International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970)
Volume-2 Number-4 Issue-6 December-2012

Review on Periodicity Mining Techniques in Time Series Data

Yogesh Malode[1], Rahila Patel[2]
PG Student[1], Assistant Professor[2]
Department of Computer Technology, Rajiv Gandhi College of Engineering, Research and Technology
Chandrapur, Maharashtra, India

Abstract

The rapid growth in data and databases increased a need of powerful data mining technique that will guide to analyze, forecast and predict behaviour of events. Periodicity mining needs to give more attention as its increased need in real life applications. In this paper, we are going to discuss on various periodicity mining techniques in Time Series Databases as well as symbolization. Here, we propose a periodicity mining technique that will detect various periodic patterns (symbol periodicity, sequence or partial periodicity, segment or full cycle periodicity) in time series databases using Fast Fourier Transform.

Keywords

Time Series Database, Periodic Pattern, Periodicity Mining, Symbol Periodicity, Sequence or Partial Periodicity, Segment or Full Cycle Periodicity

1. Introduction

The explosive growth in data and databases has generated an urgent need for new techniques and tools that can intelligently and automatically transform the processed data into useful information and knowledge. The Data Mining techniques provide an intelligent solution by discovering implicit and meaningful knowledge which can be used for further development of various applications of real life such as marketing, stock market, supermarket etc. The data mining can be described as the process of discovering patterns or trends in data [4].

Nowadays, we have different types of databases which can be distinguish on the basis of stored data like Transactional data, Legacy data, Time Series data, Spatial data, Multimedia Data, Temporal Data, Spatiotemporal data, relational data etc. In this paper we present a review on the periodicity mining techniques in Time Series Data. Time Series Database is used to store sequence of events which is obtained over repeated measurement of time such as hourly measurements, daily measurements and weekly measurements. Time series is a collection of data values or events that have been occurred and gathered at uniform time interval. This collected data at uniform time interval create a sense of time series. Time series data have a natural temporal ordering. This makes time series analysis distinct from other common data analysis problems, in which there is no natural ordering of the observations. We have many real life examples of time series like spending pattern, power consumption in particular location, weather in hill stations, transactions in superstore, network delay etc. We can define Time series as repeating sequence of events collected with specified time interval. Traditional analysis tools such as symbolization, Fourier Transform improve the analysis process of large time series database.

The rest of the paper is structured as follows: In section 2, we outline various analysis tools of data mining and briefly discussed on symbolization and various symbolic representation techniques and presented some mathematical formulation associated with symbolization. In section 3, we shed light on time series analysis and discussed on symbol periodicity, sequence periodicity and segment periodicity. In section 4, we briefly discussed related work in periodicity mining. In section 5, we have proposed an algorithm to detect periodicity using FFT. We concluded our study in section 6.

2. Symbolization

The time series database is a large volume of data, non-finite, noise interference forms [7].It is infeasible to analyze large data manually. So automatic or semi automatic tools are used for data analysis. The time series database should be symbolized in order to improve analysis that is complex [6]. ESAX [7], SAX [8], PERSIST used for symbolization of time series. Symbolic Aggregate Approximation (SAX) is time series representation based on Aggregate Approximation Representation (PAA) can reduce time series length to symbolic string length.SAX reduce the dimensionality but to some extent, sub–interval point information may be lost. Extended

SAX (ESAX) used ESSVS (ESAX statistically Vector Space) to measure time series similarity and each sub interval is added two characters and represent great value and the minimum value of range. The symbolization methods are evaluate in terms of information loss and compression factor [7].

- **Information Loss**

The symbolization may causes loss of information. The information loss can be evaluated by Mean Absolute Error (MAE). It is the difference between original data and reconstructed data after symbolization.

- **Compression Factor**

The symbolized signal consists of three factors:
1: symbolic string, **Symb**, of size N,
2: A vector of size N containing duration of symbol, **Dur**
3: A matrix of size ZxP containing symbol values or template, **Temp**
The compression factor is calculated as:
CompFac = |BitsOrig−BitsComp|/BitsOrig
where:
BitsComp = BitsSymb + BitsDur + BitsTempl
BitsOrig = sizeof(double)∗ M= 64∗M
BitsSymb = ceil(log2(Z))∗N
BitsDur = ceil(log2(max(Dur))∗N
BitsTempl = sizeof(double) ∗ Z ∗P and
M is original signal length, ceil rounds upwards, Z is the number of symbol, N is length of the symbolic string and p= 1 for SAX and persist.

3. Time Series Analysis

The mining in time series data is an analysis of time series data in order to reveal the implicit facts and predict the behaviour of system. Periodicity mining concerns with analyzing of large volume of time series or temporal data which may contain frequent pattern of different types .Periodicity mining is a tool that helps in predicting the behaviour of time series data. Periodicity mining is used for predicting trends in time series data. Periodicity detection is an essential process in periodicity mining to discover potential periodicity rates. Periodicity mining is a process of searching of repeating or recurring pattern after regular period interval within time series.

The periodic pattern is the recurring pattern that have temporal regularities in time series databases[10].The previous research work has been done in periodicity mining in time series databases, can be categorized on the basis of type of periodicity or periodic pattern. In this paper, we review three types of periodicity in symbolized time series such as symbol periodicity, sequence periodicity, segment periodicity.

A. Symbol Periodicity

The symbol periodicity mining in time series database is the searching process for repetition of a symbol in time series databases. The time series database may have symbol periodicity if any symbol repeated with equal period interval within time series. Suppose, The Time Series (T) = a**b**cd**b**ca**b**ac**b**ac**b** , the symbol b repeated at position 1, 4, 7, 10, 13 respectively. The symbol b repeated with regular period interval (p = 3). The symbol periodicity for symbol b started at 1st position and ended at 13th position. But in real world data it is almost impossible exact repetition of symbol with regular period interval.

Pos	0	1	2	3	4	5	6	7	8	9	10	11	12	13
T	a	b	c	d	b	c	a	b	a	c	b	a	c	b

Fig 1: The symbol b repeated at position 1,4,7,10,13, started at position 1 in time series T represent symbol periodicity

B. Sequence Periodicity

The sequence periodicity is also called as partial periodicity. In partial periodicity, only portion of time series is main area of interest for detection of periodicity. The partial periodicity represents behaviour of time series in some portion of time series, not in complete time series. The sequential pattern mining can be defined as extracting patterns that appear more frequently at certain threshold [5]. The partial periodicity specifies behaviour of time series at some but not all points of time [9]. Let, symbol set ∑ = {a, b, c, d…….} and time series, T is a sequence of symbols.

- A pattern S with period **n** is a sequence of **n** symbols and the first symbol in pattern must be in ∑ and other symbol must be either a symbol in ∑ or *, where * is used to introduce partial periodicity. Here, * is used to represent that pattern must start with a non “*” symbol.
- The pattern S is called a i-pattern if exactly i positions in S are symbol from ∑.for instance, S = (a, b,*) is a 2-pattern of period 3.

C. Segment Periodicity

The segment periodicity is also called as full cycle periodicity. The time series database have segment periodicity if any segment or pattern repeated in entire time series with equal period interval. The full cycle periodicity specifies behaviour of the time series at all points in the period. Suppose, The time series (T) = abcab abcab abcab

In Fig 3, segment abcab repeated three times with period (P) equals to 5. The segment periodicity for pattern abcab started at 0th position and ended at 14th position. The segment detection focuses on entire time series (full cycle) for periodicity.

Pos	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
T	a	b	c	a	b	a	b	c	a	b	a	b	c	a	b

Fig 2: The segment recurring at position 0,5,10 started at position 0 represent segment periodicity

Here, we have discussed different periodicity types that may occur in entire time series but in real world time series data, the periodic pattern such as symbol, sequence or segment periodic pattern may repeat in only particular section of large time series. In such case, main focus should be on that subsection rather than entire time series. The sub-sectional periodicity detection may take more attention in future because it will used to detect unusual and repeated behaviour in subsection of time series.

We can represent time series on the basis of patterns in time series, starting position, ending position, period, confidence measure .These are the parameter which are used to represent periodicity .The confidence measures the perfectness of periodic pattern.

- **Perfect Periodicity**

The perfect periodicity occurs in time series T if starting of first occurrence of pattern S until the end of T every next occurrence of S exits period p positions away from current occurrence of S. It is also possible that some of the expected occurrences of S is missing, causes imperfect periodicity. The partial periodic pattern with perfect periodicity means that the pattern reoccurs in every cycle of period interval. The real life patterns are usually imperfect.

- **Confidence Measure**

The confidence of a pattern can be defined as actual occurrence count over expected occurrence count. It is the ratio of actual periodicity of a pattern to its expected perfect periodicity [1]. The confidence measure for pattern S with period P started at stPos position calculated as:

$$\mathrm{conf}(P, stPos, S) = \frac{\text{Actual Periodicity}(P, stPos, S)}{\text{Perfect Periodicity}(P, stPos, S)}$$

The actual-periodicity (P, stPos, S) calculated by counting number of occurrences of pattern S.

Perfect- Periodicity (P, stPos,S) = $[\frac{|T|-stPos+1}{p}]$

E.g. T= abbcaabcdbaccdbabbca, the pattern ab is periodic with period 5 and stPos is 0 then

conf (5,0,ab)= ¾

For perfect periodic pattern, the confidence is always 1.

4. Related Work

The lot of research has been done on periodicity mining for time series data in past years. The different algorithms were proposed to detect periodicity .The periodicity has been categorized on the basis type of pattern detected as well as wherein the pattern to be search (in the complete time series or the segment of time series). In periodicity mining, main area of interest of research is to find symbol, sequence and segment periodicity in large volume of data. In 2005, Elfeky et al.[12] addressed the problem of periodicity detection in presence of noise by warping the time axis at various points to optimally remove noise. In 2011, Faraz Rasheed et al.[1] proposed an algorithm that can detect symbol, sequence, segment periodicity in single run .They used suffix tree as a underlying data structure. But the worst case complexity is $O(k.n^2)$ where k is a maximum length of periodic pattern and n is length of analyzed portion of time series. In 2009,Amruta Mahatre et al.[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 [13] Algorithm but have emphasized on privacy of data rather than efficient data mining. In 1995, Agrawal and Shrikant [11] proposed Apriori mining technique for mining sequential pattern. In 2005, M. G. Elfeky et al. [2] proposed two algorithms to detect symbol and segment periodicity separately. They suggested Fast Fourier Transform to detect periodicity. This technique required O(nlogn) computational time but it detected only two types of periodicity.

In 2012, Mala Dutta and Anjana Kakoti Mahanta [15] 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(nlogn) 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. [16] 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.N) where N is the length of input sequence and k is the length of periodic pattern.

The above discussion clearly shows that there is a need to detect periodicity (symbol, sequence, segment periodicity) using Fast Fourier Transform which will enhance and improve time complexity.

5. Proposed Plan

In this paper, we have summarized and compared various periodicity mining techniques. On the basis of this study, we can propose an algorithm that can detect symbol, sequence, segment periodicity for time series databases. The following diagrammatic representation (Fig 3) shows proposed steps to detect periodicity using Fast Fourier Transform.

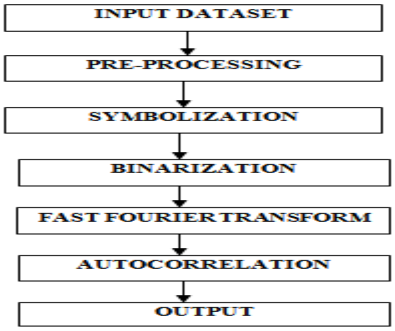

Fig 3: Periodicity detection

In proposed periodicity mining technique, we used cross correlation based approach on time series data. Cross Correlation is a mathematical tool for searching repeated patterns by analyzing degree of similarity between them in periodic signals. We proposed to compare original discretized input to shifted copy of itself, called Autocorrelation.

The method is based on FFT (Fast Fourier Transform) which has computational Complexity equals to *O* (*N log N*). The algorithm is described as follows:

1. Create a binary vector of size N for every symbol in the alphabet of the time series.
2. Compute DFT of each binary vector by using FFT algorithm
3. For each symbol of the alphabet, compute the circular autocorrelation function vector over the corresponding binary vector. This operation results in an output autocorrelation vector that contains frequency counts. Compute the dot product *r(k)* of all the N points.

   The dot product of all the N points is computed by,

   $$r(k) = \frac{1}{N}\sum_{x-1}^{N} f(x) f(x+k)$$
4. Scan only half the autocorrelation vector (maximum possible period is N/2) and filter out those values that do not satisfy the minimum confidence threshold and keep the rest as candidate periods.
5. Discover periodic patterns for the candidate periods produced in the previous step.

6. Conclusion

In this paper, we have discussed different algorithms that can detect periodicity in time series database.
Some authors proposed separate algorithm for different periodicity. Authors were proposed algorithms that can detect three types of periodicity using different data structures such as suffix trees, Fp-tree . We proposed an algorithm that can detect symbol, sequence, segment periodicity using Fast Fourier Transform which can enhance computational efficiency. Again we can approach to make it more generalized for different inputs.

References

[1] Faraz Rasheed, Mohammed Alshalalfa, and Reda Alhajj, 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, Senior Member, IEEE, and Ahmed K. Elmagarmid, Senior Member, IEEE, " Periodicity Detection in Time Series Databases", IEEE Transaction Knowledge and Data Engineering, , Vol. 17, No. 7, July 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 Wickstr¨om, "Symbolization of time-series: An evaluation of SAX, Persist, and ACA", 2011 4th International Conf. on Image and Signal Processing.

[7] Yi Jiang, Tuo Lan, Dongzhan Zhang, "A New Representation and Similarity Measure of Time Series on Data Mining", 2009 IEEE.

[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] R. Agrawal and R. Srikant, "Mining sequential patterns", In Proc. 1995 Int. Conf. Data Engineering, pages3–14,Taipei,Taiwan, March 1995.

[12] 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.

[13] W Huang, C. Y. Tseng, J.C Ou, and M. S. Chen, "A Genera Model for Sequential Pattern Mining with a Progressive Database", IEEE Transactions on Knowledge and Data Engineering, Vol. 20, No. 9,Sept 2008.

[14] Chong Zhu , Xiangli Zhang , Jingguo Sun , Bin Huang, " Algorithm for Mining Sequential Pattern in Time Series Data", 2009 International Conference on Communications and Mobile Computing.

[15] 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.

[16] Dr.Ramachandra V. Pujeri, G .M. Karthik," Constraint Based Periodicity Mining in Time Series Databases", I.J. Computer Network and Information Security, 2012, 10, 37-46.

Yogesh Malode was born in India on 28th Oct. 1984. He received his B.E. degree in information technology from the Rashtrasant Tukadoji Maharaj University of Nagpur, India, in 2006. He is at present pursuing his Master's Degree in Computer Science and Engineering from Rajiv Gandhi College of Engineering, Research and Technology, Chandrapur, Maharashtra under the supervision of Asst. Professor Rahila Patel. He is majoring in computer Science and is familiar with Data Mining. His research area includes Data Mining, Periodicity mining and Image processing.

Prof. Mrs. Rahila Patel was born on 06 May 1970. She has completed her M.Tech. from Nagpur University and is currently pursuing her Ph.D. from same University. She has over 14 years of teaching experience as Assistant Professor in RCERT, Chandrapur. Her research interest includes Data Mining, Genetic Algorithms and Optimization Techniques.