Available online at www.sciencedirect.com
ScienceDirect
Engineering Applications of Artificial Intelligence 21 (2008) 277–300
Engineering Applications of ARTIFICIAL INTELLIGENCE
www.elsevier.com/locate/engappai

Representing financial time series based on data point importance

Tak-chung Fu[a,b,*], Fu-lai Chung[a], Robert Luk[a], Chak-man Ng[b]

[a]Department of Computing, Hong Kong Polytechnic University, Hunghom, Kowloon, Hong Kong
[b]Department of Computing and Information Management, Hong Kong Institute of Vocational Education (Chai Wan), Chai Wan, Hong Kong

Received 12 September 2006; received in revised form 3 March 2007; accepted 27 April 2007
Available online 29 June 2007

Abstract

Recently, the increasing use of time series data has initiated various research and development attempts in the field of data and knowledge management. Time series data is characterized as large in data size, high dimensionality and update continuously. Moreover, the time series data is always considered as a whole instead of individual numerical fields. Indeed, a large set of time series data is from stock market. Stock time series has its own characteristics over other time series. Moreover, dimensionality reduction is an essential step before many time series analysis and mining tasks. For these reasons, research is prompted to augment existing technologies and build new representation to manage financial time series data. In this paper, financial time series is represented according to the importance of the data points. With the concept of data point importance, a tree data structure, which supports incremental updating, is proposed to represent the time series and an access method for retrieving the time series data point from the tree, which is according to their order of importance, is introduced. This technique is capable to present the time series in different levels of detail and facilitate multi-resolution dimensionality reduction of the time series data. In this paper, different data point importance evaluation methods, a new updating method and two dimensionality reduction approaches are proposed and evaluated by a series of experiments. Finally, the application of the proposed representation on mobile environment is demonstrated.

Keywords: Financial time series representation; Multi-resolution visualization; Incremental updating; Dimensionality reduction; Tree data structure; Mobile application

1. Introduction

Recently, the increasing use of temporal data, in particular time series data, has initiated various research and development attempts in the field of data and knowledge management (Last et al., 2001). A time series is a collection of observations made chronologically. The nature of time series data include: large in data size, high dimensionality and update continuously. Moreover, the time series data is always considered as a whole instead of individual numerical field.

There are varieties of time series data related research, for examples, finding similar time series (Liao et al., 2004), querying time series database (Rafiei and Mendelzon, 2000), segmentation (Wang and Willett, 2004; Feng et al., 2005), dimensionality reduction (Keogh et al., 2000; Keogh et al., 2001), clustering (Policker and Geva, 2000), classification (Wang and Willett, 2004) and forecasting (Pantazopoulos et al., 1998; Sfetsos and Siriopoulos, 2004, 2005). Those researches have been studied in considerable detail by both database and pattern recognition communities for different domains of time series data (Keogh and Kasetty, 2002). While most of the research communities have concentrated on the above issues, the fundamental problem on how to represent a time series in multi-resolution, which is also considered as information granulation as in (Bargiela and Pedrycz, 2003), has not yet been fully addressed so far. To represent a time series is essential because time series data is hard to manipulate in its original structure. Therefore, defining a more effective and efficient time series representation scheme is of fundamental importance.

The time series data used in data and knowledge management is high dimensional, but before it can be processed and analyzed, this dimensionality must be

*Corresponding author. Fax: +852 2774 0842.
E-mail address: cstcfu@comp.polyu.edu.hk (T.-c. Fu).

0952-1976/$ - see front matter
doi:10.1016/j.engappai.2007.04.009

reduced, commonly using approaches that focus on lower bounding the Euclidean distance. These approaches, however, smooth out salient points of the original time series, which is counterproductive when applied to financial time series data, as financial analysis often depends on the shape of data and the salience of data points to identify technical patterns. For these purposes, then it is important to reduce dimensionality while retaining the information associated with these points and the salient points are considered as important points to the shape of the time series.

Previous approaches to reducing dimensionality while retaining point information have included sampling. In this approach, a rate of m/x is used, where m is the length of time series P and x is the dimension after dimensionality reduction, but sampling approaches have the drawback of distorting the shape of sampled/compressed time series if the sampling rate is too low. As already noted, most other time series dimensionality reduction approaches, such as principal component analysis (PCA) (Fukunaga, 1990), singular value decomposition (SVD) (Korn, et al., 1997), discrete Fourier transform (DFT) (Agrawal et al., 1993; Rafiei and Mendelzon, 2000; Chu and Wong, 1999), discrete wavelet transform (DWT) (Popivanov and Miller, 2002; Kahveci and Singh, 2001; Chan and Fu, 1999), piecewise aggregate approximation (PAA) (Keogh et al., 2000; Yi and Faloutsos, 2000) and adaptive piecewise constant approximation (APCA) (Keogh et al., 2001), focus on lower bounding the Euclidean distance. However, because such approaches often lose important data points, they may fail to retain the general shape of the time series after compression (Fig. 1).

A time series is constructed by a sequence of data points and the amplitude of a data point has different extent of influence on the shape of the time series. That is, each data point has its own importance to the time series. A data point may contribute on the overall shape of the time series while another may only have little influence on the time series or may even be discarded. For example, frequently appearing technical time series patterns are typically characterized by a few salient points such as a head and shoulders. Time series pattern consists of a head point, two shoulder points and a pair of neck points. These points are perceptually important in the human visual identification process. These points are therefore more important than other data points in the time series. The data point with importance calculation is named as perceptually important point (PIP). The identification of PIP is first introduced by Chung et al. (2001) and used for pattern matching of technical (analysis) patterns in financial applications. The idea was later found similar to a technique proposed about 30 years ago for reducing the number of points required to represent a line by Douglas and Peucker (1973) (see also Hershberger and Snoeyink, 1992). We also found independent works by Perng et al. (2000), Pratt and Fink (2002) and Fink and Pratt (2003) which work on similar ideas. However, none of these techniques propose data structure to well-organize and store the salient points identified.

In this paper, we propose a time series representation framework which is based on the concept of data point importance. Challenges in here are like how to recognize these salient points, a data structure to represent these points which can facilitate incremental updating, multi-resolution retrieval and support dimensionality reduction. The proposed framework is capable to reduce the time series dimension to different levels of detail based on the importance of data point. On the other hand, the original accuracy can be maintained and salient points will not be distorted. A tree data structure, which stores the data points of the time series, is then proposed and efficient

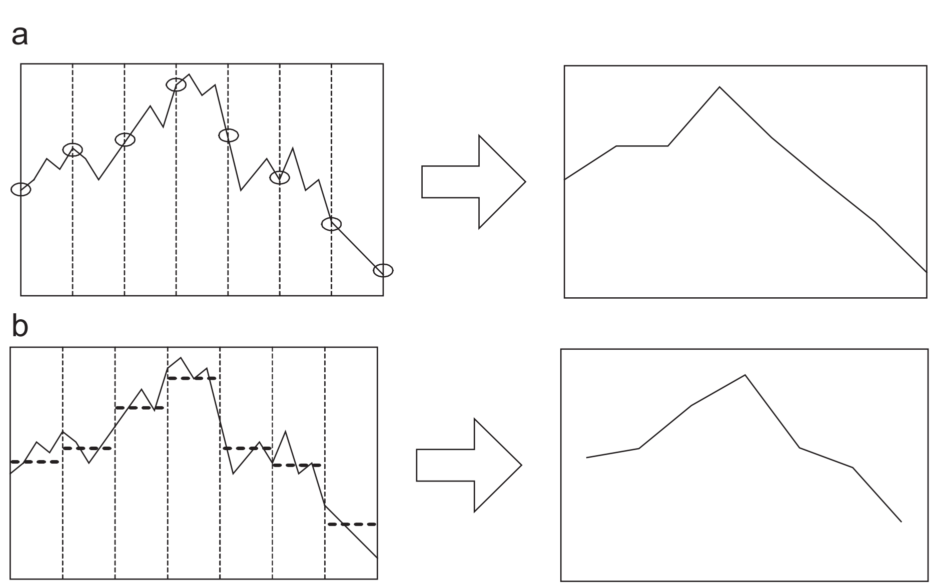

Fig. 1. Dimensionality reduction by (a) sampling and (b) PAA.