

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[†] Le Hieu Hanh[†] 松尾亮輔[†] 山崎友義[†] 荒木賢二[†] 横田治夫[‡] 小口正人[†]
お茶の水女子大学[†] 城西大学[‡]

1. 背景

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、従来の手法である KISP[1] 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

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 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。

二つ目は KB の構造についてである。KISP では今までの最小 minsup である KB.base と、頻出シーケンシャルパターンとそのサポート値を保存する。頻出シーケンシャルパターンは、候補シーケンスを生成しやすいようにサイズごとにグ

ループ化して保存している。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、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。KISP から候補シーケンス生成方法を変更したため、頻出クローズドシーケンシャルパターンはパターンサイズごとにグループ化せず、ツリー構造で管理する。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

KB についての実験環境は表 1、クローズド考慮についての実験環境は表 2に示した。使用したデータセットの詳細な内容は表 3に示した。提案手法の実装には PrefixSpan[2] のコードを参考にした。

表 1 実験環境 1

CPU	Apple M1
OS	macOS Ventura 13.3.1
メモリ	8GB
ディスク	NVMe SSD
python	3.13.1

表 2 実験環境 2

CPU	Intel Xeon Platinum 8260L x 2
OS	Ubuntu 22.04.5 LTS
メモリ	384MB
python	3.10.12

Improving the Efficiency of Interactive Sequential Pattern Mining by Closed Pattern Discovery

[†]Yui Aoyagi

[†]Hieu Hanh Le

[†]Ryosuke Matsuo

[†]Tomoyoshi Yamazaki

[†]Kenji Araki

[‡]Haruo Yokota

[†]Masato Oguchi

[†]Ochanomizu University

[‡]Josai University

表 3 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	

3.2. 実験内容

提案手法を実装し、2 種類の実験を行った。

まず、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[3] 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。刻み幅を 0.0005 に設定した場合は 120 回、0.0004 に設定した場合は 150 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しない場合の実行時間を測定し比較した。BMSWebView1 と BMSWebView2[3] 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。

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 回実行するのにかかった時間の比較を表 4 に示し、同様の範囲で minsup を増やし、150 回実行した時の比較を表 5 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

結果を表 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、

表 4 実行時間(120)

	PrefixSpan(s)	提案手法(s)
1回目	7.28	7.67
2回目	6.78	7.65
3回目	6.76	7.77
平均	6.94	7.70

表 5 実行時間(150)

	PrefixSpan(s)	提案手法(s)
1回目	8.49	7.69
2回目	8.90	7.77
3回目	8.33	7.73
平均	8.57	7.73

表 6 クローズド考慮の実行時間

	クローズド考慮あり(s)	クローズド考慮なし(s)
BMSWebView1	557.53	1158.73
BMSWebView2	434.28	441.01

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。実行時間の差は候補シーケンス数の削減量と相関があると推測されるため、候補シーケンス数の計測を行う必要がある。

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 における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] 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, Hawaii, USA, 2002. IEEE.
- [2] 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.
- [3] Fournier-Viger P. An open-source data mining library. <http://www.philippe-fournier-viger.com/spmf/index.php?link=datasets.php>, 2024. Accessed: 2025-1-9.