International Journal of Soft Computing 9 (1): 44-50, 2014
ISSN: 1816-9503

Predominant Pattern Mining using ODIP Technique with Online Time Series Data

[1]B. Sujatha and [2]S. Chenthur Pandian
[1]Department of CSE, Sengunthar Engineering College, Tiruchengode, Tamil Nadu, India
[2]Mahalingam College of Engineering and Technology, Pollachi, Tamil Nadu, India

Abstract: Extracting predominant pattern in a time series database is a major data mining problem with several applications. The existing closed sequential patterns permit us to improve efficiency without bringing down the accuracy. The narrative technique developed a previous research follows a multiplex tree pruning technique which combines both the prefix and suffix tree patterns in an activity normalized time periodicity data sequences. The combinatorial point of prefix and suffix trees is on the threshold of predominant data pattern occurrence rate which efficiently identify the regularity of all observed patterns but still obtains the interlaced unwanted data. To separate the interlaced unwanted data from the predominant pattern mining, researchers are going to implement a new technique termed Optimized Discrete Interested Pattern technique (ODIP). This technique identifies the optimal value using the repetition occurrence in the pattern. An analytical and empirical result offers an efficient and effective predominant pattern mining framework for highly dynamic online time series data. Performance of the optimized discrete interested pattern technique is measured in terms of interlaced data removal efficiency, time taken for online pattern mining based on the frequency. Experiments are conducted with online time series data obtained from research repositories of both synthetic and real data sets.

Key words: Predominant pattern mining, discrete interested pattern, online time series data, interlaced unwanted data, optimal value, multiplex tree pruning

INTRODUCTION

Predominant pattern mining plays a significant role in data mining tasks. Various patterns are introduced such as frequent item sets, sequential patterns and recurrent event for dissimilar applications. Predominant patterns are recurring patterns that have sequential regularities in time-series databases. Predominant periodic patterns exist in many kinds of data. These mining have many emerging applications such as earthquake prediction, telecommunication network error analysis, duplicate detection in DNA sequences and occurrences of recurring illnesses.

By analyzing the behavior of trajectories on road networks, Lee *et al.* (2011) order the visited locations for improving classification accuracy. Based on the analysis, it compete that sequential patterns are good feature candidates since they preserve this order information. Furthermore, when pulling out sequential patterns, researchers propose to confine the length of sequential patterns to show higher efficiency. Compared with closed sequential patterns, these partial sequential patterns allow us to significantly improve efficiency almost without losing accuracy. The comparative study over a broad range of classification approaches demonstrates that the method significantly improves accuracy over other methods in some synthetic and real trajectory data.

The discovery of predominant patterns with periodicity has been considered only synchronous periodic patterns and did not recognize the misaligned presence of patterns due to the intervention of arbitrary noise. Officially, a valid subsequence with respect to a Predominant pattern P in a sequence S is a set of non overlie valid segments where a suitable segment has at least least_rep adjacent matches of P and the distance between any two consecutive valid segments does not exceed a parameter best_dis. A valid subsequence with the most overall repetitions of P is called its fastest suitable subsequence. Conversely, this model has some problems:

- Initially, these existing models are focused on mining periodic patterns in temporal sequences of events. Still in real-world applications, it might locate several events at one time slot in terms of diverse intervals
- Second, these existing models focused on mining the highest sequence of a pattern which can only incarcerate part of the system's behavior. Therefore, exposure the highest subsequence will neglect the other subsequence

Corresponding Author: B. Sujatha, Department of CSE, Sengunthar Engineering College, Tiruchengode, Tamil Nadu, India

- Third, in order to discover the longest subsequence, a longer segment can be busted into smaller segments when two segments have common characteristics
- The ultimate problem regarding the segment's end position affects two segments overlap. Instead of using the pattern's last occurrence as the segment's end position, the end position of a segment is distinct as the period's end for the last occurrence of the pattern

To address these problems, researchers presented an efficient periodic pattern mining framework for highly dynamic online time series data. It comprises of various unwanted data interlaced with predominant pattern. In this research, researchers discuss about the Optimized Discrete Interested Pattern technique (ODIP) which identifies the optimal value of the predominant patterns of interest. Interlaced unwanted data are separated from predominant pattern using ODIP technique. The parameters namely least_rept, univ_rept and best_dist are engaged to succeed efficient and effective patterns and removal of interlaced unwanted data. These data are viewed as a directory of suitable segments of perfect repetition interleaved by a disturbance.

Each suitable segment is necessary to be of maximum with the least_rept neighboring matches of the pattern and the space of each piece of disturbance is allowed only up to best_dist. A sequence is termed valid if and only if the overall repetitions of the pattern are greater than univ_rept. Researchers propose a sequence of algorithms for separating the interlaced unwanted data from the predominant mining.

Literature review: The occurrences of a tree pattern query in an XML database is a core operation in XML query processing. Prior research display that Holistic Twig Outline Matching algorithm is an efficient technique to respond an XML tree pattern with Parent-Child (P-C) and Ancestor-Descendant (A-D) relationships as it can successfully control the size of intermediate results during query processing. Lu *et al.* (2011) uses extended XML tree pattern which may include P-C, A-D relationships, negation functions, wildcards and order restriction.

Tsai *et al.* (2011) deal with the energy conservation issue in resource-constrained environments, the algorithm only transmits the local grouping results to the sink node for further ensembling. In the cluster ensembling phase, the algorithm combines the local grouping results to gain the group relationships from a global view. Researchers further influence the mining results to track moving objects proficiently. Al-Zyadat *et al.* (2011) address the issue of efficiently monitoring the satisfaction of sequential process to obtain on the way to data storage. UpDown Directed Acyclic Graph (UDDAG) is invented by Chen (2010) for competent sequential pattern mining. UDDAG allows bidirectional pattern growth along both ends of detected patterns.

IMine index structure can be competently exploited by different item set extraction algorithms. Baralis *et al.* (2009) methods presently support the FP-growth and LCM v.2 algorithms but they can straightforwardly carry the enforcement of various constraint categories. The IMine index has been integrated into the PostgreSQL DBMS and exploits its physical level access methods. Lahiri and Berger-Wolf (2008) propose a practical, competent and scalable algorithm to find such sub graphs that takes unsatisfactory periodicity into account. Researchers demonstrate the applicability of the approach on numerous real-world networks and eliminate meaningful and interesting periodic interaction patterns.

Engler (2008) proposes a universal model for discovery of periodic patterns within datasets associated to the manufacturing of electronic goods. Three general phases are considered. The discretization of the original dataset is primarily to be discussed followed by the clustering of the dataset into state related clusters and lastly the discovery of periodic patterns in the state transitions of the tests.

Most of the existing incremental rule mining methods are extremely dependent on accessibility of main memory. If sufficient amount of main memory is not presented, they not succeed to generate the results. Jadhav *et al.* (2012) presents a novel method for incremental discovery of frequent patterns using main memory database management system to eradicate this drawback.

A frequent pattern can be said periodic-frequent if it appears at a regular interval given by the user in the database. Tanbeer *et al.* (2009) introduce a novel notion of mining periodic-frequent patterns from transactional databases. Researchers use an efficient tree-based data structure, called Periodic-Frequent pattern tree (PF-tree) that captures the database filling in a highly compact manner and enables a pattern growth mining technique to produce the complete set of periodic-frequent patterns in a database for user-given periodicity and support thresholds.

Patil and Patil (2012) focus on data preprocessing stage of the first phase Data Cleaning algorithms. Data Cleaning algorithm eliminates inconsistent or unnecessary items in the analyzed data. Malathi and Baboo (2011) look at use of missing value and Clustering algorithm for crime

data using data mining. Researchers will look at MV algorithm and Apriori algorithm with some enhancements to aid in the process of filling the misplaced value and identification of crime patterns.

A spectrum occupancy prediction model based on Partial Periodic Pattern Mining (PPPM) is introduced by Huang *et al*. (2012). The mining aims to identify frequent spectrum occupancy patterns that are concealed in the spectrum usage of a channel. Verhein (2009) solves these challenges among others by mining sequences of Spatio-Temporal Association rules. Theoretical results are exploited in order to expand an efficient algorithm which is demonstrated to contain linear run time in the number of interesting sequences discovered. To remove the interlaced unwanted data, a new technique named Optimized Discrete Interested Pattern technique (ODIP) is presented.

MATERIALS AND METHODS

Proposed ODIP technique with online time series data: The proposed research is efficiently designed for removing the interlaced data. It presents an efficient periodic pattern mining framework for highly dynamic online time series data comprising various unwanted data interlaced with predominant pattern. Interlaced unwanted data are separated from predominant pattern with Optimized Discrete Interested Pattern (ODIP) technique which identifies the optimal value of the predominant patterns of interest or of repetition occurrence.

Online time series data is analyzed to remove significant information and other individuality of the data. Time series forecasting is the use of a model to forecast future values based on previously observed values. While regression analysis is frequently employed in such a way as to test theories that the current value of one time series affects the current value of another time series. The architecture diagram of the proposed Optimized Discrete Interested Pattern technique (ODIP) with online time series data is shown in Fig. 1.

In the Fig. 1 while designing the ODIP technique, initially the event data are accessed through the preprocessor module. The data preprocess parameters are used in the preprocess module to perform the process. Preprocessing is necessary because file include noisy and interlaced data which may affect result of mining process. Some of register file data are redundant for analysis process and could affect discovery of attack. Data preprocessing is a significant steps to filter and organize only suitable information before applying several mining algorithm.

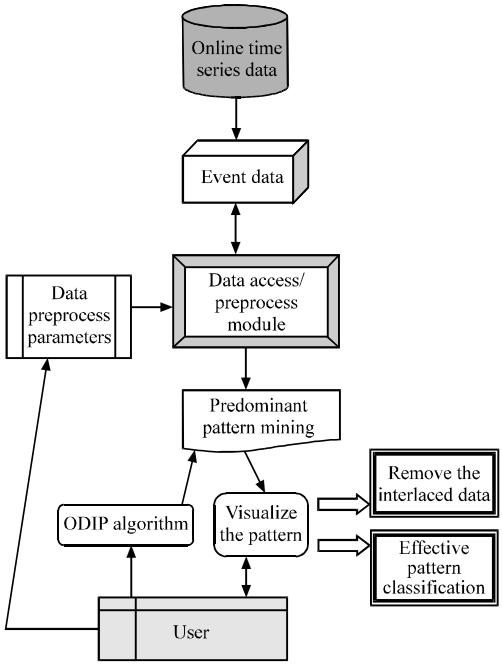

Fig. 1: Proposed Optimized Discrete Interested Pattern (ODIP) technique

Preprocessing reduce the file size and also increase quality of available data. The purpose of data preprocessing is to improve data quality and increase mining accuracy. The parameters used for the preprocessing consists of field extraction, data cleansing, user identification, session identification. These parameters are explained as:

Field extraction: The process of separating field from the single line of the file is known as field extraction. The server used different characters which work as separators. The ',' or 'space' character are most used separate characters.

Data cleansing: Data cleaning is usually site specific which involves extraneous references to embedded objects. By data cleaning, errors and contradiction will be detected and removed to improve the quality of data.

User identification: This parameter identify individual user who are using their IP address. If new IP address, there is new user. If IP address is same but browser version or operating system is different then it represents different user.

Session identification: Session defines as time duration between the log in and log out of the system. A

referrer-based method is used for identifying sessions. If IP address, browsers and operating systems are same, the referrer information should be taken.

Predominant pattern mining uses the Optimized Discrete Interested Pattern algorithm to remove the interlaced unwanted data. This mining uses to visualize the patterns and to effectively classify the patterns. The users effectively uses the preprocess parameters to visualize the patterns efficiently.

Optimized discrete interested pattern: OptimizedDiscrete Interested Pattern (ODIP) is used to specify all possible combinations. This study will demonstrate how ODIP specify can be used to find out suitable segments for multiple event singular patterns and complex patterns and also the combination of segments with respect to one pattern to form suitable sequences.

ODIP for multiple events: To discover suitable segments for multiple event 1-patterns, researchers propose two mining methods in ODIP technique. They are the Time-list Based Specification (TBS) and Sector-Based Specification (SBS). These two methods have their particular advantages and can be used in right situations.

Time-list Based Specification (TBS): For each episode 'e', researchers can specify possible event sets from events that have suitable segments with episode e. Duplicate specify are stay away from an alphabetic or numerical order on the events. For each combined event set, the time-list is obtained by the time-list intersection stage from the constituent events. The hash based procedures are used to check if suitable segments exist for the event set. Specification stops whenever no suitable segment exists for an event set.

Sector-Based Specification (SBS): Another way to discover multiple event 1-patterns is to combine suitable segments of single-event 1-patterns. Consider segments with the same episode. Two overlie segments with the same counteract can form 2-event singular patterns if the number of repetitions of the overlapped area is greater than least_rept. For efficient combination, segments of the same episode are ordered by their begin location. Two segments can be combined if they have the same counteract and the overlie area has repetitions greater than min rep. The overlie area is defined by the maximum begin location and the minimum ending location of the two segments. It is to be noted that the ending location of the segment can be determined by begin + (rept-1). Figure 2 shown below shows the tree structure for the online time series data.

Figure 2 contains three events namely W, Y and Z. Researchers can specify all combinations optimized

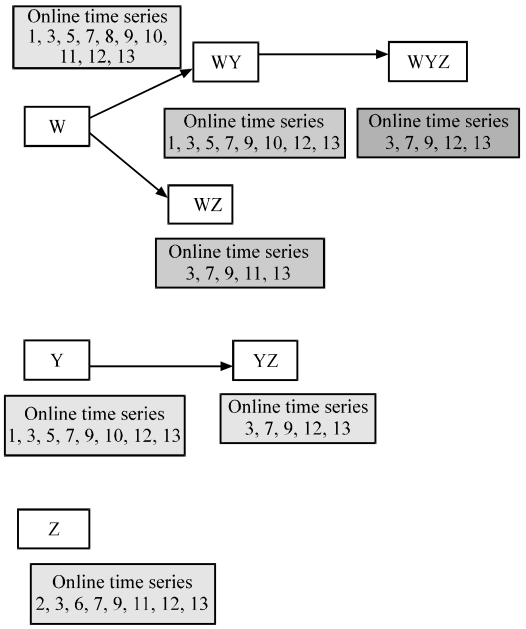

Fig. 2: Tree specification

discrete interested pattern. {W, Y} is initially specified and the time list is the intersection of W time list and Y time list. With the time list information, the hash procedure is used to check if suitable segments exist for this event set {W, Y}. Since, suitable segments exist for this event set {W, Y, Z} is then specified with time list being the intersection of WY time list and Z time list.

ODIP for complex events: Discovering complex patterns from singular patterns has a procedure similar to the Sector-Based Specification (SBS). It specifies possible combinations of suitable segments of the same episode in and checks if a combination forms a complex pattern. For two overlie segments with different counteract, they can form a 2-pattern if the replication of the overlie area is greater than least_rept. To discover a 'q'-pattern, researchers can compose it from a q-1 pattern with1-pattern. In other words, a q-pattern is composed of q segments of 1-patterns with different counteract. Two segments with the same counteract can only form a singular pattern. For an efficient combination, segments are ordered by their begin location.

Since, a pattern {W, *, X, Y} can also be represented by {X, Y, W, *} or {Y, W, *, X} which desirable to select one illustration to keep away from repetition. The idea is to select the one with the major repetitions. Therefore, the first element of the pattern is determined by the segment with the minimum ending location.

Algorithm for Optimized Discrete Interested Pattern (ODIP) Method: The steps to be performed is given below:

```
Procedure of ODIP (e, Seg_List_e with respect to episode e)
      For Each Segment Seg_i € Seg_List_e; do,
        Head node = Seg_i;
        Tail node = all Segment Seg_j € Seg_List with j>I;
        Node.Begin = Seg_i. Begin;
        Node.Ending = Seg_i. Begin + (Seg_i. rept-1) *e;
Sub_Procedure of ODIP (Head node, Tail node, Node, e)
   If (Head node == e) then return;
   For Each Seg_i € Tail node; do,
     Legal = True
      For Each Seg_i € Head node; do,
       If (Seg_i. Begin-Seg_j. Begin) %e == 0) Then
     Legal = False
  Break;
  If (legal == false) Then Continue;
    Inew.begin = Seg_i. Begin;
    Inew.Ending = Min {End node, e* Seg_i. rept};
     rept = (Inew.ending-Inew.begin)%e + 1;
  If (rept> = least_rept) Then
    Inew.Head = Head node U Seg_i;
    Inew.Tail = all Seg_k € Tail node with k>i;
  Else If (End node-Seg_i. Begin)<(e* least_rept);
  End If
  End For
  End
```

As shown in the algorithm, each procedure examines possible combinations of predominant pattern mining collected by beginning node with a section in the end node. For an achievable combination of the new segment must have a different counteract with each section in the head node. Two variables namely the start node and ending node are used to record begin and ending location for the current pattern. Therefore, the overlie area can be easily computed. Since, segments are prearranged by their begin location, the maximum begin location is determined by the new segment. Since, segments are ordered by their begin positions, once the overlie area is less than least_rept, the remaining segments can be ignored if the gap between ending node and Seg_i. start is less than e*least_rept.

Experimental evaluation: In this study, it provides the outcome of numerous experiments that have been processed using synthetic, real and ICU dataset in the UCI repository. The ICU data set consists of a sample of 200 subjects who were part of a much larger study on survival of patients following admission to an adult Intensive Care Unit (ICU). It account the outcome of trying different individuality of the proposed ODIP algorithm against other algorithms like Mining Discriminative Patterns for Classifying Trajectories (MDP-CT) and Predominant Periodic pattern mining using Multiplex Tree Pruning system (PPMTP).

The synthetic data have been produced in the similar way as done. The parameters forbidden through data creation are data allocation (uniform or normal), alphabet extent (number of exclusive signs in the data), dimension of the data (quantity of symbols in the data), time period size and the style and quantity of noise in the data. A datum might hold substitution, addition and removal noise or any combination of these kinds of noise with the interlaced data.

The ODIP algorithm detects the interlaced unwanted data and performs an efficient and effective predominant pattern mining framework for highly dynamic online time series data. Since, the confidence level of the predominant patterns increases, the noise level and interlaced data in the periodic data is diminished compared to an existing MDP-CT and PPMTP technique. The performance of the proposed Optimized Discrete Interested Pattern technique (ODIP) with various online time series data is measured in terms of:

- Interlaced data removal efficiency
- Time taken for the online pattern mining
- Repetition cycle of predominant patterns

RESULTS AND DISCUSSION

In this research, researchers have seen how the interlaced unwanted data are removed from the predominant pattern mining. Figure describe the performance of the proposed Optimized Discrete Interested Pattern technique (ODIP) in a varied online time series data. In this consequence, researchers compared ODIP against Mining Discriminative Patterns for Classifying Trajectories (MDP-CT) and Predominant Periodic pattern mining using Multiplex Tree Pruning System (PPMTP).

Removing efficiency: Removing efficiency is defined as the efficient way to remove the interlaced unwanted data from the optimized discrete interested pattern technique. It is measured in terms of percentage.

Figure 3 describes the efficient way of removing the interlaced data of the synthetic dataset. The online time series of the proposed Optimized Discrete Interested Pattern technique (ODIP) is compared with a Mining Discriminative Patterns for Classifying Trajectories (MDP-CT) and Predominant Periodic pattern mining using Multiplex Tree Pruning System (PPMTP).

Figure 3 describes the efficient way of removing the interlaced data with respect to the synthetic dataset. The set of experiments was used here to examine the impact of removing efficiency in the optimized discrete interested pattern algorithm. ODIP technique is capable to accomplish vastly precise results and its performance is normally reliable.

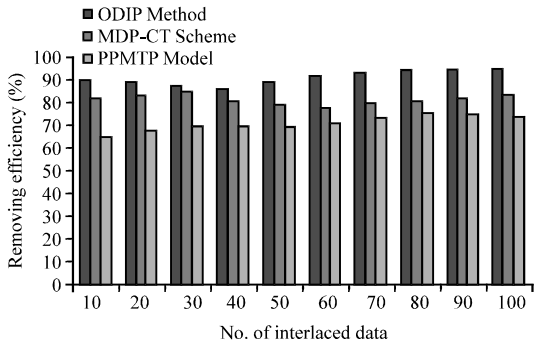

Fig. 3: Number of interlaced data vs. removing efficiency

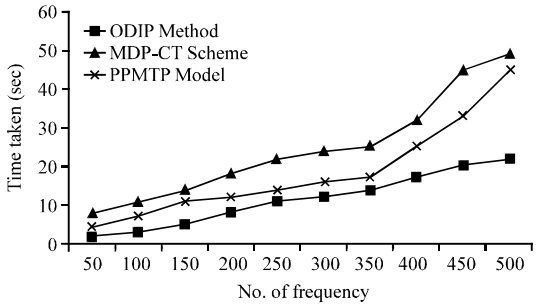

Fig. 4: Number of frequency vs. time taken

As researchers can see from Fig. 3, ODIP is more efficient by removing the interlaced data in online time series when compared with the MDP-CT and PPMTP algorithm. Experiments showed that the proposed ODIP algorithm efficiently identifies the interlaced data by performing the multiple events in a variety of situations. Compared to an existing research, the proposed ODIP technique achieves ~80-90% efficient removal of interlaced data.

Execution time: The time in which a single instruction is executed is called execution time of the online time series data. It makes up the last half of the instruction cycle. It is measured in terms of seconds.

Figure 4 describes the time taken for online pattern mining based on the frequency. The time taken to execute based on the online time series with respect to the real dataset. The execution time of the predominant pattern mining using the proposed Optimized Discrete Interested Pattern (ODIP) is compared with an existing Mining Discriminative Patterns for Classifying Trajectories (MDP-CT) and Predominant Periodic pattern mining using Multiplex Tree Pruning System (PPMTP).

Figure 4 describes the presence of time to execute based on the frequency of the predominant pattern mining with respect to the real dataset. As observed from the Fig. 4, ODIP exhibit set of experiments on data sets with

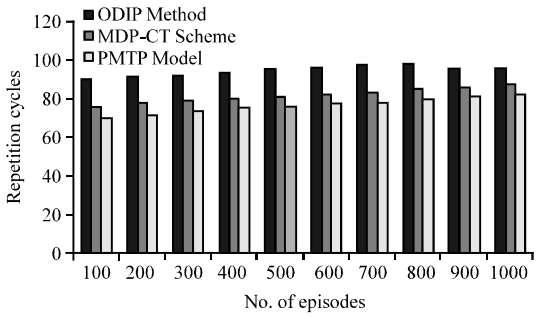

Fig. 5: Number of episodes vs. repetition cycle of predominant pattern

lesser execution time taken. In tricky cases, ODIP presents much improved results than the existing MDP-CT and PPMTP algorithm. The results stated in Fig. 4 recommend that the proposed ODIP is more interested to the proportion of data sets in execution time parameter.

The time consumption is measured in terms of seconds. Compared to the existing MDP-CT and PPMTP algorithm, the proposed ODIP consumes less time since it gives efficient classification of pattern result and the variance in time consumption is ~30-40% low in the proposed ODIP Method.

Repetition cycle of predominant pattern: It is defined as the sequence of characters to determine the round time-delay of an error detecting and feedback system. It is necessary to provide regular repetition of information in the ODIP technique.

Figure 5 describes the repetition cycle of predominant pattern with respect to the episodes. The episodes are used on the ICU dataset of the UCI repository for the proposed Optimized Discrete Interested Pattern (ODIP) technique is compared with an existing Mining Discriminative Patterns for Classifying Trajectories (MDP-CT) and Predominant Periodic pattern mining using Multiplex Tree Pruning System (PPMTP).

Figure 5 describes the repetition cycles accuracy based on the episodes with the help of the complex patterns of ODIP. This pattern forms the sophisticated pattern in the interesting directions. Compared to an existing (MDP-CT) and PPMTP, the proposed ODIP provides the efficient accuracy in the predominant pattern and the variance in is ~20-25% high in the proposed ODIP technique.

CONCLUSION

In this research, researchers efficiently remove the interlaced unwanted data using the Optimized Discrete Interested Pattern technique (ODIP). It processed using

synthetic, real and ICU dataset in the UCI repository. This technique identifies the optimal value using the repetition occurrence in the pattern. An analytical and empirical result offers an efficient and effective predominant pattern mining framework for highly dynamic online time series data. Performance of the optimized discrete interested pattern technique is measured in terms of interlaced data removal efficiency, time taken for online pattern mining based on the frequency. Experiments are conducted with online time series data obtained from research repositories of synthetic, ICU and real data sets. The experimental evaluations showed that ODIP algorithm performs well with ~80-90% efficient removal of interlaced data compared to Mining Discriminative Patterns for Classifying Trajectories (MDP-CT) and Predominant Periodic pattern mining using Multiplex Tree Pruning System (PPMTP).

REFERENCES

Al-Zyadat, W.J., R.B. Atan, H. Ibrahim and M.A.A. Murad, 2011. The Direct Impact to Pre-Filtering Process to Weather Dataset. J. Theor. Appl. Inf. Technol., 26: 1-6.

Baralis, E., T. Cerquitelli and S. Chiusano, 2009. IMine: Index support for item set mining. IEEE Trans. Knowl. Data Eng., 21: 493-506.

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

Engler, J., 2008. Mining periodic patterns in manufacturing test data. Proceedings of the IEEE Southeastcon, April 3-6, 2008, Huntsville, AL., USA., pp: 389-395.

Huang, P., C.J. Liu, L. Xiao and J. Chen, 2012. Wireless spectrum occupancy prediction based on partial periodic pattern mining. Proceedings of the IEEE 20th International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems, August 7-9, 2012, Washington, DC., USA., pp: 51-58.

Jadhav, J., L. Ragha and V. Katkar, 2012. Incremental frequent pattern mining. Int. J. Eng. Adv. Technol., 1: 223-228.

Lahiri, M. and T.Y. Berger-Wolf, 2008. Mining periodic behavior in dynamic social networks. Proceedings of the 8th IEEE International Conference on Data Mining, December 15-19, 2008, Pisa, Italy, pp: 373-382.

Lee, J.G., J. Han, X. Li and H. Cheng, 2011. Mining discriminative patterns for classifying trajectories on road networks. IEEE Trans. Knowledge Data Eng., 23: 713-726.

Lu, J., T.W. Ling, Z. Bao and C. Wang, 2011. Extended xml tree pattern matching: Theories and algorithms. IEEE Trans. Knowl. Data Eng., 23: 402-416.

Malathi, A. and S.S. Baboo, 2011. Evolving data mining algorithms on the prevailing crime trend: An intelligent crime prediction model. Int. J. Sci. Eng. Res., 2: 97-102.

Patil, P. and U. Patil, 2012. Preprocessing of web server log file for web mining. World J. Sci. Technol., 2: 14-18.

Tanbeer, S.K., C.F. Ahmed, B.S. Jeong and Y.K. Lee, 2009. Discovering Periodic-Frequent Patterns in Transactional Databases. In: Advances in Knowledge Discovery and Data Mining, Theeramunkong, T., B. Kijsirikul, N. Cercone and T.B. Ho (Eds.). Springer, Berlin, Germany, ISBN-13: 9783642013065, pp: 242-253.

Tsai, H.P., D.N. Yang and M.S. Chen, 2011. Mining group movement patterns for tracking moving objects efficiently. IEEE Trans. Knowledge Data Eng., 23: 266-281.

Verhein, F., 2009. Mining complex spatio-temporal sequence patterns. Proceedings of the 9th SIAM International Conference on Data Mining, April 30, 2009, Australia, pp: 605-616.