

MINING SEQUENCE AND INTER-SEQUENCE PATTERNS IN LARGE DATABASES

Anh Nguyen

Department of Applied Informatics, Wroclaw University of Science and Technology, Poland

1 Introduction

Inter-sequence pattern mining is an extension of sequential pattern mining that involves discovering common patterns, associations, and dependencies between sequences in a sequential database. It identifies patterns that are common not only within the same transaction but also between transactions. The inter-sequential pattern mining problem was first proposed by Wang et al. in 2009 (Wang & Lee, 2009), and several algorithms have been developed since then, including DBV-ISP (Vo et al., 2012), ISP-IC, iISP-IC and pISP-IC (Le et al., 2018), DBV-ISPMIC, and pDBV-ISPMIC (Nguyen et al., 2023).

2 Research Problem

The problem of inter-sequence pattern mining is an extension of sequential mining, with three extensions including itemset, sequence, and inter. When dealing with large sequential databases and a high user-specified *maxspan* value, a large number of candidate patterns are generated. Therefore, the inter-sequence pattern mining problem requires optimization of the candidate generation process as well as optimization of storage space, as shown in Table 1.

Table 1. Comparing the Advantages and Disadvantages of Proposed Inter-Sequence Pattern Mining Algorithms

References	Data structure	Advantage	Disadvantage
Wang & Lee (2009)	PatternList	The inter-sequence pattern problem was first introduced. The EISP-Miner only requires a single database scan and can confine the joining and support counting operations to a small number of pattern lists, thus avoiding expensive subset inter-sequence matching. It also exhibits greater efficiency compared to M-Apriori.	Many patterns are generated, requiring a significant amount of storage space.
Vo et al. (2012)	DBV- PatternList	The DBV-ISP algorithm employs the DBV-PatternList data structure to store information regarding frequent inter-sequence patterns, resulting in reduced memory usage compared to the EISP-Miner algorithm. Additionally, support calculation is	The algorithm utilizes a set of contiguous bits to store the positional information of the pattern. Therefore, when the pattern is rarely encountered in the database, positions that are absent within the pattern must be represented by a bit value of 0. This results in the utilization of more memory than necessary.

		expedited as it operates on bits.	
T. Le et al. (2018)	DBV- PatternList	The ISP-IC algorithm applies an item condition to the inter-sequence pattern mining problem. The regulatory limit is quite extensive, allowing for the restriction of the number of inter-sequence patterns that can be generated.	The algorithm has not yet addressed the limitations of the DBV- PatternList data structure.

2.1 Aim of this Thesis

The aim of this thesis is to address the limitations of inter-sequence pattern mining in terms of processing time and storage space. The thesis proposes a novel storage data structure for the inter-sequence pattern mining problem, aiming to minimize data duplication during the mining process. Additionally, it introduces an inter-sequence pattern mining model with itemset constraints to reduce the number of generated candidates, thus accelerating the search and processing of relevant information. Moreover, the thesis presents additional propositions to enhance the efficiency of the proposed methods and algorithms.

2.2 Objectives of this Thesis

The objectives of this thesis are as follows:

1. Proposing a solution to address the problem of mining inter-sequence patterns with itemset constraints, introducing the DBV-ISPMIC algorithm.
2. Developing an optimal approach for solving the inter-sequence pattern mining problem with itemset constraints, presenting the *p*DBV-ISPMIC algorithm.
3. Introducing a method for optimizing storage space in the context of the inter-sequence mining problem, utilizing the ISP-PI (Inter-Sequence Pattern mining based on Pseudo-Index) algorithm.
4. Proposing a candidate pruning technique for the inter-sequence pattern mining problem, incorporating the ISP-IC (Inter-Sequence Pattern mining with Index Intersection Checking) method.
5. The proposed algorithms and methods will be evaluated through experiments using real-world databases sourced from the data mining community's data warehouse. The experimental results will be compared based on the algorithm's running time and memory usage requirements.

3 Thesis Contributions

Based on the aim of the thesis, the main contributions are presented in two sections and are briefly outlined as follows:

3.1 The First Contribution Addresses Objectives 1 and 2 of the Thesis

Drawing from proposed inter-sequence mining problems such as EISP-Miner (Wang & Lee, 2009), DBV-ISP (Vo et al., 2012), and ISP-IC (Le et al., 2018), as well as sequential pattern mining problems with itemset constraints like MSPIC-DBV (Van et al., 2018), we introduce a problem of inter-sequence pattern mining with itemset constraints, named DBV-ISPMIC (Nguyen et al., 2023). The algorithm employs a data structure called DBV-PatternList to store candidates, along with a tree structure named ISP-Tree to store frequent patterns. Additionally, we propose a method for quickly checking the condition of generated candidate itemsets and apply parallel mining to accelerate the algorithm.

The DBV-PatternList data structure optimizes candidate information storage. Instead of using a numeric data type to represent pattern information, DBV-PatternList utilizes a bit-vector data structure. Pattern information is indicated by turning bits on or off, allowing a numeric data type to store more candidate information, thus reducing the space needed for candidates.

Checking itemset constraints for all generated samples is time-consuming for the algorithm. We suggest a method to rapidly verify that the generated candidate meets itemset constraints, using the condition information from the parent patterns that created it. This decreases the algorithm's running time.

The inter-sequence pattern mining problem employs the ISP-Tree structure to store generated frequent patterns, processing the algorithm according to the depth-first traversal method. As the handling of branches on the tree is separate, we present a parallel processing technique for branches on the tree. This enables the algorithm to optimize runtime by processing multiple branches simultaneously.

The evaluation of experimental results is presented in article [R2].

3.2 The Second Contribution Addresses Objectives 3 and 4 of the Thesis

Building upon inter-sequence pattern mining algorithms like EISP-Miner and DBV-ISP, we propose a novel algorithm called ISP-PI. This algorithm aims to optimize data mining models in the context of inter-chain mining using a data structure known as Pseudo-IDList. The ISP-PI addresses the shortcomings of previous algorithms concerning data duplication. Instead of requiring storage for all the information of a candidate, we can retrieve its information from the original pattern. This method compresses the position values of the generated candidates, allowing for the retrieval of values from the original patterns that produced the candidates and eliminating the need to save all positions.

Furthermore, the algorithm incorporates a pruning method named ISP-IC to effectively reduce the number of generated candidates. This optimization enhances

processing time and storage space, which is crucial due to the growing volume of collected data. The ISP-PI algorithm efficiently compresses data to minimize storage space and employs candidate pruning to accelerate the algorithm's runtime in inter-sequence pattern mining.

The evaluation of the research results, along with the proposed algorithm, is presented in article [R1].


4 Publications

- [R1]. **Nguyen, A.**, Nguyen, N. T., Nguyen, L. T. T., & Vo, B. (2023). An Efficient Pruning Method for Mining Inter-sequence Patterns based on Pseudo-IDList. *Expert Systems with Applications* (Accepted, Impact Factor: 8.5, Punktacja MEiN: 200)
- [R2]. **Nguyen, A.**, Nguyen, N. T., Nguyen, L. T. T., & Vo, B. (2023). Mining inter-sequence patterns with Itemset constraints. *Applied Intelligence*, 53(17), 19827–19842. <https://doi.org/10.1007/S10489-023-04514-7> (Impact Factor: 5.3, Punktacja MEiN: 70)
- [R3]. Nguyen, T. T. D., Nguyen, L. T. T., **Nguyen, A.**, Yun, U., & Vo, B. (2021). A method for efficient clustering of spatial data in network space. *Journal of Intelligent & Fuzzy Systems*, 40(6), 11653–11670. <https://doi.org/10.3233/JIFS-202806> (Impact Factor: 1.737, Punktacja MEiN: 70)
- [R4]. Huynh, H. M., Nguyen, L. T. T., Vo, B., **Nguyen, A.**, & Tseng, V. S. (2020). Efficient methods for mining weighted clickstream patterns. *Expert Systems with Applications*, 142, 112993. <https://doi.org/10.1016/j.eswa.2019.112993> (Impact Factor: 6.954, Punktacja MEiN: 140)
- [R5]. Nguyen, L. T. T., Nguyen, T. D. D., **Nguyen, A.**, Tran, P.-N., Trinh, C., Huynh, B., & Vo, B. (2020). Efficient Method for Mining High-Utility Itemsets Using High-Average Utility Measure. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12496 LNAI, 305–315. https://doi.org/10.1007/978-3-030-63007-2_24 (Core ranking: C, Punktacja referatów z wykazu MEiN: 20)
- [R6]. Nguyen, L. T. T., Vo, B., Nguyen, T. N., & **Nguyen, A.** (2019). Mining class association rules on imbalanced class datasets. *Journal of Intelligent and Fuzzy Systems*, 37(6), 7131–7139. <https://doi.org/10.3233/JIFS-179326> (Impact Factor: 1.851, Punktacja MEiN: 70)
- [R7]. Le, T., **Nguyen, A.**, Huynh, B., Vo, B., & Pedrycz, W. (2018). Mining constrained inter-sequence patterns: a novel approach to cope with item constraints. *Applied Intelligence*, 48(5), 1327–1343. <https://doi.org/10.1007/s10489-017-1123-9> (Impact Factor: 2.882, Q2)

5 References

- Le, T., Nguyen, A., Huynh, B., Vo, B., & Pedrycz, W. (2018). Mining constrained inter-sequence patterns: a novel approach to cope with item constraints. *Applied Intelligence*, 48(5), 1327–1343. <https://doi.org/10.1007/s10489-017-1123-9>

- Nguyen, A., Nguyen, N. T., Nguyen, L. T. T., & Vo, B. (2023). Mining inter-sequence patterns with Itemset constraints. *Applied Intelligence*, 1–16. <https://doi.org/10.1007/S10489-023-04514-7>
- Van, T., Vo, B., & Le, B. (2018). Mining sequential patterns with itemset constraints. *Knowledge and Information Systems*, 57(2), 311–330. <https://doi.org/10.1007/s10115-018-1161-6>
- Vo, B., Tran, M. T., Hong, T. P., Nguyen, H., & Le, B. (2012). A dynamic bit-vector approach for efficiently mining inter-sequence patterns. *Proceedings - 3rd International Conference on Innovations in Bio-Inspired Computing and Applications, IBICA 2012*, 51–56. <https://doi.org/10.1109/IBICA.2012.31>
- Wang, C. S., & Lee, A. J. T. (2009). Mining inter-sequence patterns. *Expert Systems with Applications*, 36(4), 8649–8658. <https://doi.org/10.1016/j.eswa.2008.10.008>


71/11/2023