価を行う指標の有用性について考察する．

2. 用語性判定指標に基づく系列パターン生成と評価指標群の比較

本節では，UCI KDD Archive[5]から提供される Web 閲覧ログデータの共通データセット[4]を用いて，系列パターンの生成と評価指標による整列結果の比較を行う．系列パターンの生成では，自然言語処理での用語抽出のために開発された用語性判定指標[6]を用いて連接した系列パターンを生成する．これらの系列パターンについて，系列データの部分系列である系列パターンの出現回数の計測方法を２種類想定し，代表的な４種類の指標によって得られた結果の並び替えを行った結果を比較する．

2.1 用語性判定指標に基づく系列パターン生成

用語性判定指標は，自然言語によるコーパスから意味のある語句である用語を自動抽出するための指標である．中川らは，１つ以上の連接する名詞からなる用語を抽出するため，語句内の各単語の左右にくる単語の多様性から用語性を判定する手法を提案した[6].

ここで，Web 閲覧ログデータ D 中の各ページを単語 p_i と考え，１つ以上のページアクセスから成る部分系列 $Cs=<p_i>(1 \leqq i \leqq L)$を考えたとき，この用語性判定指標は(1)式のように定義される．(1)式において，$f(Cs)$候補部分系列の出現頻度である．また，Cs の構成要素である p_i を含む bi-gram をそれぞれ見たとき，$FL(p_i)$は p_i の左側（時間順序で言うと直前）に異なるページが来た回数，同様に $FR(p_i)$は p_i の右側（時間順序で言うと直後）に異なるページが来た回数である．

$$FLR(Cs) = f(Cs) \times \{\prod_{i=1}^{L} (FL(p_i)+1)(FR(p_i)+1)\}^{\frac{1}{2L}} \quad (1)$$

本実験で用いた四半期毎の Web 閲覧ログデータの大きさ$|D_{period}|$と(1)式に示す FLR において，FLR>1.0 となる系列パターンを得た結果を表 1 に示す．

表 1 系列データセットのデータ数と FLR スコアに基づく系列パターン数

period	$\|D_{period}\|$	系列パターン数	period	$\|D_{period}\|$	系列パターン数
1996_Q3	755	731	1998_Q1	5296	4396
1996_Q4	1786	1630	1998_Q2	5502	4482
1997_Q1	5005	4277	1998_Q3	4848	4017
1997_Q2	4002	3441	1998_Q4	7956	6360
1997_Q3	3851	3354	1999_Q1	6838	5491
1997_Q4	3534	2827	1999_Q2	1299	1062

なお，この Web 閲覧ログデータでは，米国内の都市の 675 のレストランがリスト中に提示され，選択したレストランを表示して行った 9 種類の行動が１つのイベント p_j として記録されている．このため，イベントの可能な組み合わせは 6705 以上ある．訪問者は目的に合ったレストランを検索するため，訪問から目的のレストランの選択あるいは退去と判断されるまでが任意の長さをもつ１つの閲覧系列データ $s=<p_1, p_2, \ldots, p_m> \in D_{period}$ となっている．

2.2 系列パターン評価指標

表 2 に系列パターンの出現頻度に関して，各評価指標の基準となる頻度の計測方法とそれらに基づく評価指標の定義式を示す．系列パターンを不可分とした場合，出現頻度の計測方法は，２種類考えられる．１つは１つの系列データ中に１つ以上系列パターンが出現しても１つとしてカウントする方法であり，自然言語処理では文頻度(DF)に相当する．もう１つは，系列データ中に出現した系列パターンの出現回数を数え上げる方法であり，語頻度(TF)に相当する．

このように系列パターンの出現頻度が自然言語処理での文頻度，語頻度に対応することから，語の重要度指標としてよく用いられる TFIDF[7]を系列パターンに適用したものが，表 2 中の TFIDF である．

表 2 系列パターン頻度計測基準と系列パターン評価指標の定義式

	系列パターン*sp*の頻度計測基準	
	系列パターンを含む系列データ数 $DF = \mid D_{\in sp} \mid$	系列パターンの出現回数 $TF = \sum_i freq(sp, d_i)$
支持度	$DF / \mid D \mid$	$TF / \sum TF_{sp \in D}$
オッズ	$DF / (DF - \mid D \mid)$	$TF / (TF - \sum TF_{sp \in D})$
自己情報量	$(DF / \mid D \mid) \log_2 (DF / \mid D \mid)$	$(TF / \sum TF_{sp \in D}) \log_2 (TF / \sum TF_{sp \in D})$
TFIDF	$TF * \log(\mid D \mid / DF)$	

2.3 系列パターン評価指標による整列結果の比較

2.1 で生成した系列パターンのうち，1996 年第三四半期について，3.2 で述べた各指標に基づいて並び替えを行った．ここでは，基準となる指標として系列データ数に基づく支持度を利用し，上位 20 位までの結果を表 3 に示す．

表 3 各指標による降順での整列結果（表中の数値は FLR スコアの順位）

降順順位	支持度(TF)	オッズ(TF)	自己情報量(TF)	支持度(DF)	オッズ(DF)	自己情報量(DF)	TFIDF
1	1	1	1	1	1	1	1
2	2	2	2	3	3	3	239
3	3	3	3	4	4	4	688
4	4	4	4	6	6	6	6
5	5	5	5	2	2	2	3
6	6	6	6	5	5	5	4
7	7	7	7	7	7	7	144
8	8	8	8	25	25	25	25
9	9	9	9	8	8	8	2
10	10	10	10	10	10	10	5
11	11	11	11	19	19	19	7
12	12	12	12	144	144	144	19
13	13	13	13	239	239	239	10
14	14	14	14	9	9	9	46
15	15	15	15	11	11	11	8
16	16	16	16	13	13	13	413
17	17	17	17	46	46	46	11
18	18	18	18	12	12	12	13
19	19	19	19	688	688	688	9
20	20	20	20	14	14	14	140

この結果，用語性判定指標である FLR スコア，および系列データ数に基づくオッズや自己情報量は，支持度と同じ結果となった．ただし，これらの指標の間でも算出される値は異なる．一方，系列パターンの出現数に基づく結果では，支持度，オッズ，自己情報量とも同じ並び順となっているが，系列パターンを含むデータ数を基準にした指標とは並び順が異なる．さらに，系列パターンの出現数，系列パターンを含む系列データ数から得た TFIDF では，両者と異なる並び順となる結果が得られた．

図 1 に最も含まれている系列データ数が多かった系列パターンと評価指標によって順位が上昇した系列パターンを示す．

以上の結果，それぞれの指標で値は異なるが，整列結果は基準とする系列パターンの出現頻度の計測方法に拠ることが示された．しかし，ここでは系列パターンの構成要素までは考慮していない．このため，系列パターンの興味度示す指標[8]や時点毎の差分を考慮した指標[9]を用いることで，系列パターン

のさらに多様な側面を計量化し，比較する必要がある．

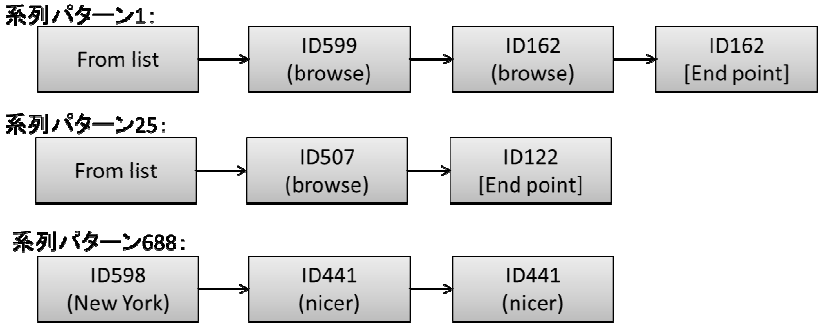

図 1 系列パターンの出現頻度計測基準と評価指標によって順位が変化した系列パターン
（系列パターン番号は FLR スコアによる順位を示し，ID はレストランの ID を示す）

3. まとめ

本稿では，情報システムのログデータにおける特徴的なイベント系列パターンとそれらに関連する特定の事象を分析するため，系列パターンに基づくプロセスマイニング手法について述べ，系列パターンを評価するために用いられてきた指標群による整列結果を比較した．比較では，Web サイトのアクセス履歴である共通ログデータについて得られた系列パターンについて，代表的な 4 種類の評価指標を適用し，系列パターンが各アクセスログデータに複数回出現することを考慮した指標とそうでない指標では，上位にくる系列パターンが異なる結果となった．このことから，指標が利用する系列パターンの頻度計測方法や各指標の値により，系列パターンの異なる側面が計量化されることが示された．

今後は，より多くの指標を実装し，系列パターンの評価値あるいは時間経過に伴う変化の時系列パターンをデータセットとして，利用者の意図や不具合事象と関連する特徴的な系列パターンを抽出する手法の開発を進めていく．さらに，得られたモデルによる提案や不具合回避方法の提示など，情報推薦システムとしての拡張を行っていくことが今後の課題と考えられる．

4. 謝辞

本研究は，日本学術振興会・科学研究費補助金（基盤研究(C) 24500175）の助成を受けたものである．

参考文献

[1] Kobayashi, T. and Hayashi, S.: Recent Researches for Supporting Software Construction and Maintenance with Data Mining, Computer Software, Vol. 27, No. 3, 2010, pp. 3 13–3 23.
[2] van der Aalst, W.: “Process Mining”, Communications of the ACM, vol.55, No. 8, 2012, pp. 76-83.
[3] 阿部秀尚，津本周作：系列パターンマイニングに基づくオーダー入力状況の分析，第 31 回医療情報学連合大会論文集, 2011, pp. 559–562.
[4] Entree Chicago Recommendation Data: http://kdd.ics.uci.edu/databases/entree/entree.html
[5] Hettich, S. and Bay, S. D.: The UCI KDD Archive [http://kdd.ics.uci.edu]. Irvine, CA: University of California, Department of Information and Computer Science.　1999.
[6] Nakagawa, H.: Automatic Term Recognition based on Statistics of Compound Nouns, Terminology, Vol. 6, No. 2, 2000, pp.195–210.
[7] Sparck Jones, K.: A statistical interpretation of term specificity and its application in retrieval. Document retrieval systems, pages 132–142, 1988.
[8] 櫻井茂明：多様なデータに対する系列パターンマイニングの適用, 人工知能学会誌, Vol.27, No.2, 2012, pp.128-135
[9] 岩沼宏治：テキスト系列マイニングにおける有用性尺度について, 人工知能学会誌, Vol.27, No.2, 2012, pp.136-145