

Efficient Periodicity Mining using Circular Autocorrelation in Time Series Data

Y. B. Malode¹, D. B. Khadse², D. V. Jamthe³

¹ Asst. Professor, Information Technology Department, PBCOE, M.H., India

² Asst. Professor, Computer Science & Engineering Department, PBCOE, M.H., India

³ Asst. Professor, Computer Science & Engineering Department, PBCOE, M.H., India

Abstract – This paper focused on symbol, segment partial periodicity mining. Here, we proposed an algorithm that can detect periodic pattern through extracting a set of candidate periods featured in time series utilizing circular autocorrelation. The proposed algorithms are used to detect all periodicities in time series without any previous knowledge of nature of data. Moreover, the proposed algorithms are discovered the periodic patterns for conservative set periods. Experimental results show that the proposed algorithms are highly accurate with respect to the discovered periodicity rates and periodic patterns. Real-data experiments demonstrate the practicality of the discovered periodic patterns.

Key Words: Time Series Database, Symbol Periodicity, Segment or Full Cycle Periodicity, Partial periodicity.

1. INTRODUCTION

The periodicity mining in time series database plays important role in data mining task. It can be used as tool for forecasting and prediction of the future behavior of time series. The researchers proposed different algorithms for periodicity detection in time series databases. A time series database is a database that contains data over time e.g. weather data that contains several measure at different times per day. The pattern mining is an approach to detect different symbol patterns which consist of combination of symbols from input symbol set (length of pattern, $L \geq 1$).

The input symbol set is the set of symbols which can be used to symbolized entire time series. Consider, the set of transactions, $X = \{15, 10, 25, 41, 13, 44, 57, 60\}$; input symbol set, $\Sigma = \{a, b, c, d, e\}$; the total symbols in Σ are 5; interval width = $X_{\max} - X_{\min} / \text{Total symbols}$ then X is discretized into symbolized time series, $T = \{aabdee\}$ where symbol a : limit 10 -20, symbol b : limit 21-30, symbol c : limit 31-40, symbol d : limit 41-50, symbol e : limit 51-60.

Periodic patterns indicate repetitive occurrence of activity(s), event(s). The repetition count indicates periodicity of pattern or a symbol. The period is term which shows interval after which pattern is regularly occurred in

time series. Periodicity mining is analysis of time series data to detect recurring patterns. Other side of periodicity mining is the symbolization which needs more attention. The time series is mostly symbolized before it is analyzed. The basic idea behind the symbolization is to shorten and speed up the analysis. The analysis of time series without symbolization is tedious stuff and time consuming because periodicity mining is a concern with analysis of large volume of time series. In this paper, we focused on symbol, segment and partial periodicity mining which specify the behavior of time series.

• Symbol Periodicity

The time series (T) may have symbol periodicity if any symbol from input symbol set Σ is recurring with period P in time series T at most of the positions specified by $stPos + I \times P$ where $P = 1, \dots, \text{length}(T) - 1$; $stPos + I \times P \leq \text{length}(T)$; $I \geq 0$.

Consider, Symbolized time series (T) = {abcdbecbdbc}

Here, symbol b is repeated with regular interval 2 and starting position (stPos) is 2 and end position (endPos) is 11. As per periodicity theory, if $P = 2$, $stPos = 2$, $\text{length}(T) = 12$ then symbol should be repeated at positions 2, 4, 6, 8, 10, 12 but practically it is repeated at position 2, 4, 6, 9, 11. This example shows that any symbol or segment which is repeated at other position than expected position but it retains same interval (period) for almost all its actual positions then it shows symbol or segment periodicity.

• Segment Periodicity

The time series (T) may have segment periodicity if any segment which can be any combination of symbol from input symbol set Σ is recurring with period P in time series, where $P = 2, \dots, \text{length}(T)/2$.

Consider, Symbolized time series

$T = \{abcabdabecedabb\}$

Here, segment ab is recurring at positions 1, 4, 7, 13; $stPos = 1$, $endPos = 14$, $P = 3$. The expected periodicity for segment ab should be 5 but actual periodicity is less than 5. It shows imperfect segment periodicity.

• Partial periodicity

The partial periodicity which associates periodic behavior with only a subset of all time points. It is less restrictive as compare to segment periodicity. A time series said to have partial periodicity for a pattern X starting at position $stPos$, if $|X| \geq 1$ and the periodicity of X in T is either perfect or imperfect with high confidence i.e. X occurs at most of positions specified by $stPos + i \times p$ where p is the period and integer $i \geq 0$ take consecutive values starting at 0.

The sequence periodicity is also called as partial periodicity. The partial periodicity represents behavior of time series. The sequential pattern mining can be defined as extracting patterns that appear more frequently at certain threshold [5]. The partial periodicity specifies behavior of time series at some but not all points of time [9]. Let, symbol set $\Sigma = \{a, b, c, d, \dots\}$ and time series, T is a sequence of symbols.

- A pattern X with period p is a sequence of p symbols and in the partial pattern, $*$ represent irregularity of repetition of the symbol at that position. The partial pattern is combination of either symbols in Σ or $*$, where $*$ is used to introduce partial periodicity.
- The pattern X is called a i -pattern if exactly i positions in X are symbol from Σ . for instance, $S = (a, b, *)$ is a 2-pattern of period 3.

The most of research work discover partial periodic pattern for a user defined period length. In previous algorithms, period length is not known in advanced then it is impossible to detect the periodicity. Otherwise, we have to check the patterns for all possible period which is exhaustive. In case, we assumed that the period is known in advanced but there is possibility of missing of unsuspected periodicity.

• Confidence Measure

Confidence measure is the ratio of actual periodicity to expected periodicity for symbol or segment. Confidence Measure = Actual Periodicity / $[(|T| - stPos + 1) / p]$. Confidence measure is a key parameter which can be used as a threshold value to determine the periodicity. If threshold to confidence measure is user specified then any pattern or symbol is consider as periodic if and only if its confidence measure is greater or equal to user specified level.

In perfect periodicity, confidence measure is 1. But in real database it is rarely possible to have perfect periodicity. Confidence measure become a major factor in periodicity mining, if user wish to collect or consider only those periodicity which satisfy user specified confidence threshold.

The proposed paper addresses the problem of discovering partial periodicity. The rest of the paper is structured as follows: In section 2, the light on various research works on partial periodicity mining technique and briefly discussed on symbolization. In section 3, we discussed our approach to detect partial periodicity. In section IV, we focused on

experimental results on three years of temperature data. In section V, we concluded our approach.

2. RELATED WORK

In periodicity mining, main area of interest of research is to find symbol, sequence and segment periodicity in large volume of data. In 2011, Faraz Rasheed, Mohammed Alshalalfa, and Reda Alhaji [1] proposed an algorithm that can detect symbol, sequence, segment periodicity in single run. They used suffix tree as a underlying data structure. M. G. Elfeky, W.G. Aref, and A.K. Elmagarmid [2] proposed two algorithms to detect symbol and segment periodicity separately. They suggested Fast Fourier Transform to detect periodicity. This technique required $O(n \log n)$ computational time but it detected only two types of periodicity. In 2005, Elfeky, Walid G. Aref, Ahmed K. Elmagarmid [3] addressed the problem of periodicity detection in presence of noise by warping the time axis at various points to optimally remove noise. But the worst case complexity is $O(k \cdot n^2)$ where k is a maximum length of periodic pattern and n is length of analyzed portion of time series.

In 2009, Amruta Mahatre, Mridula Verma, Durga Toshniwal [4] the proposed a concept to preserve privacy by adding fake data to each transaction in pre-processing before it is subjected to data miner of sequential pattern mining. Here, they used PISA Algorithm but have emphasized on privacy of data rather than efficient data mining. In 1995, Agrawal and Shrikant [7] proposed Apriori mining technique for mining sequential pattern.

Mala Dutta and Anjana Kakoti Mahanta [11] proposed calendar based approach to detect periodicity and also shows relationship between periodicity across different levels of any hierarchical timestamp such as year/month/day, hour/minute/second. Calendar based periodicity extraction works on both continuous and discrete domain. It has $O(n \log n)$ time complexity for continuous domain and $O(n)$ for discrete domain where n is the number of intervals in which pattern occurs. In 2012, Dr. Ramachandra et al. [12] proposed constraint based periodicity mining algorithm in time series databases. This algorithm is applicable to detect symbol, sequence, segment periodicity in real time data. Here, authors were addressed problem by using FP-tree as underlying data structure. The algorithm has worst case complexity is $O(k \cdot N)$ where N is the length of input sequence and k is the length of periodic pattern.

3. OUR APPROACH

3.1 Symbolization

The time series database is a large volume of data, non-finite, noise interference forms. It is infeasible to analyze large data manually. The time series database should be symbolized in order to improve analysis that is complex. Symbolization

technique can be used to reduce the number of values for a given continuous attribute, by dividing the range of attribute into interval. The interval labels (input symbols) can then be used to replace actual data values. Here, we used following steps for symbolizing large volume of time series of data.

Step I: Input the number of intervals, K which represent how many symbols we are going to use to symbolize the time series dataset.

Step II: Sort the data elements. Determine minimum attribute value (R_{min}), maximum attribute value (R_{max}) in time series dataset.

Step III: Determine interval width, W

$$W = (R_{max} - R_{min}) / K$$

Step IV: Assigning Discrete symbol to each data value according to interval width.

In time series data, all data values which come under specific interval width represented by a unique but same symbol. Symbol will vary according to interval width.

The interval boundaries are specified as $R_{min} + Wi$, where $i = 1 \dots K-1$.

Example: Consider the sequence of events $X = \{65 \ 86 \ 74 \ 79 \ 85\}$ which has to be discretized for an interval range of 3. The interval width based on the above formula will yield a value of 7. Then the values from 65 to 72 will be assigned the label "a", and the values from 73 to 79 will be assigned the label "b" and the values from 80 to 86 will be assigned the label "c". The sequences of events are thus symbolized as $X = \{a \ c \ b \ b \ c\}$

3.2 Proposed Plan

Basically, our approach guide to detect periodic pattern without prior knowledge of nature of data. The algorithm based on basic filtering step by computing circular autocorrelation using fast fourier transform. The proposed system is based on the concept of circular autocorrelation. Correlation determines the degree of similarity between two signals. If the signals are identical, then the correlation coefficient is 1; if they are totally different, the correlation coefficient is 0. *Autocorrelation* refers to the correlation of a time series with its own past and future values. Autocorrelation is also sometimes called "*lagged correlation*" or "*serial correlation*", which refers to the correlation between members of a series of numbers arranged in time.

Autocorrelation is a method which is frequently used for the extraction of fundamental frequency, F_0 : if a copy of the signal is shifted in phase, the distance between correlation peaks is taken to be the fundamental period of the signal. In

statistics, the autocorrelation of a discrete time series or a process $X(t)$ is simply the correlation of the process against a time-shifted version of itself.

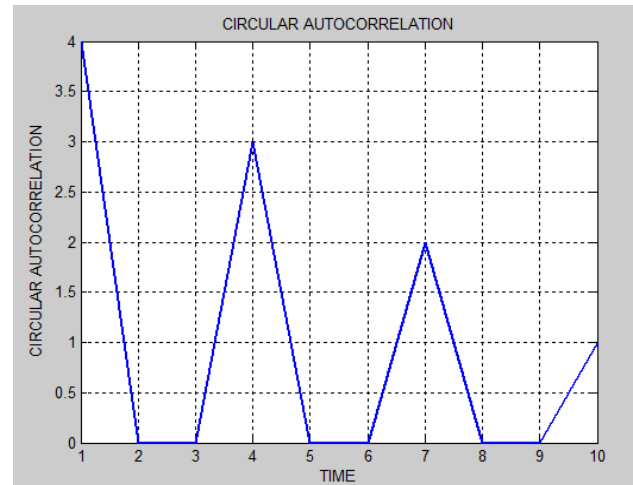


Fig 1: Circular autocorrelation graph for symbol 'a'

Consider, the time series $T = abcabdacba$, peak represent frequency of symbol at that time. The first non-zero autocorrelation values represent fundamental frequency of symbol in entire time series. The peak difference in consecutive peaks guide to compute period. In fig1, the symbol 'a' has a periodicity 4 and period is 3.

- **Symbol Periodicity Detection**

The algorithm for finding the Symbol periodicity takes the symbolized time series as input and retrieves the symbols that are periodic.

Input: A symbolized time series sequence $T = x_1, x_2, \dots, x_n$, of length n.

Output: Symbol Periodic Patterns, Number of Occurrences, Perfect and Imperfect Periodic Rates.

Algorithm:

1) For the time series T , create a binary matrix M of size $K \times n$ values, in which every row represents binary vector to a particular symbol, where K is the Interval range specified by user and n is the size of the time series. The existence of a symbol is denoted by "1" and the non existence by "0" in binary vector for each symbol.

2) Apply circular auto correlation to every row of the matrix M separately to find the circular auto correlation for every symbol using the formula,

$$r(k) = \frac{1}{N} \sum_{x=1}^N f(x) f(x+k)$$

3) Every non-zero element of the resulted sequence represents the total number of occurrences of that symbol from that position. The first non-zero element represents the total number of occurrence of that symbol.

4) The symbol that exceeds the minimum threshold percentage of occurrence is considered as a periodic symbol

5) The Index positions of the non-zero elements represent the starting position of the symbol pattern.

6) The Period are derived from the index positions ($PR_i = P_i - P_{i-1}$).

i) If the periodic rate of the symbol is the same in a minimum threshold percentage, it is considered as perfect periodic rate.

ii) If the periodic rate does not satisfy the above condition, it is considered as imperfect periodic rate.

• **Segment Periodicity Detection**

The algorithm for finding the segment periodic patterns also takes the symbolized time series together with the matrix.

Overview: The non zero elements of the Binary Matrix is auto correlated with the adjacent element of every other row until a zero value or end of the series is reached. The resulted sequence is searched in the rest of the series. If the entire time series can be represented as a repetition of the same sequence, then it is declared as a segment pattern. From this output, the number of occurrences, the index positions and periodic rates are derived.

Input: A symbolized time series sequence $T = x_1, x_2, \dots, x_n$, of length n and the Binary Matrix M with the symbols that are not eligible for candidate patterns removed based on the execution of the above algorithm.

Output: Segment Periodic Patterns, Number of Occurrences, Perfect and Imperfect Periodic Rates

Algorithm:

1) From the matrix M , remove the rows corresponding to the symbols that are not frequent.

2) Every non-zero element of the row is auto correlated with the adjacent element of every other row until a zero value or end of the series is reached.

3) The resulted sequence is searched in the rest of the series.

i) If found, it is declared as a valid sequence if it exceeds the minimum threshold percentage.

ii) If not found, the sequence is shrink and searched in the sequence until a two bit sequence is reached.

iii) If the entire time series can be represented as a repetition of the same sequence, then it is declared as a segment pattern.

4) The Index positions of the sequence represents the starting position of the sequence pattern.

5) The Periodic Rates are derived from the index positions ($PR_i = P_i - P_{i-1}$).

i) If the periodic rate of the symbol is the same in a minimum threshold percentage, it is considered as perfect periodic rate.

ii) If the periodic rate does not satisfy the above condition, it is considered as imperfect periodic rate.

• **Partial Periodicity Detection**

Overview: The algorithm for finding the partial periodicity takes the symbolized time series as input and retrieves the patterns that are partially periodic.

Input: A symbolized time series sequence

$$T = x_1, x_2, \dots, x_n \text{ of length } n.$$

Output: Periodic Patterns, Number of Occurrences, Perfect and Imperfect Periodic Rates.

Step 1: Scan the time series once and create binary vector of size N for every symbol in the alphabet of the time series, in which every row represents binary vector of size $[1 \times N]$ to a particular symbol, N is the size of the time series. The existence of a symbol is denoted by "1" and the non existence by "0" in binary vector for each symbol

Step2: For each symbol of the alphabet, compute circular autocorrelation vector over corresponding binary vector. This operation gives an output autocorrelation vector.

In *circular* autocorrelation, the point at the end of the series is shifted out of the product in every step and is moved to the beginning of the shifting vector. Hence in every step we compute the following dot product for all N points:

$$r(k) = \frac{1}{N} \sum_{x=1}^N f(x)f(x+k) \quad \text{where } k = 0,1,2,\dots,N-1$$

Step3: Scan only half of the autocorrelation vector because maximum possible period is $N/2$.

Step 4: The computed circular autocorrelation that provide conservative set of candidate period lengths for every letter in the alphabet set of our binary series.

Step 5: Detect partial periodicity for all candidate period using Han's technique.

4. EXPERIMENTAL RESULT

To detect partial periodicity there is need to scan entire time series. In the time series, segment repeated many times but sometimes not all portion of segment is repeated. Such not complete segment repetition is called the partial periodicity. every occurrence of segment will be a part of periodicity. Only that segment became a part of segment periodicity if length of segment and period are equal; segment occur at positions $stPos + Period * I \geq length(T)$; where $i=1,2,\dots$; T is a time series. In real time series data, confidence measure for segment periodicity is very less as compare to symbol periodicity. In periodicity mining stPos is a key factor that impact on the confidence measure.

- *Real Data*

For real data analysis, temperature data set are used for experimentation purpose. This dataset is one of the best set to search periodic pattern. Transactional dataset like Wal-Mart data etc are continuously used to detect periodicity by many researchers. Temperature data set comes with different patterns of temperature such as in rainy season temperature shows different patterns than the temperature pattern in winter season. Such patterns are periodic in dataset containing the yearly temperature data. Fig 2 represents graphical representation of three years of temperature data. For that temperature data which contain daily temperature are used. The record contains temperature data of 36 months.

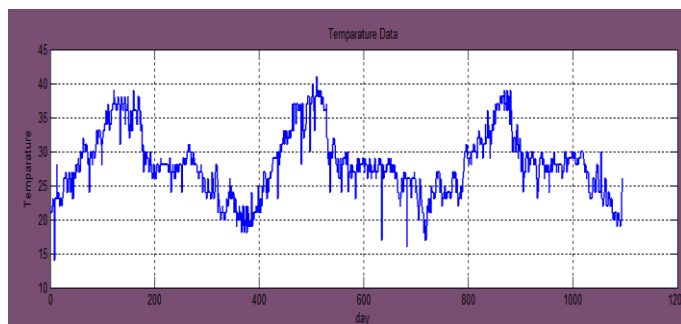


Fig 2: Temperature data of three years

If we provide less number of input symbols to discretized the time series data then periodicity (Confidence Measure) is high as compare to more input symbols. In table 1, for symbol set $\Sigma = \{a, b, c, d\}$, confidence measure for symbol a is 0.041971 but for $\Sigma = \{a, b, c, d, e, f\}$, confidence measure for symbol a is 0.041058.

Number of Symbol	Symbol	Confidence Measure	Period
4	a	0.041971	1
	b	0.413321	1
	c	0.423358	1
	d	0.120438	1
6	a	0.041058	5
	b	0.167883	1
	c	0.279197	1
	d	0.363139	1
	e	0.100365	1
	f	0.080292	1

Table1. Experimental Results for Symbol Periodicity for Different Symbol Set

In table 2, input symbol set, $\Sigma = \{a, b, c, d\}$, we detected segment periodicity for given input threshold. Here stPos & endPos represent starting index and ending position of segment in time series respectively. Confidence measure shows the periodicity count of segment, which is the important parameter to analyze the time series data. The segment periodicity concerns about repetition of patterns in entire time series. The confidence measure for pattern cc, dd for first symbol set are 0.3655, 0.0991 respectively. For symbol set of five symbols, confidence measure of patterns are changed to 0.4603, 0.2253.

Number of Symbol	Pattern	stPos	endPos	Conf. Measure
4	nnnnnnn*	1	893	0.5600
	nn***nn*	113	267	0.6000
	n*nnn*n*n	321	678	0.4500
	nnmnn*n*n	13	16	0.4800
	nn*nn*n*n	59	54	0.5400

Table2. Experimental Results for Segment Periodicity for Different Day Symbol Set with Different Threshold

The partial periodicity is a looser kind of periodicity. We tested our algorithm over different data sets. The most interesting data set we used i.e. temperature data.

Number Of Symbol	Thresh old	Pattern	stPos	endPos	Conf. Measure
4 $\Sigma=\{a, b, c, d\}$	0.3	ccc	52	1020	0.3158
		bb	1	1083	0.3403
		cc	49	1047	0.3655
	0.09	ccc	52	1020	0.3158
		bb	1	1083	0.3403
		cc	49	1047	0.3655
		dd	118	877	0.0991
		ddd	111	878	0.1116
		bbb	1	1082	0.2984
5 $\Sigma=\{a, b, c, d, e\}$	0.2	cccc	26	1046	0.3716
		cc	26	1062	0.4603
		dd	63	898	0.2253
	0.05	bbbbb	1	1087	0.1195
		cccc	26	1046	0.3716
		eeee	118	876	0.0868
		bb	1	1089	0.1788
		cc	26	1062	0.4603
		dd	63	898	0.2253

Table3. Experimental Results for Partial Periodicity

5. CONCLUSION

In this paper, we have presented algorithm that detect symbol, segment, and partial periodicity. Here circular autocorrelation is used to find the conservative candidate period. In this approach, there is no need of previous knowledge of nature of data. In previous approach, periodicities mined for user defined period length or assume the period. But these methods are lengthy and ambiguous .We used circular autocorrelation to detect periodicity. Using our approach, the periodic pattern detected for only the conservative period set. It shows the interest only portion of time series where the pattern is repeated. Our algorithm can be used as a filter to discover the candidate periods without any previous knowledge of the data along with an acceptable estimate of the confidence of a candidate periodicity. It is useful when dealing with data whose period is not known or when looking for unexpected periodicities. Algorithms such as Han’s described in [13] can be used to extract the patterns. We tried our method against various data sets and it proved to speed up linearly against different alphabets and different numbers of time points.

Interesting extension of our work would be the development of an algorithm to perform over other kinds of temporal data such as distributed and fuzzy. Finally, we intend to investigate the application of an algorithm or a function, other than the circular autocorrelation, that would require a smaller number of FFT computations. Results show that in

most of the cases, the performance of proposed algorithm is better than the other algorithms.

REFERENCES

- [1] Faraz Rasheed, Mohammed Alshalalfa, and Reda Alhadj, Associate Member, IEEE, “Efficient Periodicity Mining in Time Series Databases Using Suffix Trees”, IEEE Transaction Knowledge Data Engineering, Vol. 23, No. 1, January 2011.
- [2] M.G. Elfeky, W.G. Aref, and A.K. Elmagarmid, “Periodicity Detection in Time Series Databases”, IEEE Trans. Knowledge and Data Eng., vol. 17, no. 7, pp. 875-887, July 2005.
- [3] Mohamed G.Elfeky, Walid G. Aref, Ahmed K. Elmagarmid, “WARP: Time Warping for Periodicity Detection”, Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM’05), 2005.
- [4] Amruta Mahatre, Mridula Verma, Durga Toshniwal “Privacy Preserving Sequential Pattern In progressive databases using Noisy Data”,2009 13th International Conference Information Visualisation.
- [5] Kuo-Yu Huang and Chia-Hui Chang, Member, IEEE Computer Society, “SMCA: A General Model for Mining Asynchronous Periodic Patterns in Temporal Databases”, IEEE Transaction on Knowledge and Data Eng., vol. 17, no. 6, June 2005.
- [6] Anita Sant’Anna, Nicholas Wickström, “Symbolization of time-series: An evaluation of SAX, Persist, and ACA”, 2011 4th International Conf. on Image and Signal Processing.
- [7] R. Agrawal and R. Srikant, “Mining sequential patterns”, In Proc. 1995 Int. Conf. Data Engineering, pages3–14,Taipei,Taiwan, March 1995.
- [8] J. Lin, E. Keogh, L. Wei, and S. Lonardi, “Experiencing SAX: a novel symbolic representation of time series”, Data Mining and Knowledge Discovery, vol. 15, pp. 107–144, 2007.
- [9] Kuo-Yu Huang and Chia-Hui Chang, Member, IEEE Computer Society, “SMCA: A General Model for Mining Asynchronous Periodic Patterns in Temporal Databases”, IEEE Transaction Knowledge and Data Engineering, Vol. 17 No.6, June 2005
- [10] Jiong Yang, Wei Wang, and Philip S. Yu, Fellow, IEEE, “Mining Asynchronous Periodic Patterns in Time Series Data”, IEEE Transaction Knowledge and Data Engineering, Vol. 15, No. 3, May/June 2003

- [11] Mala Dutta¹ and Anjana Kakoti Mahanta” Detection of calendar based periodicities of interval-based temporal patterns”, International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.2, No.1, January 2012.
- [12] Dr. Ramachandra V. Pujeri, G .M. Karthik,” Constraint Based Periodicity Mining in Time Series Databases”, IJ. Computer Network and Information Security, 2012, 10, 37-46
- [13] J. Han, G. Dong, and Y. Yin. Efficient Mining of Partial Periodic Patterns in Time Series Databases. In Proc. of 1999 Int. Conf. on Data Engineering, Sydney, Australia, March 1999.

BIOGRAPHIES



Y.B.Malode received his B.E. degree in information technology from the Rashtrasant Tukadoji Maharaj University of Nagpur, India, in 2006. He has received his Master's Degree in Computer Science and Engineering from Rajiv Gandhi College of Engineering, Research and Technology, Chandrapur, Maharashtra. He is majoring in computer Science and is familiar with Data Mining. His research area includes Data Mining, Periodicity mining and Image processing.



D. B. Khadse received the B.E. Degree in Information Technology from Rashtrasant Tukadoji Maharaj University of Nagpur, India, in 2007. He has received Master of Engineering (M.E.) Degree in Wireless Communication and Computing from G. H. Raisoni College of Engineering, Nagpur, Maharashtra, India.



D. V. Jamthe received the B.E. Degree in Information Technology from Rashtrasant Tukadoji Maharaj University of Nagpur, India, in 2005. He has received Master of Engineering (M.E.) Degree in Wireless Communication and Computing from G. H. Raisoni College of Engineering, Nagpur, Maharashtra, India.