クローズドパターン抽出を用いた 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検討

青柳　結衣　(指導教員：小口　正人)

1　はじめに

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 (minsup) 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．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．しかし，従来の手法である KISP[2] 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．

2　提案手法

図 1 に本研究の提案手法の構成を示した．提案するインタラクティブな SPM では，既存の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．

一つ目はクローズド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．それにより，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 2 に示したように，候補シーケンスを生成する際，まず候補シーケンスとその位置情報を与え，minsup と closed の条件を満たしているか検証する．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 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．最後に候補シーケンスの枝刈りをし，minsup と closed 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．

二つ目は KB の構造についてである．提案手法では，クローズドシーケンシャルパターンのみを頻出か判断するために，今までの最小 minsup である KB.base と，頻出クローズドシーケンシャルパターンとそのサポート値，さらに候補シーケンスの位置情報を保存する．頻出クローズドシーケンシャルパターンはツリー構造で管理し，候補シーケンス生成時に調べたシーケンスの位置情報はハッシュ構造で保存する．頻出と見做されないシーケンスの位置情報も保存しておくことで，次

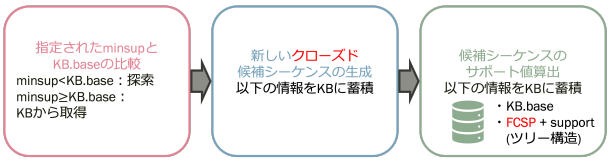

図 1: 提案手法の概要

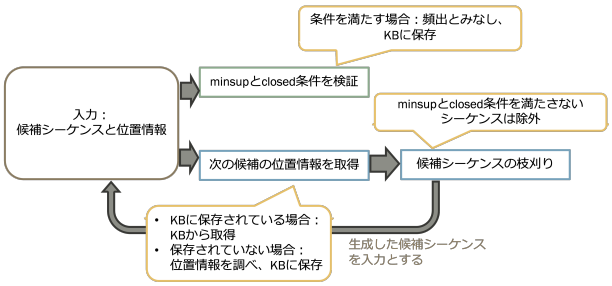

図 2: 候補シーケンス生成手順

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

回以降のマイニング時に再度候補シーケンスを生成せずにサポート値を算出可能になる．

3　実験

本実験では，提案手法において KB とクローズドと候補シーケンスの位置情報保持を考慮することの有効性を確認することを目的とする．KB の構造の有効性については今後の課題とする．

3.1　実験環境

実験環境は表 1 に示し，使用したデータセット [1] の詳細な内容は表 2 に示した．提案手法の実装には PrefixSpan[3] のコードを参考にした．

3.2　実験内容

提案手法を実装し，3 種類の実験を行った．

まず，提案手法の KB 利用による実行時間の影響を調べるため，PrefixSpan と提案手法の実行時間を比較した．BMSWebView1 を使用し，minsup を 0.01 から徐々に増やし 0.07 になるまでの実行時間を測定した．刻み幅が 0.0005 の場合は 120 回，0.0004 の場合は 150 回，0.0003 の場合は 200 回，それぞれ minsup が 0.07 になるまで実行した．

次に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

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	

増加幅	PrefixSpan(s)	提案手法(s)	提案手法/PrefixSpan
0.0005	14.86	15.16	102%
0.0004	18.43	15.29	83%
0.0003	24.75	16.02	65%

図 3: 実験 1 の結果

	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	532.73	1,154.66	46%
BMSWebView2	423.13	456.90	92%

図 4: クローズド考慮の実行時間

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 0.0001 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．

最後に候補シーケンスの位置情報の保持についても検証した．提案手法では，新しい候補シーケンスの位置情報を取得する際，KB に位置情報が保存されている場合，KB から取得し，保存されていない場合は位置情報を調べる．BMSWebView1 を使用し，minsup を 0.001 から 0.0009 になるまで 0.00001 刻みで実行し，位置情報を保存する場合としない場合の実行時間を比較した．

3.3 実験結果

3.3.1 KB 使用による実行時間の検証

minsup を 0.01 から徐々に増やし，0.07 になるまで 120 回，150 回，200 回実行した．それぞれ 3 回ずつ計測し，平均実行時間を図 3 に示した．120 回実行すると PrefixSpan の方が高速だが，150 回実行すると提案手法の方が高速である．実行回数が増えるほど，提案手法が有効であることが分かる．PrefixSpan は minsup が変わると再びマイニングを行なう．しかし，提案手法は minsup が 0.01 の時は同様のマイニングを行なうが，それ以降は KB から条件を満たすシーケンスを取得するだけなので速くなったと推測される．minsup が単調に減少する場合や，増加と減少を繰り返す場合の検証は今後の課題とする．

3.3.2 クローズド考慮による実行時間の検証

結果を図 4 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が，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 5 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．

	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 5: 候補シーケンス数の比較

	0.00100	・・・	0.00095	・・・	0.00090
パターン1	54.58	・・・	665.04	・・・	958.07
パターン2	7.73	・・・	8.96	・・・	10.81

図 6: 位置情報保持の実行時間 (s)

3.3.3 候補シーケンスの位置情報の保持

位置情報を保存する提案手法をパターン 1，毎回位置情報を調べる手法をパターン 2 として，この 2 つの実行時間を図 6 に示した．比較すると位置情報を保存しない方が高速であることがわかる．原因の追求は今後の課題とする．

4 まとめと今後の課題

SPM は候補シーケンス数が減少すると高速になる．そのため，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パターン数が減少し，アルゴリズムの高速化に繋がると推測される．本論文では既存頻出パターンの利用とクローズドの考慮によってインタラクティブな SPM の実行時間の削減への有効性を示した．

今後の課題として，クローズドの考慮と既知の頻出パターン管理におけるデータ構造の有効性をさらに検証する．また，インタラクティブな SPM においての既存研究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] 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.

[2] Lin, M. Y. and Lee, S. Y.: Improving the efficiency of interactive sequential pattern mining by incremental pattern discovery, in *International Conference on System Sciences*, pp. 68–76, Hawaii, USA (2002), IEEE.

[3] Pei, J., Han, J., Mortazavi-Asl, B., Pinto, H., Chen, Q., Dayal, U. and Hsu, M.: PrefixSpan : Mining sequential patterns efficiently by prefix-projected pattern growth, in *Proceeding of 2001 international conference on data engineering*, pp. 215–224 (2001).