

インタラクティブなシーケンシャルパターンマイニングの高速化の評価

青柳 結衣¹ Le Hieu Hanh¹ 松尾 亮輔¹ 山崎 友義¹ 荒木 賢二¹ 横田 治夫² 小口 正人¹

概要：ビッグデータの活用が進む中、頻出パターンを抽出するシーケンシャルパターンマイニング (SPM) が注目されている。最適な閾値はデータセットに依存するため、閾値を調整しながら解析を繰り返すインタラクティブな SPM が不可欠である。しかし、従来の手法は既知の頻出パターンの利用により実現しているが、他パターンに含まれないクローズド頻出パターンが十分に考慮されておらず、分析効率には課題が残る。医療カルテのようなデータセットでは、クローズド頻出パターンだけを確認することで、治療方針全体を把握できる。本稿では、クローズド頻出パターンに着目し、インタラクティブな SPM の高速化を実現する手法を提案する。また、パターンの長さに着目し、抽出するパターン数を減少させる手法も提案する。さらに、公開データセットを使用して評価を行うことで、クローズドとパターンの長さを考慮することの有意性を示す。

Evaluating Methods for Improving the Efficiency of Interactive Sequential Pattern Mining

YUI AOYAGI¹ HIEU HANH LE¹ RYOSUKE MATSUO¹ TOMOYOSHI YAMAZAKI¹ KENJI ARAKI¹
HARUO YOKOTA² MASATO OGUCHI¹

1. はじめに

1.1 研究背景

様々なビッグデータの蓄積に伴い、シーケンスを対象とする SPM(Sequential pattern mining) が注目されている。SPM とは、閾値 (minsup) を指定し、頻度の高いシーケンシャルパターンを抽出する解析手法である。購買行動分析 [12]、医療カルテ解析 [16]、ウェブページのクリックストリーム分析 [5] などのアイテムの発生順序が重要となるデータベースに対して、頻出アイテムセット抽出手法 [1] より有益な情報が得ることができる。SPM のアルゴリズムは盛んに提案されており、有名なアルゴリズムとして、ID リストを用いた SPADE[15]、ID のリストとビットマップを組み合わせた SPAM[3]、プレフィックスとポストフィックスを用いた PrefixSpan[11] などが挙げられる。

SPM に使われる minsup は、データセットのシーケンス数を 1 とした時の割合で表現する。シーケンスがデータセット内に出現する頻度をサポート値とし、サポート値が minsup より大きい場合に頻度が高いとみなされる。最適な minsup はデータセットに依存するため、指定する minsup が小さすぎると頻出シーケンシャルパターン数が増え、minsup が大きすぎるとパターンが出力されない。そのため、minsup を調整しながら解析を行うのに長けているインタラクティブな SPM が不可欠である。インタラクティブな SPM の有名なアルゴリズムとして、GSP[13] や KISP[7] が挙げられる。

1.2 本研究の目的

従来の手法である KISP は、既知の頻出パターンを知識ベース (KB) に保存し、それを適宜参照することでインタラクティブなシーケンシャルパターンマイニング (SPM) を実現している。しかし、大量に生成される頻出シーケンシャルパターンの中には、意味の重複した冗長なパターンが多数含まれており、パターンの探索やユーザの分析作業

¹ お茶の水女子大学
Ochanomizu University

² 城西大学
Josai University

において非効率が生じる。本研究では、この冗長性の問題を解消し、インタラクティブな SPM の高速化と有用なパターンの提示を両立することを目的とする。具体的には、頻出クローズドシーケンシャルパターンのみを抽出することで、サポート値が同じ冗長なサブシーケンスを排除し、候補パターン数を削減する手法を提案する。クローズドシーケンシャルパターンとは、同一のサポート値をもつスーパーシーケンスが存在しないパターンであり、情報損失を抑えながら出力を圧縮できる特徴をもつ。さらに、パターンの有用性と可読性を高めるために、パターン長に基づくフィルタリングを導入し、より意味のある長いパターンを優先的に抽出する処理も組み込んだ。提案手法の有効性を検証するため、実データを用いて従来手法と比較し、クローズド考慮およびパターン長考慮が実行時間に与える影響を評価した。

1.3 本稿の構成

本稿は以下の通りに構成される。2 節では本研究の関連研究について説明する。3 節では提案手法であるクローズドとパターン長を考慮するインタラクティブな SPM について述べる。4 節では、2 つの公開データセットを用いて提案手法の有効性を評価する実験を行い、それらの結果を述べる。最後に 5 節で本稿のまとめと今後の課題について述べる。

2. 関連研究

本節では、本研究に関連する SPM, PrefixSpan, 頻出クローズドシーケンシャルパターン, インタラクティブな SPM と KISP について説明する。

2.1 SPM

Agrawal らによって提案されたシーケンシャルパターンマイニング (SPM)[2] は、シーケンスデータベース (SDB) から頻出シーケンシャルパターンをすべて抽出するための手法である。SDB はシーケンスとシーケンス ID の組の集合で表される。SPM では入力された minsup よりも頻度が高いパターンを頻出パターンとする。minsup が 0.1 の場合、SDB 内の 10% 以上のシーケンスに含まれているパターンが出力される。minsup を小さくすると多くのパターンが出力されるが、有益な情報が埋もれてしまうことがある。逆に minsup を大きくすると頻出パターンが出力されないため、適切な minsup を指定する必要がある。

2.2 PrefixSpan

Jian Pei らが提案した PrefixSpan は、深さ優先探索で頻出シーケンシャルパターンを求める SPM である [11]。入力は SDB と minsup、出力はサポート値が minsup 以上である頻出シーケンシャルパターンとなっている。

ID	シーケンス
1	(検査A,検査B), (麻酔), (手術), (投薬D,看護)
2	(検査A), (麻酔), (手術), (投薬D,看護)
3	(投薬E), (麻酔), (手術), (投薬D)
4	(検査C), (麻酔), (手術), (投薬F), (看護)
5	(検査A,検査C), (投薬E), (麻酔), (手術), (投薬D,看護)
6	(検査A,検査B,検査C), (麻酔), (手術), (投薬F,看護)

図 1 医療データセット例

prefix : (検査A)	prefix : (検査C)	prefix : (麻酔)	prefix : (手術)
(検査B), (麻酔), (手術), (投薬D,看護)		(手術), (投薬D,看護)	(投薬D,看護)
(麻酔), (手術), (投薬D,看護)		(手術), (投薬D)	(投薬D)
	(麻酔), (手術), (投薬F), (看護)	(手術), (投薬F), (看護)	(投薬F), (看護)
(検査C), (投薬E), (麻酔), (手術), (投薬D,看護)	(投薬E), (麻酔), (手術), (投薬D,看護)	(手術), (投薬D,看護)	(投薬D,看護)
(検査B,検査C), (麻酔), (手術), (投薬F,看護)	(麻酔), (手術), (投薬F,看護)	(手術), (投薬F,看護)	(投薬F,看護)

図 2 射影 DB の一部

アルゴリズムの説明の前に、プレフィックスを定義する。シーケンス $\beta = \langle b_1, b_2, \dots, b_j, \dots, b_m \rangle$ が $\alpha = \langle a_1, a_2, \dots, a_j, \dots, a_n \rangle$ のプレフィックスであるとは、 $m \leq n$ かつ $\beta_j = a_j (1 \leq i \leq m-1), b_m \sqsubseteq a_m$ が成り立つことである。さらに、 $a'_m = a_m - b_m$ とした時、シーケンス $\gamma = \langle a'_m, a_{m+1}, \dots, a_n \rangle$ をポストフィックスという。SDB 中の全シーケンスを対象として、プレフィックス β に対して求めたポストフィックスの集合を射影データベース $SDB|_\beta$ という。図 1 の例において、minsup が 0.5 の時、長さが 1 のプレフィックスに対する射影データベースは図 2 で示した。なお、指定されたシーケンスに存在しない postfix については、該当箇所を空白で示している。

PrefixSpan では射影データベースを用いて頻出シーケンスを抽出する。以下がアルゴリズムの詳細である。

- (1) SDB 内で minsup の条件を満たす長さ 1 の頻出シーケンシャルパターンを $FSP^{l=1}$ とする。
- (2) $FSP^{l=1}$ の各要素をプレフィックス β とし、それに対するポストフィックスの集合 γ からなる射影データベース $SDB|_\beta$ を生成する。
- (3) プレフィックス β に射影データベース $SDB|_\beta$ 内の対応するポストフィックス集合 γ 中の minsup の条件を満たす要素を結合して β' とする。
- (4) ポストフィックス集合 γ 中で結合した要素をプレフィックスとしたときのポストフィックス集合 γ' からなる射影データベース $SDB|_{\beta'}$ を生成する処理を β' が空になるまで繰り返す。

これによって、全ての候補の組み合わせの生成とそれらの頻度を算出する必要がなくなる。また、項目の順位を決めておき、同時に出現する項目は順位に従った順番の並びとして処理するため、組み合わせ数が減少する。

サポート値を算出するためにシーケンスの位置情報リストを求める。図 1 において、(検査 A) の位置情報はシーケンス 1 と 2 と 5 と 6 の 0 番目となる。よって、位置情報リストは (1,0),(2,0),(5,0),(6,0) となり、このリストの要素数がそのままサポート値となる。

頻出シーケンシャルパターンを探す際、minsup の条件を満たさないシーケンスの位置情報リストも求めている。図 1 の例で minsup=0.5 の場合を考える。(検査 A) を含む長さ 2 の頻出シーケンシャルパターンは、((検査 A),(麻酔)), ((検査 A),(手術)), ((検査 A),(投薬 D)), ((検査 A),(看護)), ((検査 A),(投薬 D, 看護)) の 5 パターンとなる。これらを調べる際、minsup を満たさないシーケンス ((検査 A),(検査 B)), ((検査 A),(検査 C)) の位置情報も確認している。

2.3 頻出クローズドシーケンシャルパターン

大量に生成される頻出シーケンシャルパターンには、解析の面から見た場合に冗長なものが多数存在している。そのため、SPM では頻出クローズドシーケンシャルパターンを抽出することが一般的となっている。次の性質を満たすシーケンシャルパターン (FCSP) を頻出クローズドシーケンシャルパターンとする [14]。

$$FCSP = \{\alpha \mid \alpha \in FSP, \nexists \beta \in FSP, \alpha \subset \beta, Sup(\alpha) = Sup(\beta)\}$$

これは、頻出クローズドシーケンシャルパターンは同じサポート値であるサブシーケンスを持たないことを示している。よって、短いシーケンシャルパターンの多くが頻出クローズドシーケンシャルパターンではないため、除去する対象となる。

図 1 の例では、minsup=0.5 で抽出した頻出シーケンシャルパターンの中で、 $\alpha=((検査 C), (麻酔), (手術))$ はサポート値が 3 である。また、 $\beta=((検査 C), (麻酔), (手術), (看護))$ も頻出シーケンシャルパターンとなり、サポート値は 3 である。この場合、 β は頻出クローズドシーケンシャルパターンだが、 α は異なるため、取り除かれる。 α の情報は全て β に含まれているため、結果に含める必要がない。

2.4 インタラクティブな SPM

従来の SPM は、データベースが静的であると仮定している。実際、頻出シーケンシャルパターンを取得するためにデータベースに一回のみ適用するよう設計されている [6]。その後、データベースが更新されると、アルゴリズムを再び最初から実行する必要がある。データベースの変更が小規模であり、再び完全に探索する必要がない場合があるため、この手法は非効率である。

この問題を解決するために、いくつかの増分的な SPM アルゴリズムが設計されている [4], [8], [9]。増分的なアルゴリズムには、ユーザが minsup などのパラメータを変更



図 3 KISP の概要

することを考慮して、インタラクティブにマイニングするよう設計されたものもある。インタラクティブな SPM の一つである KISP は、GSP アルゴリズムを拡張したものである。

2.5 KISP

KISP は、知識ベース KB を使用するインタラクティブな SPM で、minsup を変更し繰り返し実行するのに適している。KB は今まで調べた頻出パターンとそのサポート値と最小の minsup である KB.base で構成されている。指定された minsup が KB.base より大きい場合は、minsup を満たすパターンを KB から単純に取得するため、パフォーマンスが大幅に向上する。

図 3 に KISP アルゴリズムの大まかな流れを示す。まず、指定された minsup と KB.base の比較を行う。次に新しい候補シーケンスを生成し、そのサポート値を算出し、結果を KB に蓄積する。全ての候補シーケンスについて算出後、minsup の条件を満たす頻出シーケンシャルパターンを取得できる。

2.5.1 指定された minsup と KB.base の比較

minsup が KB.base 以上である場合は、頻出シーケンシャルパターンを KB から取得する。minsup を満たす頻出パターンは全て KB に保存されているため、データセットにアクセスせず、頻出パターンを得ることができる。minsup が KB.base より小さい場合は新たにパターンを探索する。

2.5.2 候補シーケンスの生成とサポート値の算出

探索をする際、最初に新しい候補シーケンスを生成する。ここでは今までのマイニングで候補シーケンスとされなかったシーケンスのみを生成する。KISP は既存のインタラクティブな SPM より候補シーケンス数が少ないため、高速である。最後に候補シーケンスのサポート値を算出し、KB に保存する。

2.5.3 課題点

実際に出力される頻出シーケンシャルパターンの中には、意味の重複する冗長なパターンが多数存在し、知識ベースの肥大化や探索効率の低下を招くという課題がある。本研究では、この課題を解決するために、頻出クローズドシーケンシャルパターンのみを抽出対象とすることで、冗長なパターンの除去と探索空間の圧縮を図る。クローズドシーケンシャルパターンは、同じサポート値を持つスーパーシーケンスが存在しないため、情報を損なうことなく出力パターン数を削減できる特徴をもつ。さらに、情報量を維

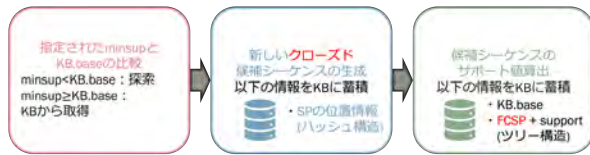


図 4 提案手法の概要

持しつつ、ユーザにとってより有用なパターンを提示するために、パターン長に着目した抽出を行う。長いパターンはしばしばより具体的かつ意味のある傾向を表すことが多く、短い部分パターンに比べて有用性が高い場合がある。そのため、パターン長を考慮したフィルタリングにより、重要なパターンの優先的提示を可能とし、探索の効率化と結果の可読性向上の両立を目指す。

3. 提案手法

図 4 に本研究の提案手法の構成を示した。提案する手法では、KISP の既存頻出パターンを利用する KB(Knowledge Base) の仕組みを応用する。主な提案は 2 つある。

3.1 候補シーケンスの生成

一つ目はクローズドなシーケンスのみを候補シーケンスとして生成することである。それによって、頻出クローズドシーケンシャルパターンのみを抽出できる。また、候補シーケンス数が減少することによってマイニングが速くなると推測される。

図 5 に候補シーケンスの生成手順を示した。まず、候補シーケンスとその位置情報を与えると、minsup とクローズドの条件を満たしているか検証する。満たしている場合は頻出シーケンスとみなし、KB に保存する。次に、新しい候補シーケンスの位置情報を取得する。KB に位置情報が保存されている場合は KB から取得し、保存されていない場合は位置情報を調べ、KB に保存する。位置情報の調べ方は PrefixSpan と同様、射影データベースを求めてそこから算出する。最後に候補シーケンスの枝刈りをし、minsup とクローズドの条件を満たさないシーケンスは除外する。これを繰り返すことで、全ての頻出パターンを求めることができる。

3.2 パターン長に着目した抽出

二つ目の提案は、出力される頻出パターンに対してパターン長に基づいたフィルタリングを行うことである。PrefixSpan を用いて抽出された頻出シーケンシャルパターンは、部分パターンを多数含んでおり、それらが KB に蓄積されることで、パターンの冗長性が高まる可能性がある。本研究では、得られた頻出パターンの中から他のより長いパターンに含まれるものを除外し、意味のある最長パターンのみを残す処理を行う。これにより、情報の重複を防ぎつつ、ユーザにとってより有用なパターン提示を可能

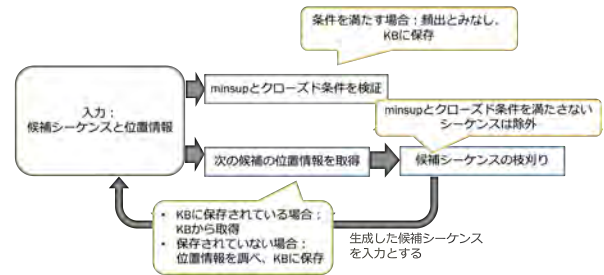


図 5 候補シーケンスの生成

にする。具体的には、パターンを長さ順に並べた上で、各パターンが他のパターンの部分シーケンスであるかどうかを判定し、そうである場合は除外するアルゴリズムを適用する。この処理を通して、マキシマルなシーケンスのみを保持し、知識ベース内の情報圧縮と可読性の向上を図る。

4. 実験

本実験では提案手法においてクローズドとパターン長考慮における有効性を確認することを目的とする。

4.1 実験環境

表 1 に実験環境を示した。提案手法の実装には、PrefixSpan の実装コードを参考にした。使用したデータセット [10] の詳細な内容は、表 2 に示した。

4.2 実験内容

提案手法を実装し、2 種類の実験を行った。

まず、クローズドのみを候補シーケンスとすることによる有意性を調べるため、クローズドを考慮した場合としない場合の実行時間を測定し比較した。BMSWebView1 と BMSWebView2 を使用し、minsup を 0.00065 から 0.00070 になるまで 0.00001 刻みで実行した。また、候補シーケンス数の変化についても調査した。

次に、パターン長によるフィルタリングをすることによる優位性を調べるために、除去を行う場合と行わない場合の抽出される頻出シーケンシャルパターン数を測定し比較した。BMSWebView1 と BMSWebView2 を使用し、

表 1 実験環境

サーバ	Dell PowerEdge R740xd
CPU	Intel Xeon Gold 5218 16 cores x 2
OS	Ubuntu 24.04.1 LTS
メモリ	64GB x 6
python	3.12.3

表 2 データセットの概要

	BMSWebView1	BMSWebView2
シーケンス数	59,601	77,512
平均要素数	2.42	4.62
サイズ (MB)	1.5	3.6
内容	EC サイトのクリックストリームデータ	

	クローズド考慮あり(s)	クローズド考慮なし(s)	考慮あり/考慮なし
BMSWebView1	392.39	816.75	48%
BMSWebView2	323.33	325.61	99%

図 6 クローズド考慮の実行時間

	クローズド考慮あり	クローズド考慮なし	考慮あり/考慮なし
BMSWebView1	85,186,590	195,240,402	44%
BMSWebView2	53,330,915	67,693,468	79%

図 7 候補シーケンス数の比較

実行時間	クローズド考慮のみ(s)	クローズド+パターン長考慮(s)	パターン長考慮あり/なし
BMSWebView1	392.39	788.62	201%
BMSWebView2	323.33	502.07	155%

図 8 頻出シーケンシャルパターン数の比較

シーケンス数	クローズド考慮のみ	クローズド+パターン長考慮	パターン長考慮あり/なし
BMSWebView1	194,657	62,112	32%
BMSWebView2	303,368	39,404	13%

図 9 パターン長考慮の実行時間

minsup を 0.00065 から 0.00070 になるまで 0.00001 刻みで実行した。

4.3 実験結果

4.3.1 クローズド考慮による実行時間の検証

結果を図 6 に示した。どちらのデータセットもクローズドありの方が高速である。これは、クローズドを考慮することにより、候補シーケンス数が減少したため速くなったと考えられる。BMSWebView1 ではクローズドの有無による実行時間の差が大きい、BMSWebView2 では差が僅かである。

続いてそれぞれの候補シーケンス生成数の比較を図 7 に示した。先ほどの結果も用いると、どちらのデータセットもクローズド考慮ありの方が高速であるのは、クローズドを考慮することによって候補シーケンス数が減少したためだとわかる。また、データセットによりクローズドの有無による実行時間の差が異なるのは、候補シーケンス数の減少量と相関があることがわかる。よって、クローズドを考慮すると候補シーケンス数が減少するため、高速化に繋がることが明らかになった。

4.3.2 パターン長考慮による検証

結果を図 8 に示す。どちらのデータセットにおいても、パターン長を考慮することで出力される頻出シーケンシャルパターン数が大幅に減少し、情報の冗長性が大きく抑制されたことが確認できる。具体的には、パターン数はおよそ BMSWebView1 では約 70%、BMSWebView2 では約 90%の削減となった。一方で、実行時間に関しては図 9 の結果より、BMSWebView1 では約 200%、BMSWebView2 では約 155%の増加が見られた。これは、各パターンに対して部分シーケンスの除去を行う処理が追加されたことに起因していると考えられる。

この結果から、出力パターン数の削減によってユーザの理解を支援できる一方、処理時間の増加というトレードオフが存在することが明らかになった。しかし、インタラクティブなマイニングにおいて重要となる「意味のあるパターンの提示」という観点では、本手法が有効であると評価できる。

5. おわりに

5.1 まとめ

SPM は適切な minsup を指定する必要がある。そのため、インタラクティブな SPM を用いることで、適切な minsup を見つけやすくなる。本研究では、インタラクティブなシーケンシャルパターンマイニングにおける処理効率と情報の有用性の両立を目的として、頻出クローズドシーケンシャルパターンに着目した手法を提案した。提案手法では、既知の頻出パターンを活用する KISP の知識ベースを応用し、候補シーケンスとしてクローズドなシーケンスのみを生成することで、候補数を削減し、マイニング処理の高速化を図った。さらに、出力されるパターンの中から部分シーケンスを取り除き、最長なもののみを残すことにより、情報量を保ちながら冗長性を抑える工夫を行った。これにより、ユーザにとってより意味のあるパターンが得られやすくなり、インタラクティブなパターン探索において有用な知見を効率的に提供できることが期待される。

5.2 今後の課題

今後の課題として、KB で頻出シーケンシャルパターンを管理する構造についての検証を行う。さらに、KISP と比較することによって、提案手法の有効性の確認を行う。

謝辞

本研究の一部は日本学術振興会科学研究費 (#24K02943) の助成からの支援によって行われた。

参考文献

- [1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB'94)*, pp. 487–499, 1994.
- [2] R. Agrawal and R. Srikant. Mining sequential patterns. In *Proceedings of the Eleventh International Conference on Data Engineering*, pp. 3–14, 1995.
- [3] J. Ayres, J. Gehrke, , and T. Yiu J. Flannick. Sequential pattern mining using a bitmap representation. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*

- (KDD'02), pp. 429–435, 2002.
- [4] H. Cheng, X. Yan, and J. Han. IncSpan: incremental mining of sequential patterns in large database. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 527–532, 2004.
 - [5] P. Fournier-Viger, T. Gueniche, and V. S. Tseng. Using partially-ordered sequential rules to generate more accurate sequence prediction. In *International Conference on Advanced Data Mining and Applications*, pp. 431–442, 2012.
 - [6] Philippe Fournier-Viger, Jerry Chun-Wei Lin, Rage-Uday Kiran, Yun-Sing Koh, and Rincy Thomas. A survey of sequential pattern mining. *Data Science and Pattern Recognition*, Vol. 1, No. 1, pp. 54–77, 2017.
 - [7] M. Y. Lin and S. Y. Lee. Improving the efficiency of interactive sequential pattern mining by incremental pattern discovery. In *International Conference on System Sciences*, pp. 68–76, 2002.
 - [8] F. Masseglia, P. Poncelet, and M. Teisseire. Incremental mining of sequential patterns in large databases. *Data and Knowledge Engineering*, Vol. 46, pp. 97–121, 2003.
 - [9] S. N. Nguyen, X. Sun, and M. E. Orlowska. Improvements of incspan: Incremental mining of sequential patterns in large database. In *Advances in Knowledge Discovery and Data Mining*, pp. 442–451, 2005.
 - [10] Fournier-Viger P. An open-source data mining library. <http://www.philippe-fournier-viger.com/spmf/index.php?link=datasets.php>, 2024. Accessed: 2024-12-20.
 - [11] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M. Hsu. PrefixSpan : Mining sequential patterns efficiently by prefix-projected pattern growth. In *Proceeding of 2001 international conference on data engineering*, pp. 215–224, 2001.
 - [12] R. Srikant and R. Agrawal. Mining sequential patterns: Generalizations and performance improvements. In *International Conference on Extending Database Technology*, pp. 1–17, 1996.
 - [13] R. Srikant and R. Agrawal. Mining sequential patterns: Generalizations and performance improvements. In *Proceedings of the 5th International Conference on Extending Database Technology*, pp. 3–17, 1996.
 - [14] Xifeng Yan, Jiawei Han, and Ramin Afshar. Clospan: Mining: Closed sequential patterns in large datasets. In *2003 SIAM International Conference on Data Mining*, pp. 166–177, 2003.
 - [15] Mohammed J. Zaki. Spade: An efficient algorithm for mining frequent sequences. *Machine Learning*, Vol. 42, No. 1-2, pp. 31–60, 2001.
 - [16] 横田治夫. 電子カルテデータ解析-医療支援のためのエビデンス・ベースド・アプローチ-. 共立出版, 2022.