Construction Pattern Mining Algorithm for Massive Construction Plans

Yu Liu[1,a], Lei Yan[1,b], Miao Li*[2,c]

[1]China Railway No. 3 Engineering Group Co, Ltd. Taiyuan, China.
[2]China Railway Electrification Engineering Group Co, Ltd. Beijing, China

Email: [a]416288149@qq.com, [b]452420271@qq.com, [c]1067707975@qq.com

Abstract. In a plethora of construction plans lie effective construction plans tailored to specific scenarios. When faced with new construction challenges, we often seek inspiration from historical construction plans, hoping to immediately derive effective solutions upon presenting construction problems. In the era of large-scale modeling, there is a growing desire for artificial intelligence models to learn effective problem-solving approaches from a vast array of construction plans. Prior to this, we need to process the massive volume of construction plans, abstracting and refining those that are effective for specific scenarios. We define these as "construction modes" and propose a Construction Plan Mining Algorithm (CPMA) aimed at extracting such modes from the vast collection of construction plans. The algorithm first serializes all documents and then employs a sequence expansion strategy to extend frequent items within the documents, generating new sequences. Utilizing a non-continuous pattern pruning strategy ensures the generation of only continuous sequences, thus maintaining the effectiveness of construction modes. Subsequently, a non-maximal pattern filtering strategy is used to filter the result set into maximal sequences, ensuring the integrity of sequence patterns. Through these three strategies, a continuous and maximal sequence pattern result set is generated. The algorithm was experimented on 1. 53 million words of publicly available construction plans, demonstrating that the construction mode mining process, with CPMA as its core, exhibits better operational efficiency and superior result outputs. This holds groundbreaking significance in the field of construction mode mining, while also providing insights and practical technological means to reduce redundancy in mining.

Keywords: Construction plan, construction pattern, data mining, serialization

1 INTRODUCTION

Mining practical patterns from a vast amount of documents is a hot research topic that has yielded abundant results[1]. Construction modes, in simple terms, are effective construction plans tailored to specific scenarios that can effectively address construction

B. Yuan et al. (eds.), *Proceedings of the 2024 8th International Conference on Civil Architecture and Structural Engineering (ICCASE 2024)*, Atlantis Highlights in Engineering 33,
https://doi.org/10.2991/978-94-6463-449-5_64

difficulties. Pattern mining is an important task in data mining and finds wide applications in various fields such as web pattern mining, drug pattern mining, bug analysis in software code, and analysis of biological gene construction data[2] [3]. We hope to effectively mine construction modes and shine in new fields.

Existing pattern mining algorithms suffer from the problem of including a large number of candidate patterns in the mining results, making it difficult for users to analyze and select results, and the algorithms also exhibit poor efficiency in terms of time and space[4]. To address these issues, concepts such as closed sequence patterns and maximal sequence patterns have been proposed. Closed sequence patterns filter out some irrelevant content but still appear redundant for large or lengthy construction databases. Maximal sequence pattern mining is faster compared to closed sequence pattern mining while reducing redundancy. However, existing algorithms still suffer from inefficiency and inefficient memory usage.

In recent years, there has been attention towards meeting the constraint of continuity in sequence patterns, and closed continuous construction patterns are the current research focus. This pattern reflects the adjacency of projects and the closure of patterns, helping to overcome the drawbacks of traditional algorithms in mining continuous information[5], [6].

To address the aforementioned issues, this paper proposes a Construction Plan Mining Algorithm (CPMA) tailored for construction plans. The algorithm combines the characteristics of maximal sequence patterns and closed continuous sequence patterns, offering advantages of high efficiency and low redundancy. It generates candidate patterns through a new pattern expansion method, enhances computational efficiency using vertical id lists, and introduces continuity constraints to produce more compact result sets[7]. Experimental results demonstrate that CPMA outperforms existing algorithms in terms of efficiency and result accuracy, particularly showing significant advantages in the field of construction mode mining.

The remainder of this paper is organized as follows: Section II discusses the properties of maximal continuous construction patterns and the algorithm execution flow; Section III verifies the superiority of the CPMA algorithm through experiments; finally, Section IV summarizes and provides outlooks for the research.

2 MAXIMUM CONTINUOUS CONSTRUCTION PATTERN MINING ALGORITHM

This chapter will detail the process of mining the maximal construction patterns using the CPMA algorithm and the mining strategies it entails. Through these strategies, CPMA can swiftly extract patterns from construction plans that adhere to continuous and maximal constraints. Additionally, this paper also provides strategies for construction pattern recovery to meet the requirements of certain applications for sub-patterns used in maximal patterns.

The entire mining process is illustrated in the figure(Fig 1), with CPMA taking construction plans SDB and minimum support minsup as input. Initially, SDB is scanned entirely to construct a vertical database (VerticalDB) (SDB), which is a key-value pair

structure where keys represent items, and values represent the vertical ID lists corresponding to the keys. Subsequently, items with support less than the minimum support minsup are filtered from VerticalDB (SDB) (these items will not appear in any frequent construction patterns), and frequent items along with their vertical ID lists are stored in the frequent single-item construction set Fs. For each frequent single-item construction s ∈ Fs, a construction extension algorithm is invoked.

The construction extension algorithm returns a construction pattern <s>, and recursively extends candidate constructions starting from the prefix <s> until no new satisfying constructions can be extended further. The extension algorithm takes the prefix construction <s> and a frequent item k ∈ Fs as input, creates a new construction by joining <s> and k, and adds a continuity constraint to ensure that only contiguous positions are retained in the new table. For each candidate construction generated through extension, its support is quickly computed using the vertical ID lists to determine if it is frequent. If the newly extended construction pattern is frequent, the new construction pattern <s, k> is again recursively executed as a prefix for the extension algorithm.

Before new continuous construction patterns are generated and prepared for output, a pattern filtering step must be performed using a maximum pattern database MP to store and check new construction patterns. Through two pattern filtering strategies, MP ensures that only maximal construction patterns are retained in the database. Upon completion of the output of all new constructions, the maximum database MP stores the final results of the CPMA algorithm.

During the execution of the algorithm, CPMA employs several strategies to ensure the mining of continuous and maximal construction patterns. This section describes in detail how to mine them. By integrating three strategies (construction extension strategy, non-continuous pattern pruning strategy, and non-maximal pattern filtering strategy), CPMA effectively filters out non-maximal and non-continuous patterns while pruning the search space to improve mining efficiency. Finally, construction recovery strategies are provided to meet the needs of some applications for sub-patterns of maximal patterns.

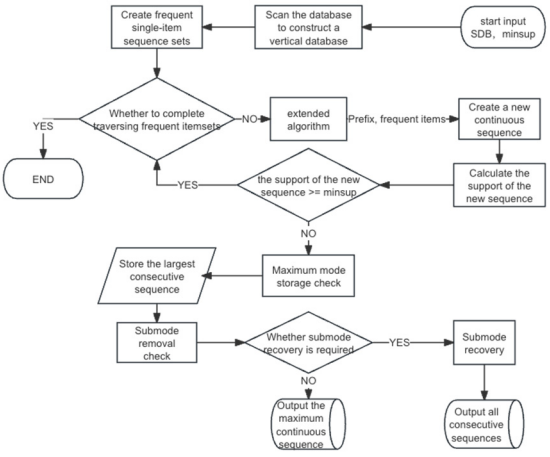

Fig. 1. CPMA algorithm mining flow chart

The construction extension strategy is an effective method for generating longer constructions from shorter ones by treating shorter constructions as prefixes, concatenating frequent items for each prefix, and utilizing pruning strategies to generate vertical ID lists for new constructions. The non-continuous construction pattern pruning strategy introduces a constraint to ensure that the positions of items in the new sequence are contiguous in the original dataset, avoiding the generation of non-continuous new construction patterns. The non-maximal pattern filtering strategy aims to reduce redundancy by filtering out non-maximal patterns before saving new constructions, thus ensuring that only maximal patterns are retained.

Finally, the sequence recovery strategy is proposed to recover all continuous construction patterns from the set of maximal continuous construction patterns[8]. This simple algorithm generates all continuous subsequences of maximal continuous patterns. To prevent duplicate outputs, an additional check is performed to determine whether a construction has already been output. If exact support for each construction pattern is required, a full-text scan can be performed after recovering all sequences to calculate their exact support.

3 EXPERIMENT

3.1 Experimental Design

For construction mode mining, this paper conducts mining experiments based on different algorithms using seven different categories of datasets commonly found in construction plans. All data are sourced from publicly available construction plans online, totaling 1. 53 million words in content. The detailed information of the experimental datasets is shown in the Table1 below:

Table 1. Construction pattern mining of proprietary domain data sets

Item category	Number of words (10,000 words)
Building construction plan	16
road construction plan	19
Water conservancy construction plan	14
Pipeline construction plan	21
Electric power engineering construction plan	33
Garden landscape construction plan	22
Marine engineering construction plan	28
total	153

To validate the excellent performance of the CPMA algorithm proposed in this paper in the field of construction mode mining, construction mode mining experiments were designed to demonstrate the effectiveness of CPMA in mining results through experiments. The experiment compared the effectiveness of CPMA with VMSP and MaxSP in handling dataset data. VMSP and MaxSP are both state-of-the-art maximal sequence

pattern mining algorithms that utilize different structural additions to enforce maximal constraints in discovering maximal construction patterns from sequence patterns.

3.2 Analysis of Results

To validate the effectiveness and superiority of the CPMA algorithm in the field of construction mode mining, this paper draws upon commonly used result evaluation methods in the information retrieval domain as well as in the construction mode mining domain. Combining the characteristics of process mode mining, precision (P) and recall (R) are defined as follows:

$$\mathrm{P}=\frac{PropR}{PropM}\times 100\% \tag{1}$$

$$\mathrm{R}=\frac{PropR}{Prop}\times 100\% \tag{2}$$

Where Prop represents the number of actual construction patterns existing in the dataset; PropM represents the number of construction patterns obtained by mining the dataset using different algorithms; and PropR represents the total number of correct patterns among those obtained by different algorithms. When calculating recall and precision, the Prop parameter is evaluated by multiple domain experts. After modeling and hash serialization of construction plans, numerical sequences can be generated. By mining construction patterns from these numerical sequences, pattern information embedded in the construction plans can be obtained. By applying the methods mentioned above to mine construction plans and analyzing the mining results of the three algorithms, the results of recall and precision are calculated as shown in the table.

Table 2. recall rate and precision rate

	Prop	*PropM*	*PropR*	*P*	*R*
C	21	18	18	100%	85. 7%
V	21	26	16	61. 5%	76. 1%
M	21	23	16	69. 5%	76. 1%

C represents the mining results with CPMA as the core algorithm. V represents the mining results with VMSP as the core algorithm. M represents the mining results with MaxSP as the core algorithm. From Table 2, it can be seen that none of the methods have mined all patterns. However, the recall rate of CPMA in this paper is higher. Due to the different writing habits of different workers, there are differences in writing styles, which leads to the discontinuity of construction pattern statements and the presence of redundant words between sentences, resulting in the failure to discover the expected construction patterns in the continuous mining process. This is also the limitation of continuous construction pattern mining.

At the same time, the precision rate of this paper reached 100%, proving that the construction pattern results mined in this paper are all correct. The number of patterns

mined by the other two algorithms exceeds the actual number of existing patterns, resulting in a precision rate of less than 70%. Through the analysis of the mining results, we found a large number of redundant patterns. CPMA only mines maximal construction patterns, while other traditional continuous construction pattern mining algorithms will mine both the maximal pattern and its sub-patterns, outputting the sub-patterns of the maximal construction pattern to the result set. Since the maximal pattern contains all sub-patterns, the output of sub-patterns is unnecessary. A higher precision rate proves the effectiveness of the maximal continuous pattern mining algorithm CPMA in the field of construction pattern mining, and CPMA has advantages over continuous pattern mining.

4 CONCLUSIONS

In this paper, we propose the CPMA algorithm, which is a new algorithm for mining maximal continuous construction patterns that can be applied in construction plan mining. It stores items and constructions' positions in the database using vertical ID lists and employs three mining strategies to ensure the extraction of maximal and continuous construction patterns. Specifically, by expanding constructions to generate new ones, the algorithm avoids the generation of non-continuous sub-constructions and prunes a significant portion of the search space through a non-continuous pruning strategy. It mines maximal continuous closed construction patterns by quickly calculating support and pruning the search space. Additionally, we provide a construction recovery strategy that can recover all continuous sub-construction patterns, preserving the ability to discover other construction pattern mining results and meeting different pattern mining needs.

Our experiments on publicly available datasets demonstrate that CPMA achieves excellent results in construction pattern mining tasks compared to the two state-of-the-art algorithms in this field[9] [10], VMSP and MaxSP. While improving mining efficiency, CPMA also enhances recall and precision of mining results, proving the significance of mining maximal construction patterns in construction plan mining.

Our work opens up a new important avenue for future research in the construction plan mining field. By mining maximal continuous patterns, redundancy can be significantly reduced, and the precision of construction pattern mining can be greatly improved. However, due to the flexibility in plan writing, improving recall remains challenging. Future algorithm research should focus on non-continuous pattern mining to discover methods for improving recall.

REFERENCES

1. Fig Chen M S, Han J, Yu P S. Data mining: an overview from a database perspective[J]. IEEE Transactions on Knowledge and data Engineering, 2016, 8(6): 866-883.
2. iel A, KLAČKOVÁ I. Safety requirements for mining systems controlled in automatic mode[J]. Acta Montanistica Slovaca, 2020, 25(3).

3. Bi L, Wang Z, Wu Z, et al. A new reform of mining production and management modes under Industry 4. 0: Cloud mining mode[J]. Applied Sciences, 2022, 12(6): 2781.
4. Fournier-Viger P, Lin J C W, Kiran R U, et al. A survey of sequential pattern mining[J]. DataScience and Pattern Recognition, 2023, 1(1): 54-77.
5. Han J, Cheng H, Xin D, et al. Frequent pattern mining: current status and future directions[J]. Data mining and knowledge discovery, 2017, 15(1): 55-86.
6. Mabroukeh N R, Ezeife C I. A taxonomy of sequential pattern mining algorithms[J]. ACM Computing Surveys (CSUR), 2020, 43(1): 1-41.
7. Mooney C H, Roddick J F. Sequential pattern mining--approaches and algorithms[J]. ACM Computing Surveys (CSUR), 2023, 45(2): 1-39.
8. Ryang H, Yun U. Top-k high utility pattern mining with effective threshold raising strategies[J]. Knowledge-Based Systems, 2021, 76: 109-126.
9. Ryang H, Yun U. High utility pattern mining over data streams with sliding window technique[J]. Expert Systems with Applications, 2022, 57: 214-231.
10. Lin J C W, Li T, Pirouz M, et al. High average-utility sequential pattern mining based on uncertain databases[J]. Knowledge and Information Systems, 2020, 62(3): 1199-1228.