随机非平稳时间序列数据的相似性研究*

赵 慧[1+], 侯建荣[2], 施伯乐[1]

[1](复旦大学 计算机信息与技术系,上海 200433)

[2](上海交通大学 安泰管理学院,上海 200052)

Research on Similarity of Stochastic Non-Stationary Time Series Based on Wavelet-Fractal

ZHAO Hui[1+], HOU Jian-Rong[2], SHI Bai-Le[1]

[1](Department of Computer Information and Technology, Fudan University, Shanghai 200433, China)

[2](School of Aetna Management, Shanghai Jiaotong University, Shanghai 200052, China)

+ Corresponding author: Phn: +86-21-64077210, E-mail: zhaohui@fudan.edu.cn, http://www.fudan.edu.cn

Received 2003-07-01; Accepted 2003-10-08

Zhao H, Hou JR, Shi BL. Research on similarity of stochastic non-stationary time series based on wavelet-ractal. *Journal of Software*, 2004,15(5):633~640.
http://www.jos.org.cn/1000-9825/15/633.htm

Abstract: Traditional dimension reduction methods about similarity query introduce the smoothness to data series in some degree, but lead to the disappearance of the important features of time series about non-linearity and fractal. The matching method based on wavelet transformation measures the similarity by using the distance standard at some resolution level. But in the case of an unknown fractal dimension of non-stationary time series, the local error of similarity matching of series increases. The process of querying the similarity of curve figures will be affected to a certain degree. Stochastic non-stationary time series show the non-linear and fractal characters in the process of time-space kinetics evolution. The concept of series fractal time-varying dimension is presented. The original Fractal Brownian Motion model is reconstructed to be a stochastic process with local self-similarity. The Daubechies wavelet is used to deal with the local self-similarity process. An evaluation formula of the time-varying Hurst index is established. The algorithm of time-varying index is presented, and a new determinant standard of series similarity is also introduced. Similarity of the basic curve figures is queried and measured at some resolution ratio level, in the meantime, the fractal dimension in local similarity is matched. The effectiveness of the method is validated by means of the simulation example in the end.

* Supported by the National Natural Science Foundation of China under Grant Nos.69933010, 70371042 (国家自然科学基金); the China Post-Doctor Science Foundation under Grant No.2003033310 (中国博士后科学基金)

ZHAO Hui was born in 1968. She is a post-doctor at the Department of Computer Information and Technology, Fudan University. Her research areas are data mining, database and the applications based on mobile agent. **HOU Jian-Rong** was born in 1965. He is a post-doctor at the School of Aetna Management, Shanghai Jiaotong University. His research areas are mathematical optimization, fractal, and wavelet analysis, with the applications in Web data mining. **SHI Bai-Le** was born in 1935. He is a professor and doctoral supervisor at the Department of Computer Information and Technology, Fudan University. His main research areas are database theory, data mining, digital library technology and optimization of stochastic system.

Key words: non-stationary time series; similarity standard; local self-similarity; wavelet transformation; fractal varying-time dimension

摘 要: 传统相似性查询的维数约简方法导致时间序列的非线性和分形这些重要特征消失,基于小波变换的匹配方法是通过某一分辨级的距离标准来度量相似性.但是,在未知非平稳时间序列分形维数的情况下,序列相似性匹配的局部误差就会增大,曲线形状的相似性查询过程在一定程度上也因此受到影响.鉴于随机非平稳时间序列在时空动力学演化过程中呈现出非线性特征和分形特征,提出了序列分形时变维数的概念,原始分数布朗运动模型被加以改造成为一个具有局部自相似性的随机过程.给出了时变 Hurst 指数的估计式和算法,提出了一种新的序列相似性判别标准.在某一分辨级水平上进行曲线形状的相似性查询和度量,同时,对于局部相似性的局部维数曲线进行匹配.最后,用仿真算例对方法的有效性加以验证.

关键词: 非平稳时间序列;相似性标准;局部自相似性;小波变换;分形时变维数

中图法分类号: TP301 **文献标识码**: A

1 Introduction

The similarity query of time series has found important applications in the aspects such as determination of the similar sales pattern of products and discovery of the similar price behavior of stocks[1]. At present, the similarity pattern query about time series is a research hotspot in knowledge discovering in the time series database[2]. Because there are more sampling points of original time series and more series in the series database, the research emphasis is how to quicken the process of querying the similarity to solve the problem of realizing the best series pattern matching.

The common method in the research on similarity query is the dimensionality reduction technique (the dimension is defined as the number of time and space sampling which distinguishes from the fractal dimension in this paper). The representative research work includes: the *F-Index* method based on discrete *Fourier* transformation (DFT) to reduce the dimension which is presented by Rakesh[2], the Karhuen Loeve(K-L) transformation method by Wu[3], the linear division method presented by Keogh Eamonn section by section , in which the complex curve subsection is represented as the straight line[4,5]. The probability is used to query similarity after the series is represented again by the Keogh Eamonn method[4,5]. The up to date research about dimension reduction is the time series similarity matching based on wavelet transformation presented by Chan and Zheng[6,7]. They use Euclidean distance standard and L-shift Euclidean distance standard as the judgement standard of series similarity respectively. The details are eliminated from the curve in order to distill the basic shape of series curve. Ordinarily, the time series reflected from an object's evolution of time-space kinetics is of non-linearity and fractal character (ragged irregularity of series). There are two kinds of behavior pattern of time series similarity. One kind displays that similarity may be strictly self-similarity or have statistic features. Almost all time series are stochastic non-stationary time series in nature[8], similarity of which means similarity in statistic features; The other kind of similarities have the difference in hierarchy structure, In nature fractal has a nesting with finite layers. Only the series objects have fractal similarity features in the non-scaled region. Otherwise similarity or self-similarity will disappear in the case that the objects go over the non-scaled region. Aforementioned dimension reduction methods about similarity query such as F-Index method, K-L transformation method and the linear division method all introduce the smoothness to data series in some degree so that the important features of time series about non-linearity and fractal are destroyed. Thus the local error of similarity matching of series increases. The latter matching method based on wavelet transformation measures the similarity by using the distance standard at some resolution level. But in the case of an unknown the fractal dimension of non-stationary time series, the process of querying the similarity of curve figure will be affected. This method is of blindness to a certain degree. In addition, previous work on similarity of sequence data mainly considers finding global patterns[9,10]. In this paper we think that the similarity of random non-stationary time series data shows the local similarity of series much more and a time series may be similar to the local shape of other sophisticated series. The similarity of the basic curve figure is queried and measured at some resolution ratio level, while the fractal dimension in local similarity is matched. The fractal determinate dimension value educed by scale relation can not depict the space-time kinetics process of object evolution completely yet, except that it can reflect the self-similarity construction rule of static structure. The dimension is always set to be constant when the similarity of some nature phenomena is studied. Actually, the evolution of nature phenomena in one dimension time world may often lead to changeable similarity. The aforementioned thought of unchangeable dimension does not accord with the objective fact in the case of having local self-similarity process. We put forward a new concept—time-varying dimension function $D(t)$ in this paper in order to describe the phenomena of evolution along with time more sufficiently and completely.

In plane space, the Hurst index H and fractal dimension D of the Fractal Brownian Motion (FBM) model have the relation: $D=2-H$, and FBM has a wide use in describing the creation of physiognomy and the stock wave of capital market[11]. So FBM model is chosen as a breakthrough to study the problem of time-varying Hurst index, where $D(t)=2-H(t)$. At the same time we note that stochastic process of the FBM model and its correlative increment process are ordinarily non-stationary because the Hurst index in the stochastic process with local self-similarity is time-varying. Wavelet analysis has been proven to be a very efficient tool in dealing with non-stationary and self-similarity. Thus, wavelet transformation will play a leading role in the process of evaluating time-varying index.

The main work in this paper is organized as follows. Section 2 presents the local self-similarity stochastic process definition of non-linear time series based on statistic self-similarity. The original FBM model is rebuilt by introducing time-varying Hurst index to make it to be a stochastic process with local self-similarity. Section 3 utilizes Daubechies wavelet to transform the local self-similarity process and establishes an evaluation expression of the Hurst index by the least square method. Section 4 introduces the algorithm of gaining the time-varying Hurst index. Section 5 describes the determinant standards of similarity. In Section 6 the effectiveness of the method is validated by simulation examples.

2 The Mathematical Model of Non-Stationary Local Self-Similarity Stochastic Process of Time Series

A statistic self-similarity process may be approximately regarded as a stochastic process that is foreign to observation distance and keeps the same behavior sample orbit. Data produced from many science domains can be modeled by this procedure.

The following stochastic process $\underline{Y}(t)$ of an integral form can be regard as a generalized fractal Browian motion (GFBM) of FBM, which includes a time-varying fractal parameter $H(t)$.

$$Y(t)=\int_{-\infty}^{0}\left[(t-u)^{H(t)-1/2}-(-u)^{H(t)-1/2}\right]\mathrm{d}B(u)+\int_{0}^{t}(t-u)^{H(t)-1/2}\,\mathrm{d}B(u) \tag{1}$$

where the real number $t\ge 0$, $B(t)$ is the standard *Brownian* motion. $H(t)\in(0,1)$.

Let $Y(t)$ be a stochastic process with zero mean value. If its covariance $\Gamma_t(s_1,s_2)$ satisfies the following expression

$$\Gamma_t(s_1,s_2)-\Gamma_t(0,0)=-q(t)|s_1-s_2|^{2H(t)}\{1+o(1)\},\ (|s_1|+|s_2|\to 0) \tag{2}$$

where $q(t)\ge 0$, then $Y(t)$ is called a stochastic process with local self-similarity. For a given $|s_1-s_2|$, local self-relativity of the process $Y(t)$ will also wear off when $H(t)$ decreases from one to zero. Thus a rough sample orbit with a gradually increased error is appeared.

When a time-varying index is smooth, the covariance function $\Gamma_t(s_1,s_2)$ in *GFBM* model (1) satisfies Eq.(2). So $Y(t)$ presents a local self-similarity behavior.

3 Wavelet Evaluation of the Hurst Index

Let $\psi(x)$ be the mother wavelet. $WY(a,t)$ is the wavelet transformation about the self-similarity process $Y(t)$ at scale a and position t. Then

$$WY(a,t)=a^{-1/2}\int\psi\left(\tfrac{u-t}{a}\right)Y(u)\mathrm{d}u=a^{1/2}\int\psi(x)Y(t+ax)\mathrm{d}x\ .$$

It is educed by Eq.(2) and the above formula that

$$E\left(|WY(a,t)|^2\right)=a^{-1}\iint\psi\left(\tfrac{u-t}{a}\right)\psi\left(\tfrac{v-t}{a}\right)E[Y(u)Y(v)]\mathrm{d}u\mathrm{d}v\ =a\iint\psi(x)\psi(y)E[Y(t+ax)Y(t+ay)]\mathrm{d}x\mathrm{d}y$$

$$\sim a\iint\psi(x)\psi(y)\{\Gamma_t(0,0)-q(t)|ax-ay|^{2H(t)}\}\mathrm{d}x\mathrm{d}y\ =C_1a^{1+2H(t)}\quad(a\to 0)\tag{3}$$

where $C_1=-q(t)\iint|x-y|^{2H(t)}\psi(x)\psi(y)\mathrm{d}x\mathrm{d}y$.

Let $y_t(a)=\log|WY(a,t)|^2$

$$\varepsilon_{t(a)}=\log\left\{|WY(a,t)|^2\ {}^{1}\!\Big/_{E\left(|WY(a,t)|^2\right)}\right\}-E\left\{\log\left[|WY(a,t)|^2\ {}^{1}\!\Big/_{E\left(|WY(a,t)|^2\right)}\right]\right\}$$

Then

$$y_t(a)=\log\left[E\left(|WY(a,t)|^2\right)\right]+C_2+\varepsilon_t(a)\tag{4}$$

where $C_2=E\left\{\log\left[|WY(a,t)|^2/E\left(|WY(a,t)|^2\right)\right]\right\}$

A regression model can be gained by (1) and (2) when a is very small:

$$y_t(a)\approx(\log C_1+C_2)+[2H(t)+1]\log^{a}+\varepsilon_t(a)\tag{5}$$

A small-scaled series is constructed as follows:

$$a_1>a_2>\ldots>a_L,a_j=2^{-j},j=1,2,\ldots,n\ .$$

Let $x_j=\log a_j$, $y_j=y_t(a_j)$, $j=1,2,\ldots,n$. The least square method is used to get an evaluator of $H(t)$ in Eq.(5) for the couples $\{(x_j,y_j), j=1,2,\ldots,n\}$:

$$\hat{H}(t)=\tfrac{1}{2}\left[\frac{\sum(x_j-\bar{x})(y_j-\bar{y})}{\sum(x_j-\bar{x})^2}-1\right]\tag{6}$$

where $\bar{x}={}^{\sum x_j}\!\big/_{n}$, $\bar{y}={}^{\sum y_j}\!\big/_{n}$

It can be proved that $\hat{H}(t)$ is a consistent result[12] of $H(t)$.

4 Algorithm Description and Analysis

Now let us observe a stochastic time series process $Y(t)$ on the discrete and equally spaced points.

The time points may be limited in [0,1]. The sample size is 2^J, $t_i=(i-1)/n$, $n=1,2,\ldots,2^J$. $y_{j,k}(k=0,1,\ldots,2^{j-1}$, $j=0,1,\ldots J-1)$ is an evaluated value of $WY\left(2^{-j},K2^{-j}\right)$. The latter is the discrete value by wavelet transformation $WY(a,t)$ in $a=2^{-j}$, $t=K2^{-j}$. Wavelet transformation is carried on by Daubechies' compactly-supported wavelet bases with M moments. Daubechies wavelet function was constructed by an American mathematician called Inrid Daubechies[13].

Step 1. [0,1) is partitioned into 2^l equal-length sub-sections I_m without interacting each other.

$$I_m=[(m-1)2^{-j},m2^{-j});1\le l\le (J-1),m=1,2,\ldots,2^l.$$

Step 2. $\hat{H}(t)$ is regarded as the average value of $H(t)$ in the corresponding sub-sections I_m. The appropriate time spot of $\hat{H}(t)$ is chosen at the point $2^{-l-1}(2m-1)$ in the middle of I_m.

The double variable set is defined as follows:

$$\{(X_m,Y_m)\}=\left\{\left[\log\left(2^{-j}\right),\log\left(\left|y_{j,k}\right|^2\right)\right]\middle| k2^{-j}\in I_m\right\}\ 0\le k\le 2^j-1,0<j\le J-1 \tag{7}$$

$\hat{H}(t)$ is evaluated by formula (6) on each I_m.

Step 3. The evaluated value of $\hat{H}(t)$ is smoothed by using local multinomial to form a curve that can be regarded as the approach of the real figure of $H(t)$.

Let $N=2^J$, $M=2^l$. The series is dealt with by wavelet transformation firstly. Reference [7] has proven that the worst time complexity of the fast wavelet transformation algorithm is $T_w=O(N)$. The worst time complexity of Step1 is $T_1=O(M)$. The worst time complexity of Step2 is $T_2=O(M^2)$. The interpolation point number of Lagrange's interpolation polynomial of Step3 is M. The time complexity of Step3 is $T_3=O(M^2)$. So the total time complexity of the algorithm is $T_w+T_1+T_2+T_3=O(N)+O(M)+O(M^2)+O(M^2)=O(N+M^2)$.

5 Determinant Standards of Similarity

Definition 1. Given two thresholds $\varepsilon_i(i=1,2)$ and two time series $\vec{X}=\{x_i\}_{i=0,1,\ldots,n}$ and $\vec{Y}=\{y_i\}_{i=0,1,\ldots,n}$ with the same length n whose fractal dimension functions are $D_1(t)$ and $D_2(t)$ respectively, where $D_i(t)\in C[a,b], i=1,2$. When the following two inequalities are satisfied at the same time the two series $\vec{X}$ and $\vec{Y}$ are similar.

$$d_1(\overrightarrow{X},\overrightarrow{Y})=(\sum_{i=0}^{n-1}(y_i-x_i)^2)^{\frac{1}{2}}\le\varepsilon_1 \tag{8}$$

$$d_2(D_1(t),D_2(t))=\underset{a\le t\le b}{\mathrm{Max}}\left|D_1(t)-D_2(t)\right|\le\varepsilon_2 \tag{9}$$

where $d_1(\overrightarrow{X},\overrightarrow{Y})$ is the Euclidean distance, and $d_2(D_1(t),D_2(t))$ is the measurement of function $D_1(t)$ and $D_2(t)$.

Definition 2. Given two thresholds $\varepsilon_i(i=1,2)$ and two time series $\overrightarrow{X}=\{x_i\}_{i=0,1,\ldots,n}$ and $\overrightarrow{Y}=\{y_i\}_{i=0,1,\ldots,n}$ with the same length n whose fractal dimension functions are $D_1(t)$ and $D_2(t)$ respectively where $D_i(t)\in C[a,b],i=1,2$. When the following two inequalities are satisfied at the same time, series $\overrightarrow{X}$ and $\overrightarrow{Y}$ are L-shift similar.

$$d_{L_1}(\overrightarrow{X},\overrightarrow{Y})=(\sum_{i=0}^{n-1}((y_i-x_i)-(y_A-x_A))^2)^{\frac{1}{2}}\le\varepsilon_1 \tag{10}$$

$$d_{L_2}(D_1(t),D_2(t))=\underset{a\le t\le b}{\mathrm{Max}}\left|(D_1(t)-\overline{D}_1)-(D_2(t)-\overline{D}_2)\right|\le\varepsilon_2$$

$$\overline{D}_1=\frac{1}{b-a}\int_a^b D_1(t)\mathrm{d}t\ ,\ \overline{D}_2=\frac{1}{b-a}\int_a^b D_2(t)\mathrm{d}t \tag{11}$$

x_A and y_A are the average values of series $\overrightarrow{X}$ and $\overrightarrow{Y}$ respectively. $\overline{D}_1$ and $\overline{D}_2$ are the average values of $D_1(t)$ and $D_2(t)$.

Lemma. Given two time series $\overrightarrow{X}$ and $\overrightarrow{Y}$ with the same length n. The two new series $\overrightarrow{S}_J$ and $\overrightarrow{T}_J$ are obtained after $\overrightarrow{X}$ and $\overrightarrow{Y}$ are transformed respectively by wavelet in the layer J, then

$$d(\overrightarrow{S}_J,\overrightarrow{T_J})\le d(\overrightarrow{X},\overrightarrow{Y}) \tag{12}$$

6 Simulation Results

In order to evaluate the effectiveness of the proposed method in searching the similarity of any two stochastic non-stationary time series, we chose two time series samples from HSI in stock market. The result reported in this section address the following issues:

- The introduction of time-varying fractal dimension (or time-varying Hurst index) can depict the non-linear irregularities of stochastic non-stationary time series.
- The new standard of similarities proposed in the above section meets the need of similarity of the basic shape, while it takes into account the similarity of fractal feature curve of the non-stationary time series data.
- Local similarity between one data series and another one.

Here are the two original non-stationary time series samples in the following figures (Fig.1 and Fig.2), and they indicate the change of HSI in the two different periods.

Figures 3 and 4 show the two series data curves produced by Fig.1 and Fig.2 after Daubechies discrete wavelet transformation. Here we adopt the wavelet base **db4**, and the two original data series are decomposed and synthesized at the fourth layer respectively. Obviously the series in Fig.3 is similar to the sub-series of series in Fig.3 when $\varepsilon_1=0.04$.

Figures 5 and 6 represent the Hurst index curves from Figs.1 and 2 respectively. Evolution of time-varying Hurst index is of great importance in stock investment strategies. Fifty points are selected in Fig.1 and treated with wavelet base **db4**, *J*=8. Hurst index discrete values are calculated by Eq.(6) and smoothed by a polynomial with order eight. So does Fig.2. Figure 7 is a segment cut from Fig.5. Hurst index curve in Fig.7 is similar to that in Fig.6 when $\varepsilon_2 = 0.02$, that is, the fractal character of the data from Fig.2 approaches to that from the local sub-series of Fig.1.

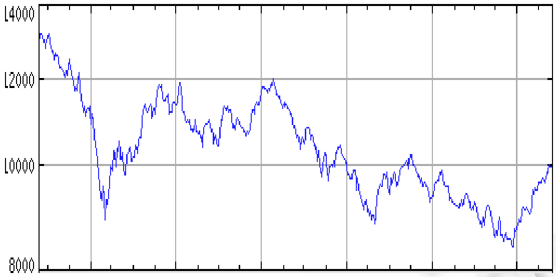

Fig.1 Curve change of HSI in the period Ⅰ

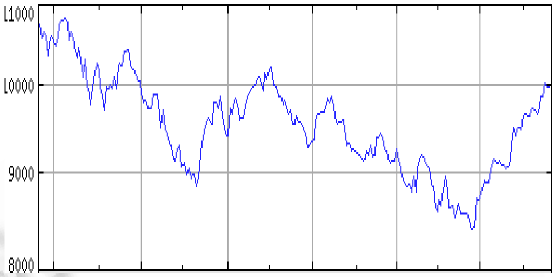

Fig.2 Curve change of HSI in the period Ⅱ

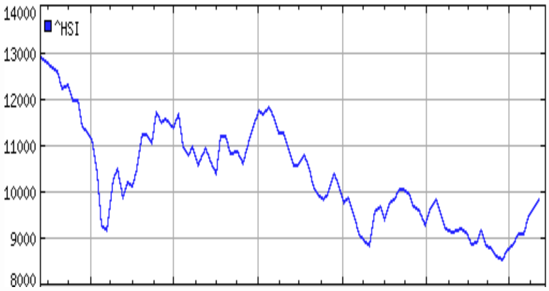

Fig.3 Synthesized curve of db4 coefficients from Fig.1

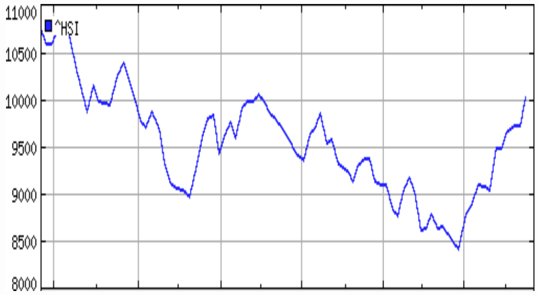

Fig.4 Synthesized curve of db4 coefficients from Fig.2

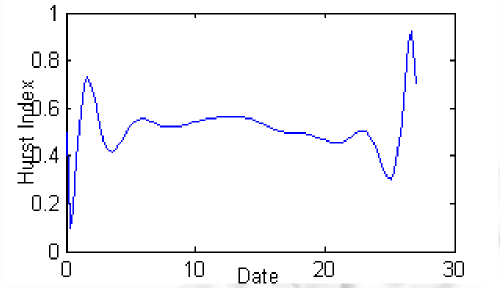

Fig.5 Hurst index curve of time series from Fig.1

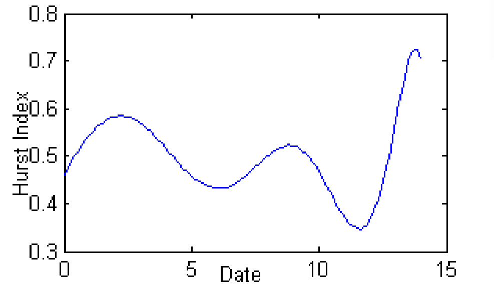

Fig.6 Hurst index curve of time series from Fig.2

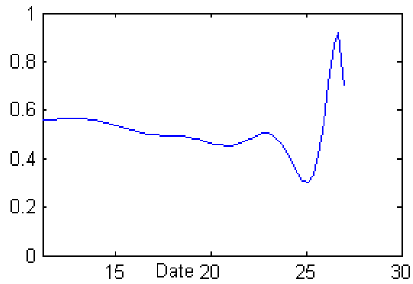

Fig.7 Hurst index curve of time sub-series from Fig.5

7 Conclusions

We have proposed a new standard of series similarity, by which the similarities of dynamic characters of data from two time series can be completely depicted.

The similarity of the basic curve figures is queried and measured at some resolution ratio level, while the fractal dimension in local similarity is matched. Daubechies wavelet is an important tool in data processing. This paper puts emphasis on algorithm and match of the fractal time-varying Hurst index curves. The effectiveness of the method is validated by means of the simulation example in the end. The work of this paper is the supplement and development of the study of similarity mentioned in the Refs.[6,7].

The above two kinds of similarity matching can be combined to depict the similarities of dynamic characters of data from two time series.

References:

[1] Agrawal R, Faloutsos C, Swami A, Efficient similarity search in sequence databases. In: Proc. of the 4th Conf. on Foundations of Data Organization and Algorithms. Chicago. 1993. 69~84.

[2] Chen MS, Han JW, Yu PS. Data mining: An overview from a database perspective. IEEE Trans. on Knowledge and Data Engineering, 1996,8(6):866~883.

[3] Wu D, Agrawal D, Abbadi AEI, Singh A, Smith TR. Efficient retrieval for browsing large image database. In: Proc. of the Conf. on Information and Knowledge Management. 1996. 11~18.

[4] Keogh E, Smyth P. A probabilistic approach to fast pattern matching in time series databases. In: Proc. of the 3rd Conf. on Knowledge Discovery in Database and Data Mining. 1997.

[5] Keogh EJ, Pazzani MJ. An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback. In: Proc. of the 4th Conf. on Knowledge Discovery in Database and Data Mining. AAAI Press, 1998. 239~241.

[6] Chan KP, Fu AWC. Efficient time series matching by wavelets. In: Proc. of the 15th IEEE Int'l Conf. on Data Engineering. Sydney, 1999. 126~133.

[7] Zheng C, Ouyang WM, Cai QS. An efficient dimensionality reduction technique for times series data sets. Mini-Macro System, 2002,23(11):1380~1383 (in Chinese of English abstract).

[8] Wang ZL. Time Series Analysis. Beijing: China Statistic Press, 1999. 143~180 (in Chinese).

[9] Wang K. Discovering patterns from large and dynamic sequential data. Journal of Intelligent Information Systems,1997,9(1):8~33.

[10] Kam PS, Fu AWC. Discovering temporal patterns for interval-based events. In: Proc. of the 2nd Int'l Conf. Data Warehousing and Knowledge Discovering (DaWaK2000). 2000.

[11] Evertsz CJG. Fractal geometry of financial time series. Fractals, 1995,3(3):609~616.

[12] Hou JR, Song GX. Application of wavelet analysis in the estimation of hurst index. Journal of Xidian University (Science Edition), 2002,29(1):113~118 (in Chinese with English abstract).

[13] Daubechies I. The wavelet transform: Time-Frequency localization and signal analysis. IEEE Trans. on Information Theory, 1990,36(5):961~1005.

附中文参考文献:

[7] 郑诚,欧阳为民,蔡庆生.一种有效的时间序列维数约简方法.小型微型计算机系统,2002,23(11):1380~1383.

[8] 王振龙.时间序列分析.北京:中国统计出版社,1999.143~180.

[12] 侯建荣,宋国乡.小波分析在 Hurst 指数估值中的应用.西安电子科技大学学报(自然科学版),2002,29(1):113~118.