T2R2 東京工業大学リサーチリポジトリ
Tokyo Tech Research Repository

論文 / 著書情報
Article / Book Information

論題(和文)	電子カルテシステムのオーダログデータ解析による医療行為の支援
Title(English)	
著者(和文)	佐々木 夢, 荒堀 喜貴, 串間 宗夫, 荒木 賢二, 横田 治夫
Authors(English)	Yume Sasaki, Yoshitaka Arahori, Muneo Kushima, Kenji Araki, Haruo Yokota
出典(和文)	第7回データ工学と情報マネジメントに関するフォーラム論文集, , ,
Citation(English)	, , ,
発行日 / Pub. date	2015, 3

Powered by T2R2 (Tokyo Institute Research Repository)

電子カルテシステムのオーダログデータ解析による医療行為の支援

佐々木　夢† 　荒堀　喜貴†† 　串間　宗夫††† 　荒木　賢二††† 　横田　治夫††

† 東京工業大学工学部情報工学科　〒 152–8552 東京都目黒区大岡山 2-12-1
†† 東京工業大学大学院情報理工学研究科計算工学専攻　〒 152–8552 東京都目黒区大岡山 2-12-1
††† 宮崎大学医学部附属病院医療情報部　〒 889-1601 宮崎県宮崎市清武町木原 5200
E-mail: †sasaki@de.cs.titech.ac.jp

あらまし　近年，医療の現場では電子カルテシステムの利用が広く普及している．電子カルテシステムに蓄積されたログデータを解析することにより，様々な面での医療の支援をすることが可能である．例えば，患者毎のオーダログ中のシーケンスを解析することで医療行為の推薦や検証を行うことができる．そのような医療行為のシーケンス解析においては，医療行為の順番と同時に，行為間の間隔等の情報も重要である．本稿では，基準となる医療行為として手術に着目し，電子カルテシステムのオーダログデータに対して，術後の医療行為間の間隔を考慮したシーケンシャルパターンマイニングを行い，得られた結果を用いて医療行為の支援を行う方法に関して検討する．

キーワード　データマイニング，シーケンシャルパターンマイニング，電子カルテ

1. はじめに

1.1 研究背景

近年，医療の現場では電子カルテシステムが広く普及している．電子カルテシステムの利用は情報の素早い検索・閲覧を可能にし，医療の現場の効率化に貢献している．電子カルテの情報は機械的な処理に通すことが容易であるため，病院に蓄積された情報の活用方法について関心が集まっている．

電子的なクリニカルパスの利用も進んでいる [1]．クリニカルパスとは特定の病気の患者に対して行われる典型的なオーダの時系列のことである．オーダとは医師が患者に実施する処置や検査のことであり，“手術” や “点滴注射” などを指す．クリニカルパスを導入し医療行為の標準化を図ることによって，医療の効率の改善，医療にかかるコストや患者の入院期間を予測しやすくなるなどの効果が期待されている．クリニカルパスを作成する際には，医療関係者が過去のオーダをもとに検討を重ねるなどしているが，事例が膨大であるためクリニカルパスの作成にかかるコストは大きい．

クリニカルパスの作成には計算機による支援が求められおり，これまでにも電子カルテデータを解析しクリニカルパスの作成を支援する研究が行われてきている．

1.2 先行研究

平野らの研究 [2] では，病院情報システムに蓄積されたオーダログを横断的に分析し，全体を通じて典型的と考えられる（適用率の高い）診療プロセスを半自動的に抽出する手法を提案した．平野らの手法では，オーダ列をある特定の期間毎にフェーズに分割し，すべての事例の情報をもとに，各フェーズにおいて各オーダ項目がどの程度の頻度で適用されているか，また，隣接するフェーズ間で，どのオーダ組が連接関係として多用されているかを頻度ベースで集計し，これらをもとに各事例のオーダ列の典型らしさを評価した．さらに，その典型らしさをもとに事例のクラスタリングを行い，クリニカルパスの候補を生成した．平野らの研究でクラスタリングされたものの中には，オーダ列を特定期間で区切ったことにより実際のクリニカルパスと外れてしまうものもあった．この結果の原因として，曜日などの都合によるオーダの実施日時の変動が考えられる．このように医療の現場ではオーダの実施日時の変動があるため，各患者の入院 k 日目に行われたオーダをもとにして頻出なオーダのパターンを抽出した場合，クラスタリングの精度が落ちる可能性がある．

牧原らの研究 [3] では，そうしたオーダの実施日時の変動による影響を回避するため，オーダ列を特定の期間ではなく，特定のオーダを基準に分割し，その前後における頻出なオーダのパターンを抽出することで，クリニカルパスの候補を提示する手法を提案した．頻出するオーダのパターンを抽出する手法として，アプリオリアルゴリズム [9] を元にした抽出アルゴリズムをアクセスログデータに適用した．牧原らは基準となるオーダを設定することで平野らの問題点の解消を図ったが，マイニングの際にオーダの順番のみを考慮したため，抽出したパターンからはオーダの実施日時に関する情報が失われてしまった．また，術後の頻出なパターン数が膨大になってしまうという問題点もあった．

1.3 本研究の目的

本研究ではクリニカルパスの作成を補助する目的で，オーダログデータを解析し，ある疾患に対して行われる頻出な術後のオーダ列を抽出する手法を提案する．前節に述べた先行研究の結果を受けて，頻出なオーダ列の抽出数を抑えながら，オーダの実施日の変動による影響を考慮した抽出を目指す．頻出なオーダ列の抽出数を抑えるために，本研究では飽和オーダ列という概念を導入する．また，オーダの実施日の変動による影響を考慮するために，オーダ間の時間間隔に注目し，同じパターンと見なす時間間隔に幅を持たせたマイニングを行う．そのようにして得られた頻出なオーダ列をクリニカルパスの候補として提示することでクリニカルパス作成のための補助を行う．

実験では宮崎大学医学部附属病院の電子カルテシステムのオーダログデータに本手法を適用し，その有用性を抽出パターン数とオーダ実施日変動の考慮の点から評価する．

1.4 本稿の構成

本稿は以下の通り構成される．2. 章ではシーケンシャルパターンマイニングの基礎とともに，本手法のベースとなったタイムインターバルシーケンシャルパターンマイニングについて説明する．3. 章ではオーダログデータを解析し頻出な術後のオーダ列を抽出する手法について説明する．続く 4. 章では実際の電子カルテシステムに対して手法を適用し，結果に関する考察を行う．最後に 5. 章で本研究のまとめと今後の課題について述べる．

2. シーケンシャルパターンマイニング

シーケンシャルパターンマイニングとは，Agrawal らによって提案されたシーケンシャルデータベースから頻出シーケンスを抽出する手法である [8]．シーケンシャルデータベースはシーケンスとその識別子の組の集合である．シーケンスはアイテムの列からなる．またシーケンスはアイテムと時間の組の列からなることもある．シーケンシャルパターンマイニングの目的は，データベースにある割合以上含まれているすべてのシーケンスを頻出シーケンスとして抽出することである．

本章では，本研究の背景知識としてシーケンシャルパターンマイニングについて説明し，本研究で用いる用語を定義する．

2.1 タイムインターバルシーケンシャルパターンマイニング

Agrawal らが初めに提案したシーケンシャルパターンマイニングではシーケンス内のアイテムの時間の情報を考慮していなかった．そのため，2015 年 1 月 1 日にコーヒーを購入し，その翌日に牛乳を購入したというシーケンスと，2015 年 1 月 1 日にコーヒーを購入し，その 1 年後に牛乳を購入したというシーケンスは区別できなかった．これらのシーケンスを区別するため，Chen らはアイテム間の時間間隔を考慮したタイムインターバルシーケンシャルパターンマイニングを提案した [11]．このマイニング手法は，シーケンシャルデータベースから頻出なタイムインターバルシーケンスを抽出する．タイムインターバルシーケンスは，シーケンス内のアイテムの順番の他に，アイテム間の時間間隔の情報を含んでいる．タイムインターバルシーケンシャルパターンマイニングによって次のようなパターンを抽出することができる．(1) ある客は 1 週間ごとにウェブサイト A を訪れる．(2) 手術 X を行った患者は，2 週間以内にウィルス Y に感染しやすい．

タイムインターバルシーケンシャルパターンマイニングでは，シーケンシャルデータベース，抽出したいタイムインターバルシーケンスの最低支持度，タイムインターバルセットを与え，サポートカウントが最低支持度以上のタイムインターバルシーケンスをタイムインターバルパターンとして抽出する．初めにタイムインターバルセット，シーケンス，タイムインターバルシーケンス，タイムインターバルサブシーケンスを定義する．

タイムインターバルセット． タイムインターバルセット TI は $r-1$ 個の定数 T_k により，下記のように与えられる $r+1$ 個のタイムインターバル I_j を用いて $TI = \{I_0, I_1, I_2, \ldots, I_r\}$ と定義する．

- $I_0 = \{t \mid t = 0\}$
- $I_1 = \{t \mid 0 < t \leqq T_1\}$
- $I_j = \{t \mid T_{j-1} < t \leqq T_j\}$，ただし $1 < j < r-1$
- $I_r = \{t \mid T_{r-1} < t\}$

シーケンス． $A = ((a_1, t_1), (a_2, t_2), \ldots, (a_n, t_n))$ とするとき，A はシーケンスである．ただし a_j はアイテム，t_j は a_j が発生した時間を表し，$2 \leqq j \leqq n$ において $t_{j-1} \leqq t_j$ である．同じ時間に発生したアイテムがある場合，それらは辞書順で並ぶものとする．

タイムインターバルシーケンス． $I = \{i_1, i_2, \ldots, i_m\}$ をアイテム集合，$TI = \{I_0, I_1, I_2, \ldots, I_r\}$ をタイムインターバルセットとする．$B = (b_1, \&_1, b_2, \&_2, \ldots, b_{s-1}, \&_{s-1}, b_s)$ とするとき，B をタイムインターバルシーケンスと定義する．ただし $b_i \in I$，$\&_i \in TI$ とする．

タイムインターバルサブシーケンス． シーケンス $A = ((a_1, t_1), (a_2, t_2), \ldots, (a_n, t_n))$ とタイムインターバルシーケンス $B = (b_1, \&_1, b_2, \&_2, \ldots, b_{s-1}, \&_{s-1}, b_s)$ について，以下の 2 つの条件を満たすような整数列 $\{j_k\}$ が存在するとき，B は A に含まれる，もしくは B は A のタイムインターバルサブシーケンスであるという．

（1） $b_1 = a_{j_1}, b_2 = a_{j_2}, \ldots, b_r = a_{j_r}$

（2） $t_{j_i} - t_{j_{i-1}} \in \&_{i-1}$

（3） $1 \leqq j_1 < j_2 < \ldots < j_s \leqq n$

トランザクションはその識別子 sid とシーケンス s を組にした (sid, s) によって表現される．シーケンシャルデータベース S はトランザクションの集合で構成される．データベース S 中における，タイムインターバルシーケンス α のサポートカウントは次のように定義される．

サポートカウント． $support_count_S(\alpha) = |\{(sid, s) | (sid, s) \in S \wedge \alpha$が s に含まれる $\}|$

タイムインターバルシーケンス α のサポートカウントが最低支持度の割合以上であるとき，α はタイムインターバルシーケンシャルパターンである．

2.2 I-プレフィックススパン

Chen らはタイムインターバルシーケンシャルパターンを抽出するために，I-プレフィックススパンというアルゴリズムを提案した．I-プレフィックススパンは，シーケンシャルパターンを効率よくマイニングする手法のプレフィックススパン [10] を，タイムインターバルを考慮するように拡張したものである．ここでは I-プレフィックススパンに関する主要な定義について説明した後，例とともにアルゴリズムを説明する．

タイムインターバルプレフィックス． シーケンス $\alpha = ((a_1,t_1),(a_2,t_2),\ldots,(a_n,t_n))$，タイムインターバルシーケンス $\beta = (b_1,\&_1,b_2,\&_2,\ldots,\&_{m-1},b_m)$ $(m \leqq n)$ が以下の条件を満たすとき，β は α のタイムインターバルプレフィックスである．

（1） $1 \leqq i \leqq m$ において $b_i = a_i$

（2） $1 < i \leqq m-1$ において $t_i - t_{i-1} \in \&_{i-1}$

射影シーケンス． シーケンス $\alpha = ((a_1,t_1),(a_2,t_2),\ldots,(a_n,t_n))$ と α のタイムインターバルサブシーケンスであるようなタイムインターバルシーケンス $\beta = (b_1,\&_1,b_2,\&_2,\ldots,\&_{s-1},b_s)$ $(s \leqq n)$ を考える．また $a_{i_k} = b_k (1 \leqq k \leqq s)$ とする．α のサブシーケンス $\alpha' = ((a'_1,t'_1),(a'_2,t'_2),\ldots,(a'_p,t'_p))(p = s+n-i_s)$ が以下の条件を満たすとき，α' は α の β に関する射影シーケンスである．

（1） β は α' のタイムインターバルプレフィックス．

（2） α' の後方 $n-i_s$ 個のアイテムと α の後方 $n-i_s$ 個のアイテムが等しい．

タイムインターバルポストフィックス． $\alpha' = ((a'_1,t'_1),(a'_2,t'_2),\ldots,(a'_p,t'_p))$ を α の $\beta = (b_1,\&_1,b_2,\&_2,\ldots,\&_{s-1},b_s)$ に関する射影シーケンスとするとき，$\gamma = ((a'_{s+1},t'_{s+1}),(a'_{s+2},t'_{s+2}),\ldots,(a'_p,t'_p))$ はプレフィックス β に関する α のポストフィックスである．

最後に，シーケンシャルデータベース S のシーケンスを α で射影したそれぞれのポストフィックスを集めたものを射影データベース $S|_\alpha$ とする．

I-プレフィックススパンのもとになったプレフィックススパンは次のアルゴリズムで頻出パターンを抽出する．初めに $\alpha = null$ に設定し，$S|_\alpha$ から頻出な長さ 1 のパターンを抽出する．$S|_\alpha$ のそれぞれのパターンを α に結合し α' とし，$S|_{\alpha'}$ を構成する．その後 $S|_{\alpha'}$ から頻出なパターンを抽出し，α' に結合して再び射影データベースを構成することを繰り返し，S 中のすべてのシーケンシャルパターンを抽出する．

I-プレフィックススパンではタイムインターバルを考慮するために，射影データベースを構成する際に表 *Table* を用いる．その表は列がアイテムに，行がタイムインターバルに対応している．表中のそれぞれのセル $Table(I_i,b)$ は，b を含み，b と α の最後のアイテムとの時間間隔が I_i を満たす $S|_\alpha$ 内のトランザクション数を示す．$Table(I_i,b)$ の値が与えられた最低支持度を超えるとき，(I_i,b) を α に結合した α' をタイムインターバルシーケンシャルパターンとして得る．さらに $S|_{\alpha'}$ を構成し，繰り返すことで S 中のすべてのタイムインターバルシーケンシャルパターンを抽出する．

例として $I_0 : t = 0, I_1 : 0 < t \leqq 3, I_2 : 3 < t \leqq 6, I_3 : 6 < t$ とし，タイムインターバルセットを $TI = \{I_0, I_1, I_2, I_3\}$，最低支持度を 50%として，表 1 に示すシーケンシャルデータベースについて考える．

はじめに，データベースを 1 度読み取り，あるアイテムがいくつのトランザクションに含まれているかを数え，長さ 1 の頻出パターンを得る．これによりアイテム数 1 の頻出パターンとして，$(a),(b),(c),(d),(e)$ の 5 パターンを得る．

表 1　シーケンシャルデータベース

sid	sequence
10	$((a,1),(c,3),(a,4),(b,4),(a,6),(e,6),(c,10))$
20	$((d,5),(a,7),(b,7),(e,7),(d,9),(e,9),(c,14),(d,14)$
30	$((a,8),(b,8),(e,11),(d,13),(b,16),(c,16),(c,20))$
40	$((b,15),(f,17),(e,18),(b,22),(c,22))$

次に，得られたそれぞれのアイテムによってデータベースを射影する．射影する際には，その時間の情報を保持する．例えば，パターン (a) によってデータベースを射影した場合，表 2 に示す射影データベースが得られる．これをもとにパターン (a) をプレフィックスとするアイテム数 2 の頻出パターンを抽出していく．

表 2　射影データベース $S|_{(a)}$

sid : time	projected sequence
10 : 1	$((c,3),(a,4),(b,4),(a,6),(e,6),(c,10))$
10 : 4	$((b,4),(a,6),(e,6),(c,10))$
10 : 6	$((e,6),(c,10))$
20 : 7	$((b,7),(e,7),(d,9),(e,9),(c,14),(d,14))$
30 : 8	$((b,8),(e,11),(d,13),(b,16),(c,16),(c,20))$

アイテム数が 2 以上のパターンの場合，アイテムが一致するほかにタイムインターバルの条件も満足する必要がある．そこで I-プレフィックススパンでは，タイムインターバルとアイテムをそれぞれ行と列にもつテーブルを作成する．射影された各トランザクションの時間差を計算し，その時間差が含まれるセルの値に 1 を加える．値が最低支持度以上になったセルのタイムインターバルとアイテムをパターンに加えることで，アイテム数が 1 多いパターンを得ることができる．表 2 からは表 3 に示す *Table* が得られる．*Table* の結果から $(I_0,b),(I_0,e),(I_1,e),(I_3,c)$ が頻出なセルと分かる．プレフィックス (a) にこれらを加えることで，$(a,I_0,b),(a,I_0,e),(a,I_1,e),(a,I_3,c)$ の 4 つのパターンが得られる．

表 3　$S|_{(a)}$ における *Table*

Table	a	b	c	d	e
I_0	0	3	0	0	2
I_1	1	1	1	1	3
I_2	1	0	1	1	1
I_3	0	1	3	1	0

続いてパターン (a,I_0,b) によって射影すると，表 4 に示す射影データベースが得られる．

先ほどと同様に *Table* を構成すると表 5 が得られる．(I_1,e) と (I_3,c) が頻出なセルとなるので $(a,I_0,b,I_1,e),(a,I_0,b,I_3,c)$ の 2 つのパターンが得られる．このように探索を続け，タイムインターバルシーケンシャルパターンのマイニングを行う．

すべての枝で，射影されるシーケンスの数が 0 になったとき

表 4　射影データベース $S|_{(a,I_0,b)}$

sid : time	projected sequence
10 : 4	$(a,6),(e,6),(c,10))$
20 : 7	$(e,7),(d,9),(e,9),(c,14),(d,14))$
30 : 8	$(e,11),(d,13),(b,16),(c,16),(c,20))$

表 5　$S|_{(a,I_0,b)}$ における *Table*

Table	b	c	e
I_0	0	0	1
I_1	0	0	3
I_2	0	1	0
I_3	1	2	0

I-プレフィックススパンの探索は終了となる．

3. 提案手法

2. 章では一般的なシーケンシャルデータベースから頻出シーケンスをマイニングする手法について説明した．本章では，術後の頻出なオーダ列を抽出するために，シーケンシャルパターンマイニングの技術を電子カルテシステムへ適用する手法について説明する．また本章では，抽出パターン数を効果的に削減するために飽和オーダ列の導入を行う．

3.1　頻出な術後のオーダ列の抽出

本研究では，オーダの実施日の変動による影響を考慮した頻出な術後のオーダ列の抽出を行うために，オーダ間の時間間隔に注目し，同じパターンと見なす時間間隔に幅を持たせたマイニングを行う．そのようなマイニングを行うために，2.1 節で紹介したタイムインターバルシーケンシャルパターンマイニングを本手法のベースにした．タイムインターバルシーケンシャルパターンマイニングの用語は，本手法において表 6 のように対応している．

本手法の概要を図 1 に示す．本手法は電子カルテシステムのログデータから必要な情報を取り出してデータベースを構築する段階と，データベースをマイニングして頻出な術後のオーダ列を抽出する段階の 2 つの段階に分かれている．以下ではそれぞれの段階について説明する．

初めにマイニングの対象とするシーケンシャルデータベースの構築について説明する．ここでは以下の 2 点の性質をもつシーケンシャルデータベースを構築する．(1) 各トランザクションはある疾患を患った患者に対して施されたオーダ列を入院期間毎に切り分けられている．(2) 各シーケンスはオーダの中に “手術” を表すオーダを含む．2 つ目の性質は，本手法の目的が頻出な術後のオーダ列の抽出のため，必要である．具体的なデータベース構築方法について説明する．まず解析の対象とする疾患をキーとして，ログデータからその疾患を患った患者の患者 ID をすべて取得する．次に，取得した患者 ID を持つ患者へ実施したオーダとそのオーダの実施日をオーダログデータから取得する．得られたデータをもとに患者 ID 毎にシーケンスを構成する．シーケンスのオーダ内に “手術” を表すオーダが含まれていない場合，そのシーケンスを破棄する．そうでない場合，そのシーケンスと患者 ID を組としてトランザクションとする．そのようにして得たトランザクションを集める，目的のデータベースを構築する．

表 6　語の対応関係

2.1 節手法	本手法
シーケンス	入院期間中のオーダ列
アイテム	オーダ
アイテム発生時間	オーダ実施日
タイムインターバルシーケンシャルパターン	頻出なオーダ列
sid	患者 ID

データベースを構築した後は，構築したデータベースから頻出な術後のオーダ列を抽出する．抽出手法にはタイムインターバルシーケンシャルパターンマイニングを用いる．頻出な術後のオーダ列を抽出するために，I-プレフィックススパンを適用する際，最初は “手術” を表すオーダのみで射影を行う．2 回目以降の射影は通常の I-プレフィックススパンと同様に，射影データベース内で頻出なオーダすべてで行う．初めの射影を “手術” を表すオーダのみで行うことにより，マイニングの結果得られるパターンは必ず “手術” を表すオーダから始まるようになる (図 2)．以上のようにして，ある疾患の頻出な術後のオーダ列をオーダの実施日の変動による影響を考慮して抽出する．なお，医療行為は 1 日を単位として患者に実施することが多いため，本手法ではタイムインターバルの単位を 1 日とする．

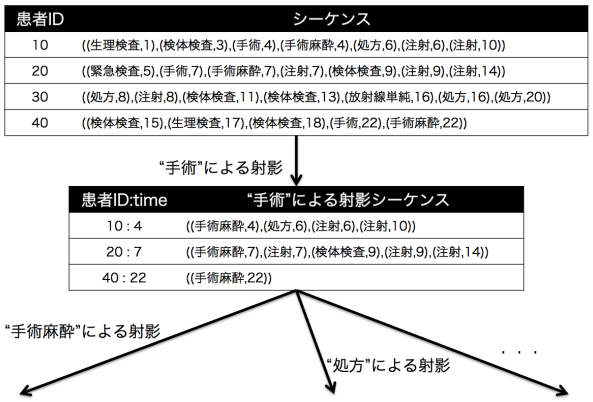

図 2　本手法のプレフィックススパンの流れ

3.2　飽和オーダ列の抽出

前節で説明した頻出なオーダ列の抽出では，図 3 に示すパターン B のように別の頻出なオーダ列に含まれているオーダ列も頻出なオーダ列として抽出するため，膨大な数のオーダ列が抽出されてしまう．本節ではそのような冗長なオーダ列を削除する手法について説明する．

シーケンシャルパターンマイニングの分野には飽和シーケンスというものがある．飽和シーケンスとは，抽出されるシーケンスの中でサポート値が等しく，かつ他のシーケンスのサブシーケンスとならないシーケンスのことである．3.1 節の手法で抽出した頻出なオーダ列はタイムインターバルシーケンスによって表現される．飽和オーダ列を抽出するには飽和タイム

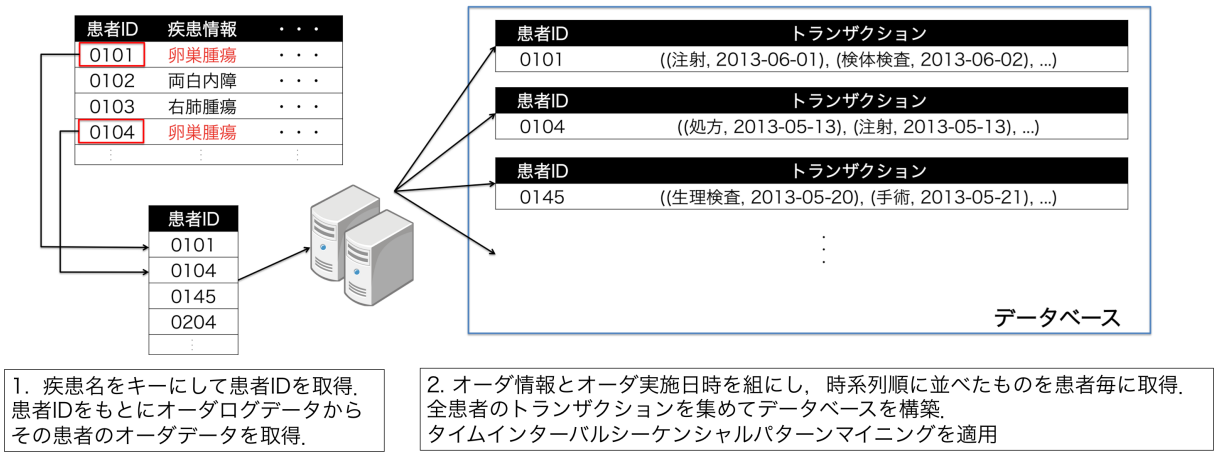

図 1　本手法の概要

図 3　抽出オーダ例

インターバルシーケンスを定義する必要がある．ここでは飽和タイムインターバルシーケンスを抽出するために，タイムインターバルシーケンスに包含関係を定義する．

定義． 2 つのタイムインターバルシーケンス $A = (a_1, \&_1, a_2, \&_2, \ldots, \&_{i-1}, a_i), B = (b_1, \&'_1, b_2, \&'_2, \ldots, \&'_{j-1}, b_j)$ について以下の条件を満たすとき B は A に含まれるという．$1 \leqq k \leqq i-1, \&_k \neq I_0$ を満たす k を昇順に並べた数列を $\{k_p\}_{p=1}^{m}$，$1 \leqq l \leqq j-1, \&'_l \neq I_0$ を満たす l を昇順に並べた数列を $\{l_q\}_{q=1}^{n}$ とし，これらを用いてタイムインターバルシーケンス A，B をそれぞれ次のように A'，B' に変形する．$A' = ((a_1, a_2, \ldots, a_{k_1}), \&_{k_1}, (a_{k_1+1}, a_{k_1+2}, \ldots, a_{k_2}), \&_{k_2}, \ldots, \&_{k_m}, (a_{k_m+1}, a_{k_m+2}, \ldots, a_i))$ $B' = ((b_1, b_2, \ldots, b_{l_1}), \&'_{l_1}, (b_{l_1+1}, b_{l_1+2}, \ldots, b_{l_2}), \&'_{l_2}, \ldots, \&'_{l_n}, (b_{l_n+1}, b_{l_n+2}, \ldots, b_j))$ A'，B' のなかでアイテムの集合を表している部分をそれぞれ a'_p, b'_q で置き換える．すなわち $A' = (a'_1, \&_{k_1}, a'_2, \&_{k_2}, \ldots, \&_{k_m}, a'_{m+1}), B' = (b'_1, \&'_{l_1}, b'_2, \&'_{l_2}, \ldots, \&'_{l_n}, b'_{n+1})$ となる．$1 \leqq x \leqq n+1, 1 \leqq y \leqq n$ において $a'_x \supseteqq b'_x$，$\&_y = \&'_y$ を満たすときタイムインターバルシーケンス B はタイムインターバルシーケンス A に含まれる．

以上の定義に基づくと，図 3 に示したパターン A はパターン B を含んでいる．そのため，ここから飽和オーダ列を抽出するとパターン A のみが得られることになる．

本手法では飽和オーダ列を以下の手順で抽出する．まず 3.1 節で説明した手法を用いて頻出なオーダ列を抽出し，パターン集合 P を作成する．P からサポート値が等しい 2 つのパターンを取り出し，その 2 つのパターンの包含関係を確認する．2 つのパターンが包含関係にあるとき，他方のパターンに含まれているパターンを P から取り除く．これを全組み合わせについて行った後，P に残ったパターンは飽和オーダ列となる．

4. 実　　験

これまで，電子カルテシステムにシーケンシャルパターンマイニングを適用し頻出な術後のオーダ列を抽出する手法について述べてきた．本章では，実際の電子カルテシステムのオーダログデータに対して手法を適用し，本手法の有用性を次の 2 点に注目して評価する．(1) 飽和オーダ列の抽出によってどれだけパターン数を絞り込めたか．(2) オーダ実施日の変動を考慮した抽出を行えたか．

4.1 実験対象データ

本研究の実験には宮崎大学医学部附属病院の電子カルテシステムに蓄積された 2012 年 1 月 1 日から 2013 年 12 月 31 日までのオーダログデータを使用する．このオーダログデータは，宮崎大学医学部附属病院で使われている電子カルテシステム WATATUMI [5], [6], [7] によって取得した．このオーダログデータは個人情報保護の点から患者の名前を含んでいない．ある患者に対して行われたオーダを抽出する際には，匿名化された患者 ID を用いて抽出を行った．なお，当該オーダログデータを医療行為改善の研究に用いることは，宮崎大学病院の HP [4] にてオプトアウトしており，宮崎大学の倫理審査委員会および東京工業大学の疫学研究等倫理審査委員会の承認を得ている．

オーダログデータの 1 レコードには患者 ID，オーダ種別，オーダ内容，オーダ実施日などの情報が含まれている (表 7)．オーダ種別には実施したオーダの内容の情報が “生理検査” や “手術” のように記録されている．オーダ内容にはオーダの手技や，処方した薬剤とその量など，オーダ種別より細かい情報が記録されている．オーダ種別は 58 種類から選択し記録されているデータであり，オーダ内容は手技の後に部位や薬剤，用量などが記録されているテキストデータである．

オーダ情報としてオーダ種別を用いる場合，クラスタリングの粒度が大きいため適切なパターンを抽出できない可能性が考えられる．一方でオーダ内容をオーダ情報として用いる場合，微妙な表記の差によって異なるオーダと判断されてしまうた

表 7　オーダログデータの例

匿名化患者 ID	オーダ種別	オーダ実施日	...	オーダ内容
16ed67321d1	手術	2012-09-01	...	子宮附属器腫瘍摘出術（両）（開腹）＜卵巣両側部分切除＞　腟式卵巣嚢腫内容排除術　大塚生食注　１Ｌ（1瓶）　イソジン液１０％　２５０ｍｌ／本（150ｍＬ）
2743e7204e1	細菌検査	2012-10-13	...	細菌培養同定検査，細菌薬剤感受性検査，塗抹顕微鏡検査（一般細菌）
3f3e1889f82	放射線単純	2012-04-23	...	単純撮影（デジタル撮影）単純撮影（イ）の写真診断 電子画像管理加算（単純撮影）単純-胸腹部　胸部　呼気のみ　Ａ→Ｐ　坐位　ダミーフィルム（1 枚）
...	...	...	...	...

表 8　実験で使用したトランザクション数

	卵巣腫瘍	右白内障	心不全
トランザクション数	29	52	47

表 9　タイムインターバルセット

タイムインターバルセット TI1	0 日, 1 日後, 2 日後, 3 日後, 4 日後, 5 日後, 6 日以降
タイムインターバルセット TI2	0 日, 1 日以降 2 日以内, 3 日後, 4 日後, 5 日後, 6 日以降
タイムインターバルセット TI3	0 日, 1 日以降 3 日以内, 4 日後, 5 日後, 6 日以降

め，必要以上にクラスタリングされてしまう可能性がある．そこで本研究では，オーダ内容の先頭部分にオーダの手技が記録されることが多いことに着目し，オーダ内容を空白で区切ったときの先頭部分を短縮オーダと名付け，オーダの分類に使用した．例えば表 7 の匿名化患者 ID“16ed67321d1” の短縮オーダは “子宮附属器腫瘍摘出術（両）（開腹）＜卵巣両側部分切除＞” となる．

表 8 に実験に使用した疾患名とトランザクション数を，表 9 にタイムインターバルセットの詳細を示す．タイムインターバルセットは TI1 が時間に関する制約が最も厳しく，TI2，TI3 の順に制約が緩くなるように設定した．

4.2　実験内容

本節では，飽和オーダ列の抽出による抽出オーダ数の削減の効果と，オーダ実施日の変動を考慮した抽出であるかを確認するための実験内容について説明する．

飽和オーダ列の抽出による抽出オーダ数の削減の効果を評価するため，実際の電子カルテシステムのオーダログデータを用いた 2 種類の抽出実験を行った．1 つは 3.1 節の手法を適用した頻出なオーダ列の抽出実験，もう 1 つは 3.2 節の手法を適用した飽和オーダ列の抽出実験である．飽和オーダ列を抽出する場合としない場合を比較し，飽和オーダ列を抽出した場合に抽出数をどれだけ削減できたか評価する．実験は卵巣腫瘍・右白内障・心不全の 3 種類の疾患に対して最低支持度とタイムインターバルセットの設定を変化させて行った．

オーダ実施日の変動の考慮については，飽和オーダ列の抽出実験で得られたパターンの内容を異なるタイムインターバルセット間で比較することにより確認する．

4.3　飽和オーダ列抽出による削減効果

3.1，3.2 節の手法により得られた頻出オーダ列と飽和オーダ列の抽出数を図 4 から図 6 に示す．それぞれのグラフ内の同じ色の点線と実線の差が冗長なパターンの数を表す．これら 3 つの疾患について，最低支持度の低下に伴い冗長なパターンが増加していくことが確認できる．

表 10 に，各疾患の最低支持度 85%における頻出オーダ列数から飽和オーダ列数への削減率を示す．

表 10　飽和シーケンスによる抽出数の削減率

タイムインターバルセット	卵巣腫瘍	右白内障	心不全
TI1	89.5%	98.9%	35.1%
TI2	89.7%	99.5%	38.9%
TI3	90.0%	99.7%	33.3%

表 10 から，3 つの疾患すべてについて抽出パターン数を削減できていることが分かる．飽和オーダ列の抽出はパターン数の削減に有効であることが分かった．特に右白内障での抽出パターン数の削減効果が顕著であった．一方，心不全に対しては抽出パターン数の削減効果が低い結果となった．

飽和オーダ列抽出の削減効果の違いについて考察する．図 7 に最低支持度 85%の条件における疾患毎の頻出オーダ列の平均アイテム数を示す．図 7 によると右白内障の頻出オーダ列の平均アイテム数が他の疾患と比べ多いことが分かる．その後に卵巣腫瘍が続き，最も平均アイテム数が少ないのは心不全となっている．この順番は飽和オーダ列抽出によるパターン数の削減効果が大きかった疾患の順である．アイテム数の多いオーダ列ほど包含関係にあるオーダ列が多くなる．そのため飽和オーダ列抽出によるパターン数削減効果が高くなると考えられる．

4.4　オーダ実施日の変動の考慮

同一パターンと見なす時間の幅を広く取ることによりオーダ実施日の変動を考慮できた例として図 8 と図 9 を示す．これらの図は最低支持度 85%とし，図 8 はタイムインターバルセットを TI2，図 9 はタイムインターバルセットを TI3 として抽出した心不全の飽和オーダ列をチャートのような形で表現し，“手

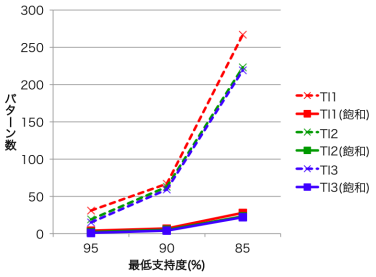

図 4　抽出数（卵巣腫瘍）

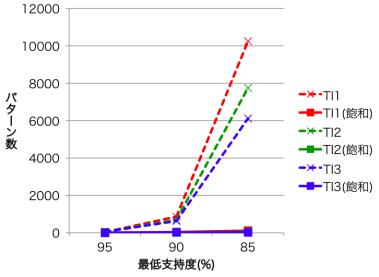

図 5　抽出数（右白内障）

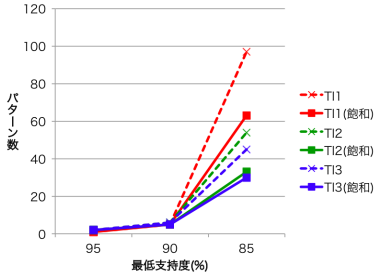

図 6　抽出数（心不全）

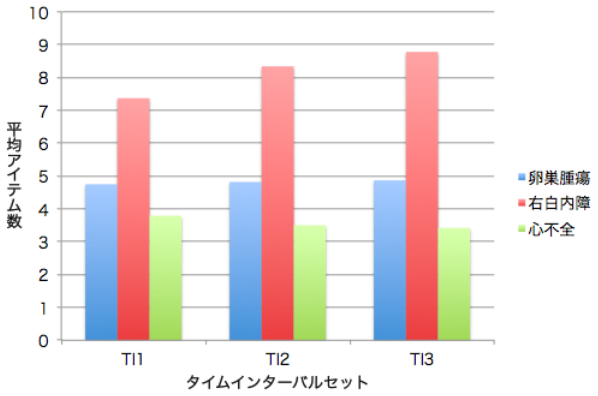

図 7　最低支持度 85%における頻出オーダ列の平均アイテム数

術” のオーダの後に続く飽和オーダ列の一部を拡大したものである．ノードはオーダを表し，ノードとノードは矢印によって接続されている．矢印はノード間のタイムインターバルとサポート値の情報を持つ．例えばノード A が “1~3 86” の矢印でノード B へ接続されている場合，オーダ A が行われてから 1 日以降 3 日以内にオーダ B が行われるパターンはデータベース中の 86%のトランザクションにサポートされていることを示す．また，ノード C が “6~ 94” の矢印でノード C へ接続されている場合，オーダ C が行われてから 6 日以降にノード C が行われるパターンはデータベース中の 94%のトランザクションにサポートされていることを示す．同一のオレンジのブロック内にあるノードは同日に行われることを示している．

図 8 と図 9 の “中心静脈注射” から次のブロックへの遷移を比較すると，図 8 には図 9 で青で囲った “内服薬剤” にあたる遷移が存在しない．同一パターンと見なす時間の幅を広く取ることによりオーダ実施日の変動を考慮できた一例であると言える．

4.5　タイムインターバルセット

本節ではタイムインターバルセットの設定が抽出できるパターンに与える影響について議論する．

頻出オーダ列の抽出実験の結果によると，どの疾患も抽出されたパターン数は TI1 が最も多く，TI3 が最も少なかった．これは疾患のオーダログデータに，同じ内容のオーダが何日にも渡って連続する特徴があることが原因と考えられる．実際に得られたパターンを観察すると，卵巣腫瘍では “点滴注射” のオーダが，右白内障では “眼圧測定・精密” と “細隙燈顕微鏡検査（前眼部及び後眼部）” のオーダが，心不全では “中心静脈注射” のオーダが抽出されるパターンの後半で連続して現れる．このとき例えば卵巣腫瘍において，TI3 で < 点滴注射 , 1 日以降 3 日以内，点滴注射 > と抽出されるパターンは TI1 では，< 点滴注射, 1 日後, 点滴注射 >, < 点滴注射, 2 日後, 点滴注射 >, < 点滴注射, 3 日後, 点滴注射 > のように時間的に細かい分類で抽出されやすくなる．そのためどの疾患においても TI1 での抽出数が他のタイムインターバルセットと比べて多くなったと考えられる．

タイムインターバルセットの制約が厳しいほど得られるパターンの時間間隔の情報が詳細になるが，オーダの実施日の変動による影響を受けやすくなる．タイムインターバルセットの選択は，抽出シーケンスの種類とオーダ間の時間の正確さに影響を与える．パターン抽出の際には適切なタイムインターバルを設定することが重要である．

5.　まとめと今後の課題

本研究ではクリニカルパスの作成を補助する目的で，電子カルテシステムのオーダログデータを解析し，ある疾患に対して行われる頻出な術後のオーダ列を抽出する手法を提案した．本手法ではオーダの実施日の変動による影響を考慮した抽出を目指すため，オーダ間の時間間隔に注目し，同じパターンと見なす時間間隔に幅を持たせたマイニングを行った．また，オーダ列の抽出数を効果的に削減するために飽和オーダ列を導入した．

実際の電子カルテシステムのオーダログデータに手法を適用し，飽和オーダ列の抽出による抽出オーダ数の削減の効果の評価と，オーダ実施日の変動を考慮した抽出ができているかを確認する実験を行った．

実験の結果，すべての疾患について飽和オーダ列の抽出による抽出数の削減は有効であることが分かった．また，同一の疾患の異なるタイムインターバルセットの抽出パターンを比較した結果，制約の緩いタイムインターバルセットによる抽出パターンの中に，制約の厳しいタイムインターバルセットによる抽出では得ることのできなかったパターンを抽出できている例があることを確認できた．

今後の課題として，今回の研究で抽出したパターンの医学的な有用性に関する評価が必要である．評価の方法としては，クリニカルパスが十分に確立された疾患を用いた実験や，医療関係者からの反応をもとにした評価などが考えられる．また，今回の研究では 3 つの疾患へ手法を適用したが，他の疾患につ

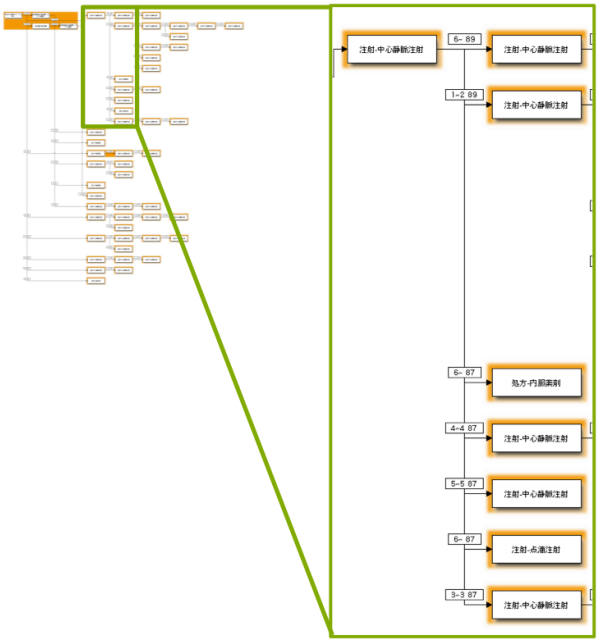

図 8　飽和オーダ列グラフ：心不全（85%，TI2）

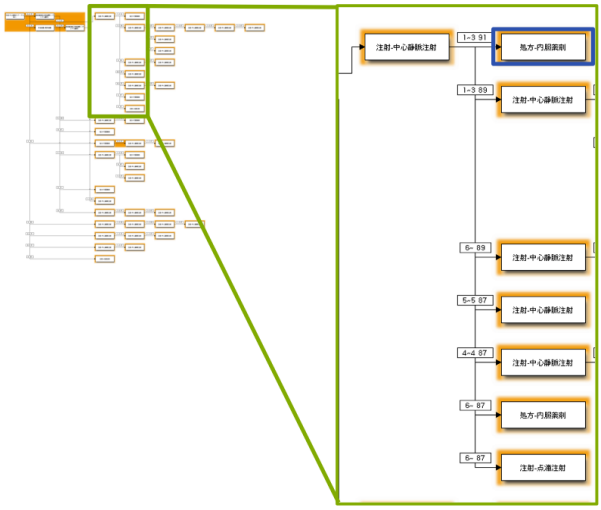

図 9　飽和オーダ列グラフ：心不全（85%，TI3）

いても手法を適用し，検証を重ね抽出手法の改善を目指していきたい．最後に，今回の研究では 3 種類のタイムインターバルセットを用いて進めたが，最適なタイムインターバルの決め方については議論できていない．適切なタイムインターバルを決定する手法についての検討も行っていきたい．

謝　　辞

本研究の一部は，日本学術振興会科学研究費補助金基盤研究（A）（#25240014）の助成により行われた．

また，本研究で宮崎大学医学部附属病院の電子カルテオーダログを医療行為改善の研究に用いることは，宮崎大学の HP にてオプトアウトしており，宮崎大学の倫理審査委員会および東京工業大学の疫学研究等倫理審査委員会の承認を得ている．関係各位の協力に感謝する．

文　　献

[1] Osamu Okada, Naoki Ohboshi, Tomohiro Kuroda, Keisuke Nagase, and Hiroyuki Yoshihara. Electronic Clinical Path System Based on Semistructured Data Model Using Personal Digital Assisant for Onsite Access. Journal of Medical Systems, Vol. 29, No. 4, pp. 379-389, 2005

[2] Shoji Hirano and Shusaku Tsumoto. Clustering of Order Sequences Based on the Typicalness index for Finding Clinical Pathway Candidates. ICDM Workshops, 2013.

[3] 牧原健太郎，荒堀喜貴，渡辺陽介，串間宗夫，荒木賢二，横田治夫．電子カルテシステムの操作ログデータの時系列分析による頻出シーケンスの抽出．DEIM Forum 2014, F6-2, 2014.

[4] 宮崎大学医学部附属病院医療情報部. http://www.med.miyazaki-u.ac.jp/home/jyoho/

[5] 電子カルテシステム WATATUMI. http://www.corecreate.com/02_01_izanami.html

[6] 串間宗夫，荒木賢二，鈴木斎王，荒木早苗，田村宏樹，淡野公一，外山貴子，石塚興彦，池田満．マイニング技法を活用した電子カルテ (IZANAMI) のネットワーク可視化．電子情報通信学会技術研究報告．CAS，回路とシステム 108(338), pp. 187-192, 2009.

[7] 串間宗夫，荒木賢二，鈴木斎王，荒木早苗，仁鎌照絵．電子カルテネットワークシステム (IZANAMI) のテキストデータ分析．電子情報通信学会技術研究報告．IN, 情報ネットワーク 109(449), pp. 383-388, 2010.

[8] Rakesh Agrawal and Ramakrishnan Srikant. Mining Sequential Patterns. Proceedings of 1995 International Conference on Data Engineering, pp. 3-14, 1995.

[9] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. Proceeding of the 20th Intervational Conference on Very Large Data Bases, pp. 487-499, 1994.

[10] Jian Pei, Jiawei Han, Behzad Mortazavi-Asl, Helen Pinto, Qiming Chen, Umeshwar Dayal and Mei-Chun Hsu. PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth. Proceeding of 2001 International Conference on Data Engineering, pp. 215-224, 2001.

[11] Yen-Liang Chen, Mei-Ching Chiang, Ming-Tat Ko. Discovering time-interval sequential patterns in sequence databases. Expert Systems with Applications 25, pp. 343-354, 2003.