A Comparative Study of Frequent and Maximal Periodic Pattern Mining Algorithms in Spatiotemporal Databases

O Obulesu, Assistant Professor, Department of IT, SVEC, Tirupati, A.P., India
Dr A Rama Mohan Reddy, Professor of CSE, SVUCE, Tirupati, A.P., India
Mahendra M, Assistant Professor, Department of IT, SVEC, Tirupati, A.P., India

Abstract: Detecting regular and efficient cyclic models is the demanding activity for data analysts due to unstructured, vigorous and enormous raw information produced from web. Many existing approaches generate large candidate patterns in the occurrence of huge and complex databases. In this work, two novel algorithms are proposed and a comparative examination is performed by considering scalability and performance parameters. The first algorithm is, EFPMA (Extended Regular Model Detection Algorithm) used to find frequent sequential patterns from the spatiotemporal dataset and the second one is, ETMA (Enhanced Tree-based Mining Algorithm) for detecting effective cyclic models with symbolic database representation. EFPMA is an algorithm grows models from both ends (prefixes and suffixes) of detected patterns, which results in faster pattern growth because of less levels of database projection compared to existing approaches such as Prefixspan and SPADE. ETMA uses distinct notions to store and manage transactions data horizontally such as segment, sequence and individual symbols. ETMA exploits a partition-and-conquer method to find maximal patterns by using symbolic notations. Using this algorithm, we can mine cyclic models in full-series sequential patterns including subsection series also. ETMA reduces the memory consumption and makes use of the efficient symbolic operation. Furthermore, ETMA only records time-series instances dynamically, in terms of character, series and section approaches respectively. The extent of the pattern and proving efficiency of the reducing and retrieval techniques from synthethic and actual datasets is a really open & challenging mining problem. These techniques are useful in data streams, traffic risk analysis, medical diagnosis, DNA sequence Mining, Earthquake prediction applications. Extensive investigational outcomes illustrates that the algorithms outperforms well towards efficiency and scalability than ECLAT, STNR and MAFIA approaches.

Keywords: Maximal Patterns, Periodicity, Sequential patterns, Spatiotemporal Databases.

1. Introduction

Data mining refers to extracting or “mining” knowledge from large amounts of data. The term is actually a misnomer. Remember that the mining of gold from rocks or sand is referred to as gold mining rather than rock or sand mining. Thus, data mining should have been more appropriately named “knowledge mining from data,” which is unfortunately somewhat long. “Knowledge mining,” a shorter term may not reflect the emphasis on mining from large amounts of data. Many people treat data mining as a synonym for another popularly used term, Knowledge Discovery from Data, or KDD. Alternatively, others view data mining as simply an essential step in the process of knowledge discovery.

1.1 Basic Pattern Mining Algorithms

The first algorithm is proposed by R. Agrawal et al. (1994), for market basket analysis in the form of association rule mining [1]. There are five types of issues such as efficient and scalable methods for mining frequent patterns, mining interesting frequent patterns, Impact to data analysis and mining applications, Applications of frequent patterns and research directions. The efficient and scalable methods for mining frequent patterns are Apriori [2] and FP-Growth [7] and ECLAT algorithm (M. Zaki) [32]. A k-itemset is frequent only if all of its sub-itemsets are frequent and no superset of an infrequent itemset can be frequent.

A large pattern will contain an exponential number of smaller, frequent sub-patterns. If the min_sup is low then huge Frequent Itemsets will be generated. To restrict those patterns we develop
closed frequent patterns and maximal frequent patterns mining techniques. A pattern \( X \) is a closed frequent pattern in a data set \( D \) if \( X \) is frequent in \( D \) and there exists no proper super-pattern \( Y \) such that \( Y \) has the same support as \( X \) in \( D \). i.e. \( C = \{ X | X \in F \text{ and no } Y \supset X \text{ such that } \sup(X) = \sup(Y) \} \). \( X \) is closed if all supersets of \( X \) have strictly less support, i.e., \( \sup(X) > \sup(Y) \), for all \( Y \supset X \). A pattern \( X \) is a maximal frequent pattern, if \( X \) is frequent, and there exists no super-pattern \( Y \) such that \( X \subset Y \) and \( Y \) is frequent in \( D \). i.e. \( M= \{ X | X \in F \text{ and no } Y \supset X, \text{ such that } Y \in F \} \). Eg: There are only 2 Maximal Frequent Itemsets with min_sup = 3 in the table.4, i.e. ABDE and BCE.

The various closed pattern mining algorithms developed are A-Close (1999), Closet [5], CHARM [10], Closet+ [14], FPClose (2003), AFOPT (2003) and the algorithm designed for mining max-patterns is first proposed by Bayardo [3] and the other one is Max-Miner algorithm (Apriori-based, level-wise, BFS method), Max-itemset-superset frequency & subset infrequency pruning. The algorithms designed for mining high-dimensional datasets & mining colossal patterns are CARPENTER (2003) [45], COBBLER (2004), TD-Close (2006) and the algorithms develop for mining sequential patterns are Generalized Sequential Patterns-GSP (1996), SPADE [13], Concept – BFS & Apriori Pruning, PrefixSpan (2004) [42] is better than SPADE than GSP, CloSpan (2003), BIDE (2004) [44] , Acyclic Graphs of events (1997), Regular Expressions specifications [4], Multilevel & Multidimensional Sequential Pattern Mining (2001), CLUSEQ-a sequence clustering algorithm (2003), IncSpan – an incremental Sequential Pattern Mining algorithm (2004), SeqIndex - sequence indexing (2004), Parallel mining of closed sequential patterns (2005), MSPX - Maximal sequential patterns by using multiple samples (2005).

1.2 Periodic Pattern Mining Basics

A full periodic pattern is a pattern where every point in time contributes (precisely or approximately) to the cyclic behaviour of a time-related sequence. For example, all of the days in the year approximately contribute to the season cycle of the year. A partial periodic pattern specifies the periodic behaviour of a time-related sequence at some but not all of the points in time. For example, Sandy reads the New York Times from 7:00 to 7:30 every weekday morning, but her activities at other times do not have much regularity. Partial periodicity is a looser form of periodicity than full periodicity and occurs more commonly in the real world [6].

1.3 Parameters Identified for Pattern Mining Research

The various parameters used for frequent pattern mining area are mentioned below.

- \(|D|\) - Number of Transactions
- \(|T|\) - Average size of Transactions
- \(|I|\) - Average size of maximal large Itemsets
- \(|L|\) - Number of maximal large Itemsets
- \(|N|\) - Number of Items
- Execution Time (Secs) Vs Algorithms
- Number of transactions Vs Time (Secs)
- Number of Items Vs Time (Secs)
- \(|D_c|\) - Number of Customers
- \(|C|\) - Average number of Transactions per Customer
- \(|T|\) - Average number of Items per Transaction
- \(|S|\) - Average length of maximal large sequences
- \(|I|\) - Average size of Itemsets in maximal large sequences
- \(|N_s|\) - number of maximal large sequences
- \(|N_i|\) - number of maximal large Itemsets
- Period Length Vs Time & Shifted/Distorted Instances Vs Time
- Data set Vs Min_support Vs Execution Time Vs memory
- Number of Transactions Vs Relative Time
- Number of Items Vs Time (Secs) Vs Memory
- Transaction size Vs Time (Secs)
- The Length of the Sequence & Maximal Pattern Vs Time
Number of Transactions Vs Customer Vs Time

1.4 Datasets Identified for Pattern Mining Research

The following synthetic and real datasets are used for frequent pattern mining research such as Mushrooms, Iris, Chess, Connect-4, T5.I2.D100K, T10.I4.D100K, T20.I6.D100K, T40.I10.D100K, Cancer, Pumsb, Airlines, Census, Gazelle, Bus Movements in Greece.

2. Overview of Proposed Work

The proposed two algorithms EFPMA and ETMA used for pattern mining approach such as given below.

The main contributions accomplished in this research work are:

i. Detecting Regular models using Extended Regular Model Detection Algorithm: An example scenario between forestfires dataset is demonstrated to explain the importance and the requirement of detecting regular models useful in the forestfire prevention, Telecommunication fraud detection and DNA Sequence mining applications etc.,

ii. Mining Maximal patterns using Enhanced Tree Mining Algorithm: A Review on all the existing maximal pattern mining techniques are analyzed and the taxonomy is framed for the techniques studied. Finally, ETMA finds effective maximal patterns from various synthetic and Accident, Retail datasets.

iii. A Methodology for spatiotemporal data mining in Big Data: A new framework has designed to apply knowledge gained from identified cyclic models in huge data.

Evaluation Criteria: Series of experiments conducted with each consisting of seven different test data sets are evaluated to assess the efficiency of the ETMA and EFPMA algorithms towards CPU
time consumption Vs Number of Transactions/items, Average transaction length; Minimum support and the obtained patterns between synthetic and real databases are analyzed.

**Conception and Design of EFPMA algorithm:** The research outlined a novel strategy to introduce the issue of cyclic model detection. This technique covers the fundamental part of balancing the database size and quantity of records Vs number of distinct itemsets and knowledge discovery. Apart from the algorithms, few measures to evaluate the effect of the ETMA and EFPMA techniques on the synthetic and actual data and on the results of model extraction are considered.

**Experimentation:** Some current outperformance [10, 26] algorithms, methods were implemented, tested and applied the framework on real datasets to investigate the scalability and run time issues to estimate the functionality of the algorithms.

**Evaluation:** Assessment criteria are specified to study the EFPMA algorithm. The effect of this method on a knowledge mining activity was assessed by evaluating the outcome of the task related to a dataset with and without conversion. The newly intended algorithm is studied with published results. Standard metrics are utilized to assess the efficacy of these techniques. Scalability of the algorithm is also assessed.

### Table 2.1: Characteristics of the Synthetic and Actual Datasets

| Dataset    | Average Quantity of Items | Average Transaction Length | #Transactions | Average size of maximal patterns |
|------------|----------------------------|----------------------------|---------------|---------------------------------|
| Accident   | 45                         | 15.5                       | 65,536        | 4.56                            |
| Retail     | 16,470                     | 10.3                       | 65,536        | 15.6                            |
| Mushroom   | 119                        | 23                         | 8,124         | 22.5                            |
| Pumsb_star | 7,116                      | 10.5                       | 49,046        | 50.5                            |
| T10I4D100K | 100                        | 10                         | 65,536        | 40.5                            |
| T40I10D100K| 100                        | 20                         | 65,536        | 25.5                            |

### Table 2.2: Performance of EFPMA with SPADE and Prefixspan algorithms

| Dataset     | Quantity of Transactions in Database | Execution Time (Secs) | SPADE | Prefixspan | EFPMA |
|-------------|--------------------------------------|------------------------|-------|------------|-------|
| T10I4D100K  | 65536                                |                        | 8.9   | 6.5        | 2.7   |
| T40I10D100K | 65536                                |                        | 7.5   | 5.5        | 2.5   |
| Forestfires | 517                                  |                        | 8.9   | 6.5        | 2.8   |

### Table 2.3: Performance of ETMA with MAFIA and ECLAT strategies

| Dataset     | Quantity of Transactions in Database | Execution Time (Secs) | MAFIA | ECLAT | ETMA |
|-------------|--------------------------------------|-----------------------|-------|-------|------|
| Accident    | 65K                                  |                       | 210   | 110   | 100  |
| Retail      | 65K                                  |                       | 420   | 100   | 95   |
| Mushroom    | 8K                                   |                       | 593   | 150   | 145  |
| Pumsb_star  | 49K                                  |                       | 1056  | 200   | 180  |
| T10I4D100K  | 65K                                  |                       | 1113  | 250   | 220  |
| T40I10D100K | 65K                                  |                       | 1500  | 280   | 250  |
3. Results and Discussions

EFPMA, an algorithm grows the models from both ends (prefixes and suffixes) of detected patterns, which results in faster pattern growth because of less levels of database projection compared to existing approaches. If the Min_sup is large then the performance is similar to existing ones, but supports efficient pruning of invalid candidates. This is faster than SPADE and prefixspan algorithms towards performance and scalability but more verification work is needed.

The proposed ETMA approach can find the patterns that were repeated in a time-series database (Full & Subsection also). The three kinds of patterns such as character, series and section periodicity and how to decrease the duplicated data there by reducing the memory consumption and complexities through the removal method is also proposed. The performance efficiency is achieved through proposed two algorithms. Assess the running time; by changing the minimum support threshold for each dataset. The lower the minimum support threshold is, the larger the quantity of regular models is, and thus the more the running time is also consumed. The execution time compared on the two synthetic datasets. These synthetic datasets normally have too many distinct items. Therefore, although their transaction length is small on average, they typically have many transactions. On T10I4D100k, when the minimum support threshold varies from 1% to 7%, ETMA is 40%, 59% and 1.2 times faster than ECLAT, STNR and MAFIA algorithms. On T40I10D100k, when the minimum support threshold varies from 6% to 10%, EFPMA is 34%, 59% and 72% faster than SPADE, Prefixspan approaches.

A comparative study on the execution time of EFPMA, ETMA algorithms on synthetic and real datasets are as follows:

![Fig. 2: Traffic Accident Data set](image-url)
4. Conclusion and Future Work

From the above comparative study work, it is proved that EFPMA is faster than SPADE and prefixspan algorithms towards performance and scalability but more verification work is needed further. ETMA algorithm is experimented on real datasets accident and retail and the performance is better than the existing algorithms STNR, ECLAT and MAFIA towards execution time, quantity of records and the size of the maximal pattern. From the above two figures, it is observed that the ETMA, DBSCAN algorithms outperforms well towards execution time of results for various minimum support instances. These sequences can be applicable in any field particularly in earthquake prediction, weather forecasting and fraud detection applications. Detecting effective cyclic models in large information for dynamic Instances is useful in the huge traffic data. Predict the Travel information by using the best Periodical Pattern Detection algorithm.

References

[1] Agrawal R , Srikant R, “Fast Algorithms for Mining Association Rules”, Proceedings of the 20th VLDB Conference Santiago, Chile, pp. 1-32, 1994.
[2] Agrawal R , Imielinski T A, Swami S. “Mining Association rules between sets of items in large databases” Proceedings of the ACM SIGMOD Conference on Management of data, Washington, D.C., pp. 1-10, 1993.
[3] Houtsma M, Swami A. “Set-oriented mining of association rules”, Research Report IBM Almaden Research Center, San Jose, California, , pp. 1-10, 1993
[4] Agrawal R,Srikant R. “Mining Sequential patterns”, Proceedings of the International conference on Data Engineering (ICDE), USA, 1995.
[5] Roberto J, Bayardo B. “Efficiently Mining Long Patterns from Databases”, Proceedings of the ACM-SIGMOD International Conference on Management of Data, USA, pp. 85-93, 1998.
[6] Minos N G, Rajeev R , Shim K SPIRIT “Sequential pattern mining with regular expression constraints”, Proceedings of the 25th VLDB Conference, Edinburgh, Scotland, pp. 223-234, 1999.
[7] Pei J, Han H, Mao R “CLOSET: An efficient algorithm for mining Frequent closed itemsets”, Proceedings of the ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, Dallas, TX, Canada, pp. 1-10, 2000.
[8] Bastide Y, Rafik T, Nicolas P, Stumme G, Lakhal L. “Mining Frequent Patterns with Counting Inference”, Proceedings of ACM SIGKDD, USA, 2(2), pp. 66-75, 2000.

[9] Han J, Pei J, Yin Y, “Mining frequent patterns without candidate generation”, SIGMOD Rec., USA, 2000.

[10] Burdick D, Calimlim M, Gehrke J. “MAFIA: A Maximal Frequent Itemset Algorithm for Transactional Databases”, International conference on Data Engineering (ICDE), USA, pp. 1–10, 2001.

[11] Mohammed J, Ching-Jui Hsiao Z. “CHARM: An Efficient Algorithm for Closed Itemset Mining”, 2nd {SIAM} International Conference on Data Mining, pp.1-17, 2002.

[12] Mohammed J, Zaki Z. “SPADE: An Efficient Algorithm for Mining Frequent Sequences”, Machine Learning Journal, Taiwan, 42, pp. 31–60, 2001.

[13] Lin M Y, Lee S Y. “Fast Discovery of Sequential Patterns through Memory Indexing and Database Partitioning”, Journal of Information Science and Engineering, 21(1), pp. 109-128, 2002.

[14] Wang J, Han J, Pei J. "CLOSE+: Searching for the Best Strategies for Mining Frequent Closed Itemsets, Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining Washington, D.C, pp. 1-10, 2003.

[15] Nizar R, Mabroukeh C I, Ezeife E. “A Taxonomy of Sequential Pattern Mining Algorithms”, ACM Computing Surveys, 43(1), p. 1-4, 2010.

[16] Li Z, Ji M, Lee J G, Kays R. “MoveMine: Mining moving object databases”, Proceedings of the ACM SIGMOD International Conference on Management of Data, Indiana, USA, pp. 1-4, 2010.

[17] Mohan P, Shekhar S, James A, Shine S, James P, Rogers R. Cascading Spatio-Temporal Pattern Discovery, IEEE Transactions on Knowledge and Data Engineering, 24(11), pp. 1977-1992, 2012.

[18] O. Obulesu and Dr. A. Rama Mohan Reddy, “Fast and efficient frequent and maximal periodic pattern mining in Spatiotemporal Databases”, 6th IEEE International Advanced Computing Conference, Bhimavarm, A.P., INDIA, pp. 35-39, February, 2016.

[19] O.Obulesu and Dr.A.Rama Mohan Reddy, “Effective Sequential Pattern Mining in Fewer Levels of Recursion”, International Journal of Computer Science and Information Technology, 2(1), pp1-9, 2012.

[20] O. Obulesu and Dr. A. Rama Mohan Reddy, “Finding Maximal Periodic Patterns and Pruning Strategy in Spatiotemporal Databases”, International Journal of Advanced Research in Computer Science and Software Engineering, 2(4), pp. 423-426, 2012.

[21] O. Obulesu and Dr. A. Rama Mohan Reddy, “Finding Frequent and Maximal Periodic Patterns in Spatiotemporal Databases towards Big Data”, International Journal of Emerging Technology and Advanced Engineering, 4(1), pp. 266–275, 2014.

[22] O. Obulesu and Dr. A. Rama Mohan Reddy, “Finding Frequent and Maximal Periodic Patterns in Spatiotemporal Databases for Shifted Instances”, International Journal of Data Mining and Knowledge Engineering, 6 (5), pp. 224-232, 2014.

Ooruchintala Obulesu, received his B.Tech degree from Sri Venkateswara University, Tirupati in 2005, Masters in Computer Science from JNTU, Hyderabad in 2008 and pursuing Ph.D. in Computer Science and Engineering from JNTU University, Anantapuramu in 2009. His areas of interest are Spatial Data mining. He published more than 10 research papers in Data mining indexed in Scopus.
Dr. A. Rama Mohan Reddy was born in 1958, received his B.Tech degree from JNT University Anantapur in 1986, Masters in Computer Science and Engineering from NIT, Warangal in 1991 and Ph.D. in Computer Science and Engineering from Sri Venkateswara University, Tirupati in 2007. He is currently working as a Professor of Computer Science and Engineering, S V University College of Engineering, Tirupati, India. His research interests are Software Engineering, Software Architecture, Cloud Computing, Operating Systems and Data Mining.

Mahendra M received his B.Tech degree from SIT, Tumkur in 2008, Masters in Computer Networks from UVCE, Bangalore in 2010. He worked as a Lecturer in the department of CSE, NITC, Calicut. His areas of interest are Wireless Networks and Data Mining.