Research Journal of Applied Sciences 7 (7): 354-358, 2012
ISSN: 1815-932X

Asynchronous Pattern Mining for Mixed Sequences Using Ant Colony Optimization

[1]M. Parimala and [2]S. Sathiyabama
[1]Department of Computer Applications, M. Kumarasamy College of Engineering, Karur, Anna University, Tamilnadu, India
[2]Department of Computer Science, Thiruvalluvar Govt. Arts and Science College, Rasipuram, Periyar University, Tamilnadu, India

Abstract: In data mining, sequential pattern mining is a fundamental and essential field since of its broad possibility of applications spanning from forecasting the user shopping patterns and scientific discoveries. In many applications, the most vital part of mining problem is mining periodic patterns in time series databases. It can be visualized as a tool for forecasting and identifying the potential activities of time-series data. Previous studies have measured hash based asynchronous periodic patterns where uneven occurrences of datasets are sequenced. Still, asynchronous periodic pattern mining has expected less attention and in effective. Most researches are explained about an occurrence of long and closed sequence in an asynchronous pattern by a form of numerical derivatives. To improve the asynchronous pattern mining more effective in this study, researchers are going to present an Ant Colony Optimization technique to identify the pattern matching in a mixed sequences (long and closed sequences) raised in the asynchronous pattern mining. The Ant Colony Optimization algorithm (ACO) is an optimized technique for resolving computational problems which can be condensed to discovering good paths and identified the best pattern from the given database. Using ACO, the optimal pattern matching is made efficiently for mixed sequences (closed or long set of sequences) in the given database. Experimental evaluation will conduct with time series dense data sets in arff format in Weka tool and show the performance improvement of asynchronous pattern mining for mining mixed sequences in the given database in terms of running time, pattern matching factor, optimality rate and scalability.

Key words: Pattern mining, asynchronous pattern mining, long sequence, closed sequence, ant colony optimizations

INTRODUCTION

In the field of data mining, sequential pattern mining (Fig. 1) is a vital part due to its broad variety of applications. An illustration for such applications consists of examination of web access, stock markets development, customers shopping guides and so on. In a database, a set of sequences is presented. Every sequence has a set of elements. Each element consists of a list of items and a user-specified the lowest amount support threshold. The main job of sequential pattern mining is to determine all of the recurrent subsequences. The subsequences which occurred frequently in the set of sequences are not smaller than the lowest amount support. The main strategy for sequence mining increasing from this study is sequence development Candidate generation does not require for the development of sequence approach. It gradually elevates

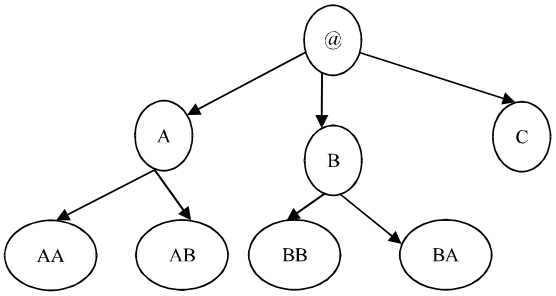

Fig. 1: Sequential pattern mining

the recurrent sequences. At first, it determines the common single items. Then, it generates a set of estimated databases for each frequent item, one database is assigned. After that, each databases in the frequent item is then mined recursively by integrating the repeated single items into a frequent sequence. These algorithms

Corresponding Author: M. Parimala, Department of Computer Applications, M. Kumarasamy College of Engineering, Karur, Anna University, Tamilnadu, India

implement well in databases comprising of small or short frequent sequences. However, even mining databases comprising of long frequent sequences, e.g., stocks values, machine observing the data, their overall performance aggravates by an order of magnitude.

To improve the efficiency of the process, a constraint integrating is the process of mining sequential patterns which avoid ineffective and remaining output. These constraints are completely included within the mining process with no post-processing step. The main problem of sequential pattern mining is pushing various constraints. They spot the prefix-monotone property as the general property of constraints for sequential pattern mining and give a structure (Prefix-growth) that includes these constraints into the mining process. Prefix-growth inclines the sequence growth approach for mining closed and long sequences.

In this study, researchers introduce ACO, an Ant Colony Optimization technique for mining mixed sequences in an asynchronous pattern mining. ACO is developed to categorize a class of domains as long sequences and closed sequences to present a high performance. ACO for asynchronous pattern mining comprises of two phases, long sequence and closed sequence which utilize an iterative procedure of candidate-generation pursued by frequency-testing. The long sequence creates the numerous sequences and imposes constraints between events within the long sequences. The closed sequence builds the sequence frequency which is not at all present in the asynchronous pattern mining. This partition permits the preamble of a novel pruning approach which minimizes the size of the search space significantly.

LITERATURE REVIEW

In data mining, pattern mining plays a vital role. The pattern mining are of various types, sequential pattern, frequent pattern, episode mining, inter transaction pattern, etc. In data engineering, the process of identifying the sequential knowledge (Chiang *et al*., 2009) is done with the help of introducing cyclic model analysis. To perform a pattern mining concepts for a multi interval of time (Hu *et al*., 2009), developed a technique for data engineering concepts in a given database. To improve the structural framework of pattern mining (Ahmed *et al*., 2009) presented a technique for high utility patterns using incremental database. It followed a tree structure format for mining patterns in a sequential format.

To monitor the frequent patterns arises in the sequential pattern mining, several heuristics have been followed. The heuristics used for identifying the frequent patterns are GSP (Chen, 2010), SPADE (Rasheed *et al*., 2011), PrefixSpan. The GSP algorithm (Chen, 2010) used the subsequences of frequent sequence in an antimode property. The SPADE discovers common sequences using the web search (Rasheed *et al*., 2011) and connection based strategy. In this strategy, the sequence pattern database is altered to a perpendicular format.

The PrefixSpan algorithm used a prototype method. It used probable databases for accomplishing this. Prefix is a probable sequential pattern mining. It examined prefix subsequences and plans the postfix subsequences into the databases (Esmaieli and Tarafdar, 2010). Determining attractive patterns in long sequences is an important area in data mining (Gouda and Hassaan, 2011). Most of the approaches described patterns as remarkable if they happen commonly adequate in a suitably cohesive form. The algorithm presented in (Ykhlef and ElGibreen, 2009) used hybrid evolutionary algorithm for mining long sequences which described dominancy of the sequences and applied for decreasing the inspection of the data set (Chen *et al*., 2011).

In order to mine the both long and closed sequence efficiently for pattern matching in an asynchronous pattern mining is done efficiently with the Ant Colony Optimization technique.

ASYNCHRONOUS PATTERN MINING FOR MINING MIXED SEQUENCES USING ANT COLONY OPTIMIZATION

The proposed APM (Asynchronous Pattern Mining) for mining the best mixed sequence pattern identification is searched by using Ant Colony Optimization (ACO). The ant colony optimization technique used here is used to identify the best pattern sequence from asynchronous pattern mining by matching the patterns occurred in the sequential pattern mining. The architecture diagram of APM using ACO is shown in Fig. 2.

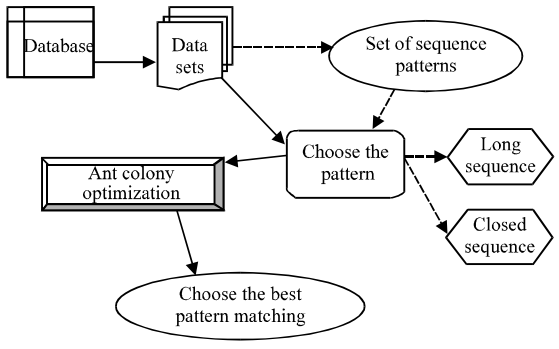

Fig. 2: Archite cture diagram of APM using ACO

A sequence is defined as a sorted list of event, where the order of the events in the sequence is the order of their occurrences. Researchers denote a sequence S as:

$$S=\{b_1, b_2, ..., b_k\}$$

Where:

b_i = An event

b_{i-1} = Happened before b_i

An event can occur once or multiple times in the same set of sequence. The proposed ACO for mining the best mixed sequence pattern identification is done by three operations. The first operation is to choose the pattern from a set of sequence. After choosing the pattern, match the pattern with the other set of patterns present in the sequence. Obtain a list of patterns which are closely related. Using ACO, choose the best pattern from a list of matched sequence pattern in a given database.

Pseudo code for mining long and closed sequence: Consider a set of sequences as S1, S2, ..., Sn and the set of events be e1, e2, ..., en in the asynchronous pattern mining. Given a frequent pattern α and it's minimal occurrence set MO (α). The event sequence S is scanned and all distinct single events are generated with their Minimal Occurrences (MO).

After identifying the frequent set of pattern sequences, the pattern matching is done with ACO technique by selecting the pattern to be matched. The pseudo code will describe about the ACO technique for mining long and closed set of sequences by pattern matching present in the asynchronous pattern mining. The process of ACO for pattern matching is shown in Fig. 3:

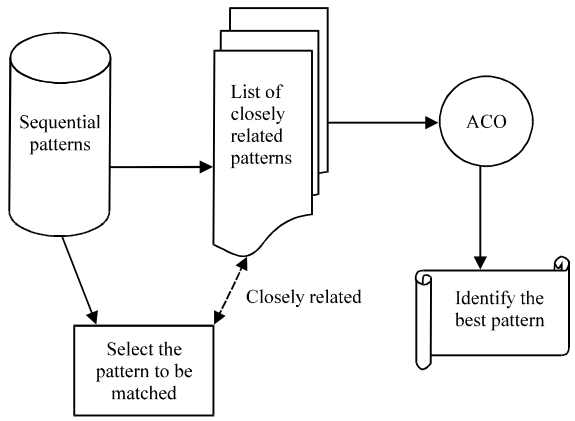

Fig 3: Process of ACO for pattern matching

Input event sequences

Output best sequence pattern:

- Step1: Let the sequences be Si where i = {1, 2,…, n} Events in the sequence Si be e1, e2, ..., en Ci-set of length -1 in sequences S
- Step 2: For each Si, do
- Step 2.1: Compute MO (Minimum Occurrence)
- Step 2.2 :Compute Sequence frequency (Sf)
- Step 2.2: End For
- Step 3: Based on Sf
- Step 3. 1: Identify the long sequence LS
- Step 3.2: End
- Step 4: Based on MO
- Step 4.1: Identify the Closed Sequence CS
- Step 4.2 :End
- Step 5: Maintain a set of LSi and CSi
- Step 6: Choose the pattern P to be matched either with patterns Pi of LSi or CSi
- Step 7: Use ACO: To find the best pattern matched from Si Step 8 Search the pattern P with a set of sequences LSi and CSi which are closely related
- Step 9: Match the pattern Pi to the set of pattern sequences Si
- Step 10: Identify the best pattern from a set of matched patterns
- Step 11: End

Through ACO, the pattern matching is efficiently done with the long sequence and the closed sequence patterns by selecting the best pattern by instructing the sequence of patterns in an asynchronous pattern mining. The ACO chose the pattern exactly matched with a set of sequences efficiently. For closed sequence pattern matching, the minimum occurrence and the sequence frequency are computed. Based on MO and SF, the closed sequence pattern matching is done. The closed pattern matching is done with respect to minimum occurrence. It first selects the patterns Pi which are closely related with the pattern to be matched. Then, from a set of matched patterns, it chooses the one which is closely related.

EXPERIMENTAL EVALUATION

The proposed APM using ACO for mining mixed sequences is implemented by using the Weka tool in the Java platform. The experiments were run on an Intel P-IV machine with 2 GB memory and 3 GHz dual processor CPU. The performance evaluation tests aimed at comparing the existing APM using mathematical derivates with challenging interactions through APM using ACO for mining long and closed sequence. The proposed APM using ACO for mining mixed sequence

infrastructure framework is depended on interception (Fig. 2). At set up it build the patterns in an asynchronous sequence pattern. Then, select the pattern to be best matched from a list of sequence patterns using ACO. An existing research done the pattern matching in an asynchronous sequence pattern using some mathematical derivations but the efficiency of the sequential patterns is less when compared to a pattern matching done with ACO. ACO efficiently chose the optimal pattern which is matched exactly with a set pf patterns. The proposed APM using ACO for mining mixed sequence framework carry three types of operations in general (selects the pattern, ACO technique, best pattern found from a list of closely related patterns). Operations can be assigned to different database available in the datasets.

The performance of the proposed APM using ACO for mining long and closed sequence infrastructure framework is evaluated by the following metrics:

- Running time
- Pattern matching factor
- Optimality rate and scalability

RESULTS AND DISCUSSION

In this research, researchers have seen that how the pattern matching is done efficiently with the ACO technique in an asynchronous pattern mining for mining mixed set of sequences. By using ACO, researchers have efficiently designed the proposed APM using ACO for mining long and closed sequence infrastructure framework for an optimal pattern matches. An independent test has been conducted with the datasets available in the time series dense datasets in the arff format.

The time series dataset has synchronous and asynchronous time stamp were sustained in main memory during the ACO process. The performance evaluation accomplished using datasets with different values of the record instances. To carry out the analysis over a large variety of diverse characteristics, researchers used real dense dataset of time series from Weka Tool Machine Learning Repository. The time series dataset records has the characteristics of a variety of time utilization factors and the number of records, items and the average record length are set to 1230, 49 and 11 correspondingly. It shows the scalability and reliability of the proposed APM using ACO for mining long and closed sequence infrastructure framework for an optimal pattern matches. The table and the graph describe the efficiency of the proposed APM using ACO for mining long and closed sequence.

Table 1: Record instances vs. running time

	Running time for mining long and closed sequence (sec)	
Record instances	Existing APM using mathematical model	Proposed APM using ACO
100	0.0030	0.0011
200	0.0032	0.0023
300	0.0045	0.0034
400	0.0052	0.0037
500	0.0070	0.0040

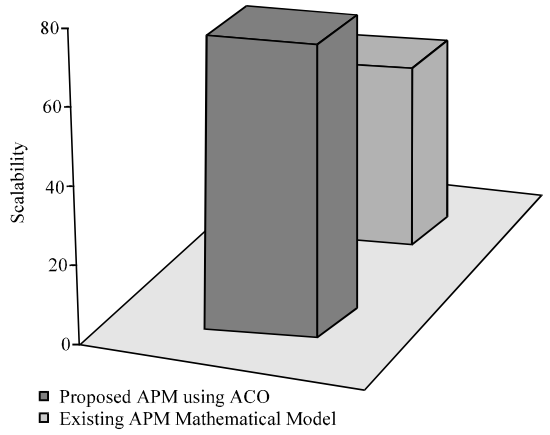

Fig. 4: Scalability factor

Table 1 shows the running time consumption for mining long and closed sequence in APM. Several time series datasets are used in the experimentation to validate the proposed APM using ACO for mining mixed sequence. Comparison result of the proposed APM using ACO for mining long and closed sequence with an existing APM using mathematical model based on running time consumption variance, measured in terms of seconds (sec). When number of record instances increases, the running time taken for mining long and closed sequence is less in the proposed APM using ACO for mining mixed sequence contrast to an existing APM using mathematical model. The performance graph of the proposed APM using ACO for mining long and closed sequence is shown in the Fig. 4. The variance in the running time consumption for mining long and closed sequence would be 12-20% low in the proposed APM using ACO for mining long and closed sequence.

Table 2 shows the pattern matching factor for mining long and closed sequence in APM using ACO. Several sequential patterns are selected from a list of APM which are closely related with each other to validate the proposed APM using ACO for mining mixed sequence. Comparison result of the proposed APM using ACO for pattern matching factor with an existing APM using mathematical model. When number of sequential patterns increases, the pattern matching factor is high in the proposed APM using ACO for mining long and

Table 2: No. of sequential record instances vs. pattern matching factor

	Pattern matching factor	
No. of sequential record instances	Proposed APM using ACO	Existing APM using mathematical model
100	0.07	0.10
200	0.10	0.15
300	0.12	0.20
400	0.18	0.25
500	0.20	0.30

Table 3: No. of matched record patterns vs. optimality rate

	Optimality rate (%)	
No. of matched record patterns	Proposed APM using ACO	Existing APM using mathematical model
25	10	8
50	15	10
75	16	12
100	20	15
125	18	14

closed sequence contrast to an existing APM using mathematical model. The variance in the pattern matching factor for mining long and closed sequence would be 15-20% high in the proposed APM using ACO for mining long and closed sequence.

Table 3 shows the optimality rate for pattern in mining long and closed sequence using ACO. The matched record patterns are selected from a list of APM which are closely related with each other to validate the proposed APM using ACO for mining mixed sequence. Comparison result of the proposed APM using ACO for optimality rate by selecting the best pattern from a list of closely related patterns with an existing APM using mathematical model. The variance in the optimality rate for selecting the best pattern from a list of patterns would be 10-15% high in the proposed APM using ACO for mining long and closed sequence in an optimality rate.

Figure 4 describes the scalability factor of the proposed APM using ACO with an existing APM using mathematical model. Since, ACO technique is used in the proposed, the scalability factor is high and it chose the best pattern from a list a sequence patterns in an asynchronous format.

Finally, it is observed that the proposed APM using ACO for mining mixed sequence in selecting the best pattern from a list of patterns outperforms well when compared to an existing APM using generic mathematical model in terms of running time, optimality and the scalability.

CONCLUSION

In this study for an asynchronous pattern mining; researchers have implemented an ACO technique for identifying the best pattern mining to fulfill the interoperability and scalability requirements of time series data sets from machine learning repository. The ACO technique identified the best pattern from a list of closely related patterns obtained from a list of closed and long sequence patterns. In addition the performance of the proposed APM using ACO for mining long and closed sequence framework is measured with metrics such as runnig time, pattern matching factor, optimality rate and scalability factor. Standard dense time series data sets are taken from Machine Learning repository to evaluate the performance evaluation of the proposed APM using ACO for mining long and closed sequence framework. The results showed that the proposed APM using ACO for mining long and closed sequence framework for pattern matching is 70% better in results compared to an existing APM using generic mathematical model.

REFERENCES

Ahmed, C.F., S.K. Tanbeer, B.S. Jeong and Y.K. Lee, 2009. Efficient tree structures for high utility pattern mining in incremental databases. IEEE Trans. Knowledge Data Eng., 21: 1708-1721.

Chen, J., 2010. An updown directed acyclic graph approach for sequential pattern mining. IEEE Trans. Knowledge Data Eng., 22: 913-928.

Chen, Y., R.F. Bie and C. Xu, 2011. A new approach for maximal frequent sequential patterns mining over data streams. Int. J. Digital Content Technol. Appl., 5: 104-112.

Chiang, D.A., C.T. Wang, S.P. Chen and C.C. Chen, 2009. The cyclic model analysis on sequential patterns. IEEE Trans. Knowledge Data Eng., 21: 1617-1628.

Esmaieli, M. and M. Tarafdar, 2010. Sequential pattern mining from multidimensional sequence data in parallel. Int. J. Comput. Theory Eng., 2: 730-733.

Gouda, K. and M. Hassaan, 2011. Mining sequential patterns in dense databases. Int. J. Database Manage. Syst., 3: 179-194.

Hu, Y.H., T.C.K. Huang, H.R. Yang and Y.L. Chen, 2009. On mining multi-time-interval sequential patterns. Data Knowledge Eng., 68: 1112-1127.

Rasheed, F., M. Alshalalfa and R. Alhajj, 2011. Efficient periodicity mining in time series databases using suffix trees. IEEE Trans. Knowledge Data Eng., 23: 79-94.

Ykhlef, M. and H. ElGibreen, 2009. Mining sequential patterns using hybrid evolutionary algorithm. World Acad. Sci. Eng. Technol., 60: 863-870.