ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.21, No.9, September 2010, pp.2161−2172
doi: 10.3724/SP.J.1001.2010.03633

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于小波的时间序列流伪周期检测方法*

李晓光[1+], 宋宝燕[1], 于 戈[2], 王大玲[2]

[1](辽宁大学 信息学院,辽宁 沈阳 110036)

[2](东北大学 信息科学与工程学院,辽宁 沈阳 110004)

Wavelet-Based Pseudo Period Detection on Time Series Stream

LI Xiao-Guang[1+], SONG Bao-Yan[1], YU Ge[2], WANG Da-Ling[2]

[1](School of Information, Liaoning University, Shenyang 110036, China)

[2](College of Information Science and Engineering, Northeastern University, Shenyang 110004, China)

+ Corresponding author: E-mail: xgli@lnu.edu.cn

Li XG, Song BY, Yu G, Wang DL. Wavelet-Based pseudo period detection on time series stream. *Journal of Software*, 2010,21(9):2161−2172. http://www.jos.org.cn/1000-9825/3633.htm

Abstract: A period detection method called MPD(memory-constrain period detection) is proposed naively on a time series stream, where the Haar-wavelet synopsis of series stream is adopted, and an estimated period based on partial fragments is proposed to improve the detection efficiency, and the cubic spline is used to detect period of arbitrary length. The time and space complexity error bound of MPD are validated through theoretical and experimental analysis.

Key words: pseudo period; time series stream; period detection

摘 要: 提出一种有效的时间序列流伪周期检测方法 MPD(memory-constrain period detection).它采用 Haar 小波技术构建时间序列流大纲,利用部分片段估计周期方法提高检测效率,采用基于三次插值的周期估计方法检测任意长度的周期.通过对 MPD 误差的理论分析和实验分析,验证了 MPD 的时间和空间复杂度以及检测误差的有效性.

关键词: 伪周期;时间序列流;周期检测

中图法分类号: TP311 **文献标识码**: A

与静态的时间序列相比较,时间序列流是一种动态的时间序列,流中的元素是按时间顺序的、快速变化的、海量的和潜在无限的.在现实生活中存在大量的具有周期特性的时间序列流,如在天气检测中的温度数据、ICU 病人呼吸、脉搏、心电图实时监控数据、太阳黑子监控数据等等.时间序列流的周期检测既可以提供流数据的周期波动特性,也是流变化检测、异常分析等时间序列流分析技术的基础.

由于各种干扰、噪音和其他复杂因素的影响,通常无法获得传统意义上的周期,即以周期间隔的数据间是相等的(如图 1 所示).伪周期是非精确的周期,其定义是以伪周期间隔的数据片段最相似(如图 2 所示).一般来说,为了检测伪周期,至少需要保存两个周期以上的数据.对于长周期的时间序列流来说,保存全部数据进行检测是不可行的.如:太阳黑子监控中的黑子活动周期为 11 年,当数据采集间隔为秒级时数据量非常大;对于短周期时

* Supported by the National Natural Science Foundation of China under Grant Nos.60703068, 60873068 (国家自然科学基金)
Received 2008-09-05; Revised 2009-02-13; Accepted 2009-04-10

间序列流来说,如 ICU 监控中的脉搏等,周期往往为秒级.在给定窗口内,需要大量的片段比较,其时间复杂度为片段数的平方,无法满足实时检测周期的要求.

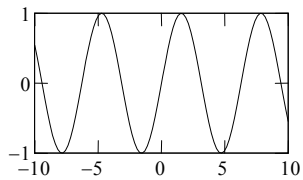


Fig.1 Time series stream with precise period
图 1 精确周期下的时间序列流

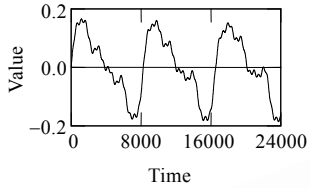


Fig.2 Time series stream with pseudo period
图 2 伪周期下的时间序列流

本文提出一种有效的伪周期检测方法——MPD(memory-constrain period detection),它通过采用小波技术和部分片段比较方法很好地解决了上述问题.本文主要的贡献有:

(1) 针对内存约束问题,提出一种基于小波的伪周期检测方法 MPD,它只保存低频系数,抛弃高频系数,从而构建数据大纲.给出 MPD 的误差理论分析,包括 MPD 误差期望分析,长、短周期检测误差分析和误差约束下的数据量分析.

(2) 针对检测效率问题,提出一种利用部分片段检测周期的方法 PMPD.在给定的误差上限,给出了选取片段数量下限的方法.PMPD 通过降低参与计算的片段数,从而提高小片段下的计算效率.

(3) 针对 MPD 只能发现 2^h 倍数周期的问题,提出基于三次插值的任意长度周期检测方法.

本文第 1 节给出问题定义.第 2 节总结相关工作.第 3 节给出 MPD 方法,包括检测算法、误差分析、基于插值的周期检测和数据量分析.第 4 节给出利用部分片段检测周期的 PMPD 方法,以及误差约束下片段数量确定方法.第 5 节是实验分析.第 6 节为本文的结论和未来工作.

1 问题提出

本文定义时间序列流为序列 $S=(x_0,x_1,\ldots,x_n)$,其中,n 随时间增加,$\forall x_t\in S,x_t$ 为实数.

定义 1.1(周期流). 如果$\exists T\in\mathbf{Z}^+,\forall x_t\in S,x_{t+T}=x_t$且不存在 $T'<T,\forall x_t\in S,x_{t+T'}=x_t$,称时间序列流是周期 T 的周期流.

定义 1.2(伪周期流). 给定周期 T 的周期流 S,如果 $x_{t+T}-x_t=\varepsilon_t,\varepsilon_t$ 为独立的随机变量,则称 S 为伪周期流.

定义 1.3(T 片段). 称序列 $F=(x_t,x_{t+1},\ldots,x_{t+T-1})$为 T 片段.给定时间序列流 S,称序列集合$\{(x_{n-kT},x_{n-kT+1},\ldots,x_{n-(k-1)T-1})|k=1\sim\lfloor n/T\rfloor\}$为 S 的 T 片段集合,记作 S_T.

定义 1.4(伪周期). 给定伪周期流 S,令

$$D(T)=\left(\frac{2}{\lfloor n/T\rfloor(\lfloor n/T\rfloor-1)}\sum_{F_i,F_j\in S_T,i<j}d(F_i,F_j)\right)^{1/2} \tag{1}$$

则 S 的伪周期 T^*定义为

$$T^*=\arg_T\min\{D(T)|T=1\sim\lfloor n/2\rfloor\}\text{且 }D(T)\leqslant o,$$

其中,o 为阈值,$d(F_i,F_j)$为计算片段相似性函数.

给定一个流 S,为了求解伪周期 T^*,一种朴素方法就是令 $T=1\sim\lfloor n/2\rfloor$,依次计算 $D(T)$,并取小于阈值 o 的最小 D 值时的 T 为 T^*.然而,这种方法存在以下问题:

(1) 如果一个流 S 的伪周期为 T^*,那么至少需要 $2T^*$的数据才能求解.那么给定内存约束 M,对于长周期来说,如果 $2T^*>M$,则无法找到 T^*.实际上,随着数据的无限到达,可以认为 $n>>M$.

(2) 给定 T 时,朴素方法的时间复杂度为 $O(\lfloor n/T\rfloor^2)$.对于短周期或者当 T 较小时,其计算代价相当高.

针对问题(1),通常的解决方法是构建数据流的大纲(synopses),即用一个远小于原数据流的数据结构来近似描述流.本文采用了多分辨率方法中的 Haar 小波[1]技术构建数据流的大纲,通过抛弃高频信息(细节信息)近似

描述原数据.当采用大纲后,那么与原数据相比,在大纲上所获得伪周期的误差是多少?针对问题(2),我们提出采用随机选取片段方法来降低计算的基数,那么关键在于“选取多少片段数才能保证伪周期以高置信度在允许误差范围内”.

2 相关工作

据我们所知,迄今为止与本文最为相关的周期检测研究是文献[2]提出的在大规模静态时间序列上的趋势检测方法.为了甄别时间序列上的具有代表性的片段,文献[2]提出松弛周期和平均趋势概念来衡量代表性片段.本文的伪周期定义与松弛周期定义一致.为了提高计算松弛周期的效率,文献[2]采用了 sketch 作为时间序列的大纲,其存在以下不足:

(1) 对给定长度 l 的片段,sketch 的大小为 $k=(9\log l)/\varepsilon^2$,如果 $\varepsilon=0.01$,那么其 k 值很大.

(2) 必须存储双倍的 sketch,且需要计算 $l=2^r$ 的 sketch.尽管提高了片段距离计算速度,但其误差上界增加了 1 倍.

(3) 需要事先计算 sketch,无法处理增量数据,只适用于静态时间序列.

(4) 需要对时间序列上任意长度为 l 的片段存储 sketch,不适用于时间序列流上的周期检测.

(5) 实验显示周期检测的误差率很高.

另外,文献[3]重点研究了周期模式管理问题,其周期检测采用了动态设置阈值方法.相比之下,本文的伪周期定义更具一般性.

时间序列流上周期检测的关键点数据大纲构建和片段相似性计算效率.当前提出许多静态序列大纲构建方法,如 SVD[4],DFT[5],DWT[6],PLA[7],PAA[8],随机抽样[9],滑动窗口[10],小波[1,11],梗概(sketch)[2].根据本文对问题的定义,小波和梗概是合适的.但正如上面阐述的,由于空间约束,sketch 不适用于本文,本文方法是基于 Haar 小波技术的,但只保留第 h 层上的低频信息并抛弃高频信息来近似序列流,可以线性更新大纲.文献[12]采用了同本文相似的大纲,主要用于对齐多数据流.与片段相似性计算相关的是时间序列的相似性查找,其重点是如何高效率地计算两个片段的相似性.当然,在计算给定的两个片段相似性时可以完全借鉴已有的工作,但本文的重点是如何高效地计算任意两个片段的向量性.为此,我们采用了片段选取的策略,通过降低基数提高计算效率.

3 基于小波的伪周期检测方法

本文借鉴了 Haar 小波思想,设计了基于内存限制的周期检测方法 MPD,进而给出 MPD 方法的误差分析.

3.1 Haar小波

Haar 分解定理给出了逐层分解数据信息的方法.为了方便后续问题的讨论,本文将 Haar 层次分解法表示为下面的矩阵形式.详细的 Haar 小波定义和分解定理见文献[1].令 $(a_{m,0},a_{m,1},\ldots,a_{m,2^m-1})$ 为数据向量,$n=2^m$,A_h 和 D_h 分别为 $2^{m-h}\times 2^m$ 矩阵(见文献[1]),h 为分解层次.令 $M_h=[D_1^T,D_2^T,\ldots,D_h^T,A_h^T]^T$ 为层次 h 的小波变换矩阵.那么有

$$[d_{m-1,0},\ldots,d_{m-1,2^{m-1}-1},d_{m-2,0},\ldots,d_{m-2,2^{m-2}-1},\ldots,d_{m-h,0},\ldots,d_{m-h,2^{m-h}-1},a_{m-h,0},\ldots,a_{m-h,2^{m-h}-1}]^T=M_h\times[a_{m,0},a_{m,1},\ldots,a_{m,2^m-1}]^T.$$

这里,我们称 $(d_{m-1,0},\ldots,d_{m-1,2^{m-1}-1},d_{m-2,0},\ldots,d_{m-2,2^{m-2}-1},\ldots,d_{m-h,0},\ldots,d_{m-h,2^{m-h}-1})$ 为高频系数向量(细节向量),$(a_{m-h,0},\ldots,a_{m-h,2^{m-h}-1})$ 为低频系数向量(均值向量).本文利用低频系数向量来拟合原始数据,即在第 h 层小波变换时,在第 $h-1$ 层的低频系数基础上,计算并保留第 h 层的低频系数,抛弃了高频系数.

3.2 MPD检测算法

$d(F_i,F_j)$ 为计算片段相似性的函数.目前,常用的相似性函数有 L_p,相关系数、余弦距离等,其中最常用的为欧式距离(L_2).为了消除比较长度对欧式距离的影响,我们采用了平均欧式距离,具体公式如下:

$$d(F_i,F_j)=\frac{1}{T}\sum_t(x_t-y_t)^2,x_t\in F_i,y_t\in F_j \tag{2}$$

MPD 周期检测方法见算法 1,其基本思想是:设置大小为 M 的缓冲区 S',当数据量大于内存约束 M 时,对当前缓冲区中数据进行次低分辨率的小波变换,并且抛弃变换后的高频信息,仅保留当前层的低频信息来近似估计伪周期(步骤 4~步骤 6).注意,如步骤 2、步骤 3 所示,如果当前层为 t,且$|S'|<M$,则需等待 2^t 个数据计算它们的低频系数并保存.

算法 1. MPD.

输入:时间序列流 S,内存大小 $M=2^h$(h 为一给定正整数),阈值 o;

输出:伪周期 T.

1) $t=0$; let S' be a array of size M, where $|S'|$ is the element count of S';

2) for each new arrival sequence X of size 2^t in S

3) if $|S'|<M$ then append $\sum_{x_i\in X} x_i \Big/ \sqrt{2^t}$ into S' else

4) for $i=0\sim 2^{h-1}-1$ do update x_i in S' with $(x_{2i}+x_{2i+1})/2^{1/2}$;

5) delete the elements from $M/2-1$ to $M-1$ in S'; t++;

6) for $T=1\sim|S'|/2$ compute $D(T)$ and output the 2^tT as current period with respect to the minimal value of $D(T)$ and less than the threshold o.

3.3 MPD误差分析

类似于文献[1]中的引理证明,我们有:

引理 3.1. 设 M_h 是小波变换矩阵,M_h^T 为 M_h 的转置矩阵,则 M_h^T 是 M_h 的逆矩阵.

定理 3.1. 给定片段 F_i 和 F_j,M 为小波变换矩阵,则 $d(F_i,F_j)=d(F_i^a,F_j^a)+d(F_i^d,F_j^d)$,$F^a$ 和 F^d 分为片段 F 的低频系数向量和高频系数向量.

证明: $F_i'^T=MF_i^T$, $F_j'^T=MF_j^T$,则 $M^TF_i'^T=M^TMF_i^T\Rightarrow M^TF_i'^T=F_i^T\Rightarrow F_i'M=F_i$ (因为 M 的逆矩阵和转置矩阵相同,即 $MM^T=I$(见引理 3.1));同理, $F_j'M=F_j$,那么有

$$\begin{aligned}Td(F_i,F_j)&=T\|F_i-F_j\|^2=T(F_i-F_j)(F_i-F_j)^T=T(F_i'M-F_j'M)(M^TF_i'^T-M^TF_j'^T)\\&=T(F_i'-F_j')MM^T(F_i'^T-F_j'^T)=T\|F_i'-F_j'\|^2=T\|(F_i^a,F_i^d)-(F_j^a,F_j^d)\|^2\\&=T\|F_i^a-F_j^a\|^2+T\|F_i^d-F_j^d\|^2=Td(F_i^a,F_j^a)+Td(F_i^d,F_j^d).\end{aligned}$$

由于算法 MPD 只是利用低频系数求解 T,则 MPD 算法中的片段距离为 F^a 上的距离 $D^a(T)$,其与真实 $D(T)$ 的差值$\omega=D(T)-D^a(T)$,称ω为 MPD 的距离误差. □

定理 3.2. 给定 T,算法 MPD 的距离误差$\omega=D(T)-D^a(T)\leqslant D^d(T)$,其中

$$D^a(T)=\left(\frac{2}{\lfloor n/T\rfloor(\lfloor n/T\rfloor-1)}\sum_{F_i,F_j\in S_T,i<j}d(F_i^a,F_j^a)\right)^{1/2},$$

$$D^d(T)=\left(\frac{2}{\lfloor n/T\rfloor(\lfloor n/T\rfloor-1)}\sum_{F_i,F_j\in S_T,i<j}d(F_i^d,F_j^d)\right)^{1/2}.$$

证明:根据定理 3.1,有

$$D(T)=\left(\frac{2}{\lfloor n/T\rfloor(\lfloor n/T\rfloor-1)}\sum_{F_i,F_j\in S_T,i<j}d(F_i,F_j)\right)^{1/2}$$

$$=\left(\frac{2}{\lfloor n/T\rfloor(\lfloor n/T\rfloor-1)}\left(\sum_{F_i,F_j\in S_T,i<j}d(F_i^a,F_j^a)+\sum_{F_i,F_j\in S_T,i<j}d(F_i^d,F_j^d)\right)\right)^{1/2}$$

$$\leqslant\left(\frac{2}{\lfloor n/T\rfloor(\lfloor n/T\rfloor-1)}\left(\sum_{F_i,F_j\in S_T,i<j}d(F_i^a,F_j^a)\right)\right)^{1/2}+\left(\frac{2}{\lfloor n/T\rfloor(\lfloor n/T\rfloor-1)}\left(\sum_{F_i,F_j\in S_T,i<j}d(F_i^d,F_j^d)\right)\right)^{1/2}$$

$$=D^a(T)+D^d(T).$$ □

定理 3.3. 不失一般性,设 $n=2^m$,小波分辨率为 h,$T=r2^h$,$r=1\sim2^{m-h-1}$,令 $z_{ijk}=(x_{ik}-x_{jk})^2$,$k\in[1,T]$, $x_i\in F_i^d$, $x_j\in F_j^d$. z_{ijk} 为随机变量,设对任意 i,j,k,z_{ijk} 之间独立且 $E(z_{ijk})=\mu$,$D(z_{ijk})=\sigma^2$.那么在给定 T 下,$D^d(T)$期望值为

$$E(D^d(T))\approx\frac{1}{4}+\frac{\sqrt{l(2^m-2^{m-h})}\mu}{2\sqrt{2\pi}\sigma},$$

其中,$l=\lfloor2^{m-h}/r\rfloor(\lfloor2^{m-h}/r\rfloor-1)/2$.

证明:给定 T,对任意片段的高频系数向量大小为(2^m-2^{m-h}),片段数为$\lfloor n/T\rfloor=\lfloor2^{m-h}/r\rfloor$,根据流特性,$z_{ijk}$ 数量为 $|T|\lfloor n/T\rfloor(\lfloor n/T\rfloor-1)/2>>30$,且 z_{ijk} 间独立,$E(z_{ijk})=\mu$,$D(z_{ijk})=\sigma^2$.根据中心极限定理,有 $D^d(T)^2\sim N(((2^m-2^{m-h})/T)\mu$, $((2^m-2^{m-h})/T^2l)\sigma^2)$;又 $D^d(T)\geqslant0$,则有 $P(D^d(T))=0$,$D^d(T)<0$.令$\mu'=((2^m-2^{m-h})/T)\mu$,$\sigma'=((2^m-2^{m-h})/T^2l)^{1/2}\sigma$,那么

$$E(D^d(T))=\int_{-\infty}^{+\infty}t\frac{1}{\sqrt{2\pi}\sigma'}\mathrm{e}^{\frac{-(t^2-\mu')^2}{2\sigma'^2}}\mathrm{d}t=\int_0^{+\infty}t\frac{1}{2\sqrt{\pi}}\mathrm{e}^{\frac{-(t^2-\mu')^2}{2\sigma'^2}}\mathrm{d}\frac{(t^2-\mu')}{\sqrt{2}\sigma'}$$

$$=\int_{-\frac{\mu'}{\sqrt{2}\sigma'}}^{+\infty}\frac{1}{2\sqrt{\pi}}\mathrm{e}^{-x^2}\mathrm{d}x=\int_0^{+\infty}\frac{1}{2\sqrt{\pi}}\mathrm{e}^{-x^2}\mathrm{d}x+\int_{-\frac{\mu'}{\sqrt{2}\sigma'}}^{0}\frac{1}{2\sqrt{\pi}}\mathrm{e}^{-x^2}\mathrm{d}x,$$

其中,$\int_0^{+\infty}\mathrm{e}^{-x^2}\mathrm{d}x=\frac{\sqrt{\pi}}{2}$.而 $\int_{-\frac{\mu'}{\sqrt{2}\sigma'}}^{0}\mathrm{e}^{-x^2}\mathrm{d}x$ 为不能积分的超越函数.根据 Taylor 展开式可得

$$\int_{-\frac{\mu'}{\sqrt{2}l\sigma}}^{0}\mathrm{e}^{-x^2}\mathrm{d}x=\left[\sum_{i=0}^{\infty}(-1)^i\frac{x^{2i+1}}{(2i+1)i!}\right]_{-\frac{\mu'}{\sqrt{2}\sigma'}}^{0}=\sum_{i=0}^{\infty}(-1)^i\frac{\left(\frac{\mu'}{\sqrt{2}\sigma'}\right)^{2i+1}}{(2i+1)i!},$$

则

$$E(D^d(T))=\frac{1}{4}+\frac{1}{2\sqrt{\pi}}\sum_{i=0}^{\infty}(-1)^i\frac{\left(\frac{\mu'}{\sqrt{2}\sigma'}\right)^{2i+1}}{(2i+1)i!}.$$

这里,i 取值为 0,则 $E(D^d(T))\approx\frac{1}{4}+\frac{\mu'}{2\sqrt{2\pi}\sigma'}=\frac{1}{4}+\frac{\sqrt{l(2^m-2^{m-h})}\mu}{2\sqrt{2\pi}\sigma}$. □

实际上,对于绝大部分数据来说,高频系数是非常小的[12].而 z_{ijk} 为高频系数的差值平方,那么 z_{ijk} 值大多落在很小的区间内.这里,在 z_{ijk} 分布未知的情况下,根据 z_{ijk} 的上述特性,不妨认为其分布为指数分布.我们通过大量实验也验证了这一点.

引理 3.2. 如果 z_{ijk} 符合区间参数为λ的指数分布,则 $E(D^d(T))\approx\frac{1}{4}+\frac{\sqrt{l(2^m-2^{m-h})}}{2\sqrt{2\pi}}$.

证明:$E(z_{ijk})=\mu=1/\lambda$,$D(z_{ijk})=\sigma^2=1/\lambda^2$,根据定理 3.3 可证. □

3.4 MPD讨论

3.4.1 影响因素

从引理 3.2 或定理 3.3 可以看出,影响 $E(D^d(T))$的有数据量 $n=2^m$ 以及分辨率 h 和片段比较次数 l,且 $E(D^d(T))$ 是 n,h 和 l 值的递增函数.根据引理 3.2,$E(D^d(T))$分别对 n,h 和 l 求偏导,可得如下公式(如果根据定理 3.3,只需在

下式前乘以μ/σ即可,且不影响分析结果):

$$\frac{\partial E(D^d(T))}{\partial n}=\frac{\sqrt{l(1-2^{-h})}}{4\sqrt{2\pi n}} \tag{3}$$

$$\frac{\partial E(D^d(T))}{\partial l}=\frac{\sqrt{(2^m-2^{m-h})}}{4\sqrt{2\pi l}} \tag{4}$$

$$\frac{\partial E(D^d(T))}{\partial h}=\frac{l2^{m-h}}{4\ln 2\sqrt{2\pi l(2^m-2^{m-h})}} \tag{5}$$

对于给定的 T,随着数据量 n 增加,l 也增加(但 l 最多为$|W|(|W|-1)/2$,$|W|$为窗口大小),那么 $E(D^d(T))$是增加的.但从公式(4)可以看出,其增加幅度非常小.从公式(5)可以看出,h 的增加对 $E(D^d(T))$的影响较大.h 与窗口大小有关,即在给定的 T 下,窗口越大,其 h 值越小,反之则越大.对于长周期来说,h 值较大,而 l 值较小.对于短周期来说,h 值较小,而 l 值偏大.因此总体来说,$D^d(T)$的期望值受 h 和 l 值影响较小.对于数据量 n 来说,从公式(3)可以看出,当 n 较大时,其对 $E(D^d(T))$的影响较小.但当数据量 n 较小时,其对 $E(D^d(T))$具有较大影响.

3.4.2 数据量下限

定理 3.4. 给定误差上限θ,对于周期 $T=r2^h$ 和置信度δ,如果当前数据量为 $n\geqslant\frac{2\pi(1-\delta)^2\sigma^2(2-2^{1-h})}{\theta^4}-r2^h$,则算法 MPD 满足 $P(D^d(T)\leqslant\theta)\geqslant 1-\delta$.

证明:根据定理 3.3 中的证明可知,$D^d(T)^2\sim N(((2^m-2^{m-h})/T)\mu,((2^m-2^{m-h})/T^2l)\sigma^2)$,又 $D^d(T)\geq 0$,则有 $P(D^d(T))=0$,$D^d(T)<0$.令$\mu'=((2^m-2^{m-h})/T)\mu$,$\sigma'=((2^m-2^{m-h})/T^2l)^{1/2}\sigma$,那么

$$P(D^d(T)\leqslant\theta)=P(D^d(T)^2\leqslant\theta^2)=\int_0^{\theta^2}\frac{1}{\sqrt{2\pi}\sigma'}\mathrm{e}^{\frac{-(t-\mu')^2}{2\sigma'^2}}\mathrm{d}t=\int_0^{\theta^2}\frac{1}{\sqrt{\pi}}\mathrm{e}^{\frac{-(t-\mu')^2}{2\sigma'^2}}\mathrm{d}\left(\frac{(t-\mu')}{\sqrt{2}\sigma'}\right)$$

$$=\frac{1}{\sqrt{\pi}}\int_{\frac{-\mu'}{\sqrt{2}\sigma'}}^{\frac{\theta^2-\mu'}{\sqrt{2}\sigma'}}\mathrm{e}^{-x^2}\mathrm{d}x=\frac{1}{\sqrt{\pi}}\left[\sum_{i=0}^{\infty}(-1)^i\frac{x^{2i+1}}{(2i+1)i!}\right]_{\frac{-\mu'}{\sqrt{2}\sigma'}}^{\frac{\theta^2-\mu'}{\sqrt{2}\sigma'}}=\frac{1}{\sqrt{\pi}}\sum_{i=0}^{\infty}(-1)^i\frac{(\theta^2-\mu')^{2i+1}+(\mu')^{2i+1}}{(2i+1)i!(\sqrt{2}\sigma')^{2i+1}},$$

则 $\frac{1}{\sqrt{\pi}}\sum_{i=0}^{\infty}(-1)^i\frac{(\theta^2-\mu')^{2i+1}+(\mu')^{2i+1}}{(2i+1)i!(\sqrt{2}\sigma')^{2i+1}}\approx\frac{\theta^2}{\sqrt{2\pi}\sigma'}$.令 $\frac{\theta^2}{\sqrt{2\pi}\sigma'}=1-\delta$,则有

$$\frac{\theta^2}{\sqrt{2\pi}\sqrt{\frac{(2^m-2^{m-h})}{r^2 2^{2h}l}}\sigma}=1-\delta\Rightarrow\frac{\theta^4}{2\pi(1-\delta)^2\sigma^2}=\frac{2^m-2^{m-h}}{2^{2m-1}+r2^{m+h-1}}\Rightarrow\frac{2\pi(1-\delta)^2\sigma^2(2-2^{1-h})}{\theta^4}-r2^h=2^m=n,$$

又,$\frac{\theta^2}{\sqrt{2\pi}\sigma'}$ 为 $2^m=n$ 的递增函数,因此定理 3.4 可证.证毕. □

3.4.3 周期插值

算法 MPD 利用低频系数计算周期,但仅能发现 $r2^h$ 长度的周期,并且根据讨论,随着 h 的增加,误差的期望值增加.针对上述问题,我们采用插值方法来解决,那么此时,一方面可以发现非 $r2^h$ 长度的周期,另一方面,误差可以进一步减少(见实验).具体过程是:利用算法 MPD 求解一系列 T 值,然后对该系列进行 Lagrange 三次插值,最后取插值后的最小值所对应的 T 为 T^*.Lagrange 三次插值函数如下:

$$\begin{aligned}D_3(t)=&\frac{(t-t_{i+1})(t-t_{i+2})(t-t_{i+3})}{(t_i-t_{i+1})(t_i-t_{i+2})(t_i-t_{i+3})}D_i+\frac{(t-t_i)(t-t_{i+2})(t-t_{i+3})}{(t_{i+1}-t_i)(t_{i+1}-t_{i+2})(t_{i+1}-t_{i+3})}D_{i+1}+\\&\frac{(t-t_i)(t-t_{i+1})(t-t_{i+3})}{(t_{i+2}-t_i)(t_{i+2}-t_{i+1})(t_{i+2}-t_{i+3})}D_{i+2}+\frac{(t-t_i)(t-t_{i+1})(t-t_{i+2})}{(t_{i+3}-t_i)(t_{i+3}-t_{i+1})(t_{i+3}-t_{i+2})}D_{i+3}\end{aligned} \tag{6}$$

其中,$t_{i+1}<t\leqslant t_{i+2}$,$t_i$ 为 MPD 中片段距离 D_i 对应的 T 值,$i=1\sim\lfloor n/2\rfloor$.

4 基于部分片段的伪周期检测

给定周期 T,MPD 的时间复杂度为 $O(\lfloor n/T\rfloor^2)$,那么当 T 值较小时,MPD 的运行时间是相当高的.这里,我们提出一种基于部分片段的 MPD 算法——PMPD.基本思想是:给定周期 T,在$\lfloor n/T\rfloor$个片段中随机选取 $m<\lfloor n/T\rfloor$个片段来计算 D 值,记为 D_s.设此时的比较片段比较次数为 $s(s=m(m-1)/2)$,那么,PMPD 的关键在于 s 的取值.如果 s 过小,时间复杂度会大幅降低,但 D_s 与 D 的误差会很大;如果 s 取值较大,计算结果更准确,但仍然会很耗时.这里,s 的取值应满足:给定误差γ和置信度δ,选择 $m<\lfloor n/T\rfloor$个片段,使得 $P(|D_s-D|\leqslant\gamma)\geqslant 1-\delta$.

定理 4.1. 不失一般性,对任意 $i,j,i\neq j$,令 $d(F_i^a,F_j^a)$ 为同分布、相互独立的随机变量,且有 $b\geqslant d(F_i^a,F_j^a)\geqslant a$.那么,给定误差上限$\gamma$和置信度$\delta$,当选择 $s\geqslant\dfrac{b-a}{\gamma}\sqrt{\dfrac{1}{2}\ln\left(\dfrac{2}{\delta}\right)}$ 次片段比较时,$P(|D_s-D|\leqslant\gamma)\geqslant 1-\delta$.

证明:令 $d_{ij}\equiv d(F_i^a,F_j^a)$,$E(d_{ij})=\mu$.$I$ 为所选取片段集合,$s=|I|(|I|-1)/2$,则 $D_s=\dfrac{1}{s}\sum_{i,j\in I,i\neq j}d_{ij}$.因为 d_{ij} 为同分布、相互独立的随机变量,那么,

$$E(D_s)=\frac{1}{s}E\left(\sum_{i,j\in I,i\neq j}d_{ij}\right)=\frac{1}{s}\sum_{i,j\in I,i\neq j}(d_{ij})=\mu .$$

由于$\lfloor n/T\rfloor(\lfloor n/T\rfloor-1)/2$ 非常大(>>30),根据辛钦定理, $D\xrightarrow{p=1}\mu$,即 D 以概率 1 近似μ.

根据 Hoeffding 不等式[13],有

$$P(|D_s-\mu|\geqslant\gamma)=P(|D_s-D|\geqslant\gamma)\leqslant 2e^{-2s\gamma^2/(b-a)^2}\Rightarrow 1-P(|D_s-D|\geqslant\gamma)\geqslant 1-2e^{-2s\gamma^2/(b-a)^2}$$
$$\Rightarrow P(|D_s-D|\leqslant\gamma)\geqslant 1-2e^{-2s\gamma^2/(b-a)^2}$$

令 $1-2e^{-2s\gamma^2/(b-a)^2}\geq 1-\delta$,则 $s\geq\dfrac{b-a}{\gamma}\sqrt{\dfrac{1}{2}\ln\left(\dfrac{2}{\delta}\right)}$. □

具体的 PMPD 算法见算法 2.与 MPD(算法 1)的不同之处在于,PMPD 在计算 D 值时,利用定理 4.1 计算最小比较次数 s(步骤 6).如果 s 小于在所有片段上的比较次数,则随机选取部分片段,使其比较次数为 s(步骤 8),即用 Ds 近似 D;否则,在全部片段上计算 D(步骤 9).

算法 2. PMPD.

输入:时间序列流 S,内存大小 $M=2^h$,阈值 o,误差上限γ和置信度δ;

输出:伪周期 T.

1) $t=0$; let S' be a array of size M, where $|S'|$ is the element count of S';
2) for each new arrival sequence X of size 2^t in S
3) if $|S'|<M$ then append $\sum_{x_i\in X}x_i\Big/\sqrt{2^t}$ into S' else
4) for $i=0\sim 2^{h-1}-1$ do update x_i in S' with $(x_{2i}+x_{2i+1})/2^{1/2}$;
5) delete the elements from $M/2-1$ to $M-1$ in S'; t++;
6) compute s by Theorem 4.1 and γ,δ;
7) for $T=1\sim|S'|/2$ do
8) if $s<\dfrac{|S'|}{2T}\left(\dfrac{|S'|}{T}-1\right)$, then compute $D(T)$ by randomly selected fragments of size T from S', where s' distance computations are required.
9) else compute $D(T)$ by the whole fragments of size T from S'
10) output the 2^tT as current period with respect to the minimal value of $D(T)$ and less than the threshold o.

5 实验分析

5.1 实验方法

主要分析比较了本文提出的检测方法和文献[2]中的基于 Sketch 的检测方法.将三次插值后的 MPD 记做 MPDs,基于部分片段的 MPD 记做 PMPD,基于 sketch 的检测方法记做 Sketch.实验采用两种测试数据:人工生成数据和真实数据.人工数据分别为 SinSeries 和 RandomSeries,其中,SinSeries 数据产生器为

$$X_t=A\times\sin(t/T)+\varepsilon_{t \bmod 2T\pi},\varepsilon_i\sim N(\mu_i,\sigma_i).$$

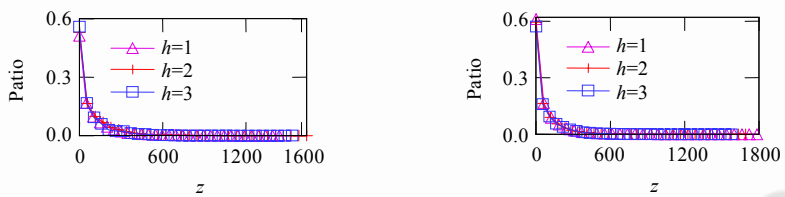
μ_i,σ_i 为随机数,i=1~2$T\pi$,2$T\pi$为周期,A 为振幅.这里取 10.RandomSeries 数据产生器为 $X_t=R_{t \bmod T}+\varepsilon_{t \bmod T}$,$R_t\in R[1\ldots T]$,$R[1\ldots T]$是长度为 T 且每个元素值位于区间[0,1500]中的随机序列,ε同上.真实数据为 62 年间太阳黑子的检测数据,数据量为 22 735 天,周期约为 11 年(4 033 天).评价指标采用 T 值相对误差(α_T)和 D 值相对误差(α_D),公式如下:

$$\alpha_T=\left|\frac{\overline{T}-T^*}{T^*}\right|,\alpha_D=\left|\frac{\overline{D}-D^*}{D^*}\right|,$$

其中,T^*和 D^*为真实值,$\overline{T}$ 和 $\overline{D}$ 为计算值.

5.2 MPD误差实验分析

(1) z_{ijk} 直方图.由于空间限制,只给出了部分数据和 h 下的 z_{ijk} 直方图.如图 3 所示,z_{ijk} 符合指数分布,其中,RandomSeries 下的均值平均为 100.36 和方差平均为 141.657,SinSeries 下的均值平均为 82.39 和方差平均为 117.52.

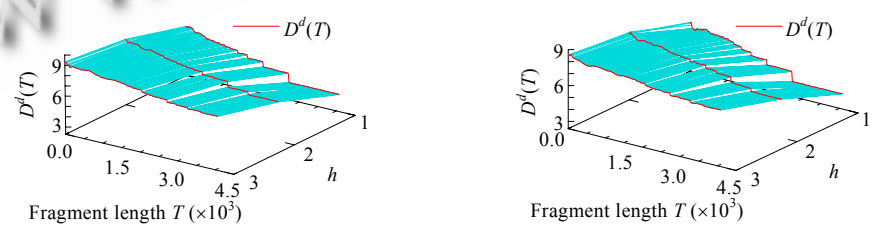


(a) RandomSeries (b) SinSeries

Fig.3 Histogram of z_{ijk}

图 3 z_{ijk} 直方图

(2) h 和 l 与 MPD 的距离误差上限 $D^d(T)$.主要验证 h 和 l 对 MPD 的距离误差上限 $D^d(T)$的影响.这里,数据大小为 16K,最大分辨率为 3,真实周期为 1 375.结果如图 4 所示.可以看出,由于 D^d 是 h 和 l 的增函数,且相对于 l,随着 h 的增加 D^d 增加幅度更大.在 h 较大 l 较小(T 较大)时,或者 l 较大(T 较小)h 较小时,D^d 均较小.那么对于给定窗口大小的 MPD,总体来说,D^d 值受 h 和 l 值影响较小,如第 5.3 节中的比较分析实验所示,MPD 的α_T 值非常小.

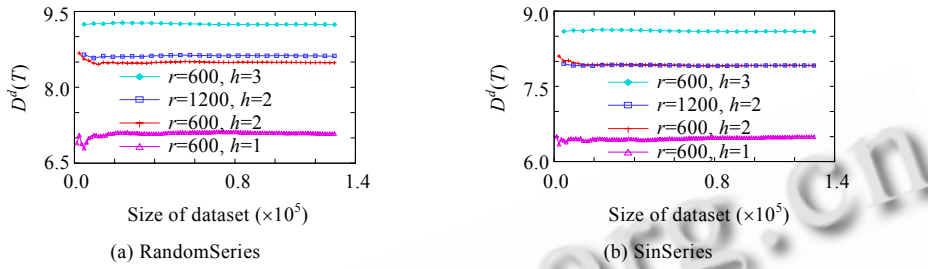


(a) RandomSeries (b) SinSeries

Fig.4 Values of MPD's $D^d(T)$ varied with h and l

图 4 h 和 l 与 MPD 的 $D^d(T)$

(3) 数据量 n 下限与 MPD 距离误差上限 $D^d(T)$.主要验证数据量 n 对 MPD 距离误差上限 $D^d(T)$的影响.这里,θ=0.01 和δ=0.025,最大分辨率为 3,真实周期为 1 375,如图 5 所示.根据定理 3.4 和所设置的参数可得,n 下限最大为 13 120,从图 5 可以看出,当数据量大于下限时,随着 n 的增加,$D^d(T)$的增幅非常缓慢.与 h 相比,当 r 增加时(T 增加),对 $D^d(T)$影响更大,这也进一步验证了上面的实验结论.



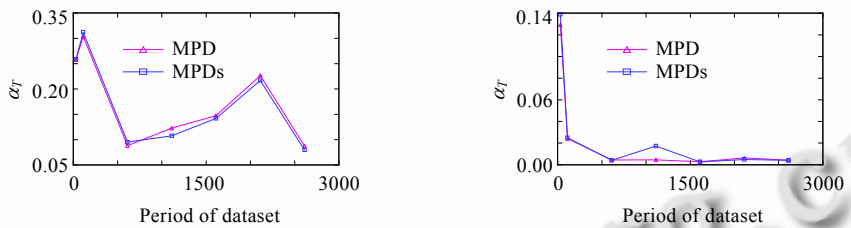
(a) RandomSeries (b) SinSeries

Fig.5 Values of MPD's $D^d(T)$ varied with the lower bound of n

图 5 数据量 n 的下限与 MPD 的 $D^d(T)$

5.3 比较分析

(1) 三次插值.主要验证三次插值后 MPD 的α_T值变化.这里,数据大小 1M,窗口大小 1 024,真实周期分别为 111,611,1 111,1 611,2 111 和 2 611.从图 6 可以看出,三次插值前后,MPD 的α_T值比较接近.当周期较小时,三次插值方法的α_T值略差.而当周期较大时,三次插值方法的α_T值较好.因为 MPD 只能发现 $r2^h$ 长度的周期,且随着 h 的增加,未插值时的 MPD 的误差 D^d 增大;而三次插值通过拟合和预测,可以获得任意长度的周期,在一定程度上降低了这种误差.

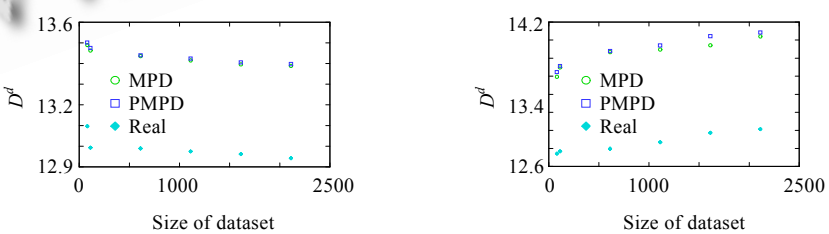


(a) RandomSeries (b) SinSeries

Fig.6 α_T after cubic spline

图 6 三次插值后的α_T值

(2) 片段选取.主要验证基于部分片段的 MPD 方法 PMPD 的α_T值.这里,数据大小为 1M,窗口大小为 1 024,真实周期同上.γ分别取 5,10 和 20,δ=0.05.对于 RandomSeries 来说,$b-a$=1489;对于 SinSeries 来说,$b-a$=1844.经过计算片段数分别为 272,136 和 68.如图 7 所示,对于 RandomSeries 数据和 SinSeries 数据来说,MPD 的平均α_D分别为 0.033 和 0.081,PMPD 的平均α_D分别为 0.034 和 0.082.可以看出,MPD 的α_D很小,PMPD 与 MPD 非常接近.

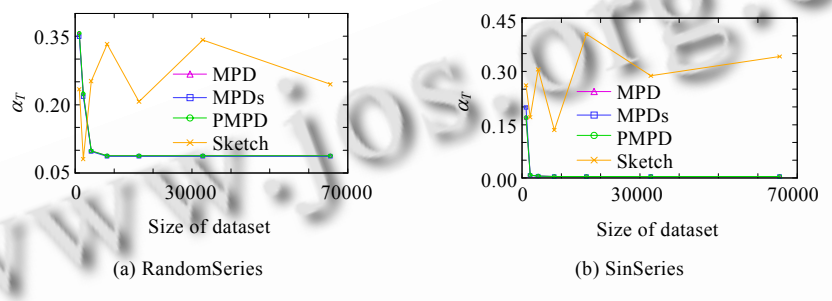


(a) RandomSeries (b) SinSeries

Fig.7 Scatter graph of D^d of PMPD and MPD

图 7 PMPD 与 MPD 的 D^d 值散点图

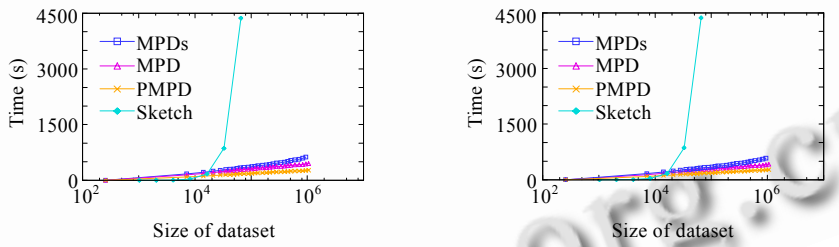
(3) Sketch vs. MPD.因为 sketch 方法为面向固定数据长度的方法,所以实验中数据序列长度固定.周期为 611,数据量大小为 1M.关于 Sketch 的 k 取值,文献[2]中分别根据公式计算 $k=(9\log T)/\varepsilon^2$,$\varepsilon=0.01$ 和 $2\log T$ 来计算,并通过实验显示后者效果更好,这里我们采用了后者.如图 8 所示,同 Sketch 相比,MPD,MPDs 和 PMPD 的 α_T 值在两个数据集上分别平均提高 99.9968%,99.9969%,99.9968%和 99.9995%,99.9994%,99.9995%.由于太阳黑子数据没有精确周期,这里我们采用 Naive 方法获得的伪周期值 3 841 作为参照标准.MPD 的检测结果为 3 836,α_T 为 0.013;MPDs 为 3 845,α_T 为 0.011;Sketch 为 3 645,α_T 为 0.051.从图 9 可以看出,Sketch 的时间复杂度是呈指数,其主要原因在于 Sketch 计算的时间复杂度为指数;而 MPD,MPDs 和 PMPD 的时间复杂度是 $O(n^2)$.其中,由于片段选取,PMPD 的效率最高,并且其 α_T 与其他两种方法十分接近.那么,随着窗口大小的增加,对于一个给定的 T 来说,窗口越大则 h 越小,则 α_T 越小,但运行时间却呈多项式增加.从图 10 和图 11 可以看出,较小的窗口($|W|=256$)与较大窗口($|W|=2048$)相比,它们的 α_T 相差很小,但 PMPD 的运行时间的增长率要远远低于 MPD 和 MPDs.



(a) RandomSeries (b) SinSeries

Fig.8 α_T comparison among Sketch, MPD, MPDs and PMPD

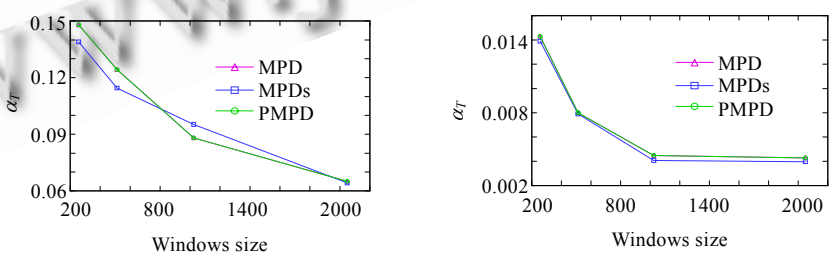
图 8 Sketch,MPD,MPDs 和 PMPD 的 α_T 值比较



(a) RandomSeries (b) SinSeries

Fig.9 Runtime comparison among Sketch, MPD, MPDs and PMPD

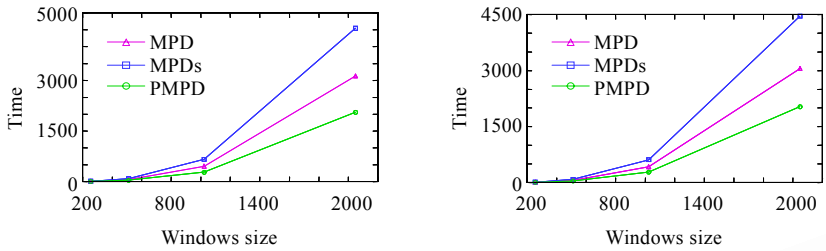
图 9 Sketch,MPD,MPDs 和 PMPD 运行时间比较



(a) RandomSeries (b) SinSeries

Fig.10 α_T comparison among MPD, MPDs and PMPD with varied windows

图 10 不同窗口下 MPD,MPDs 和 PMPD 的 α_T 值比较



(a) RandomSeries (b) SinSeries

Fig.11 Runtime comparison among MPD, MPDs and PMPD with varied windows

图 11 不同窗口下 MPD,MPDs 和 PMPD 运行时间比较

6 结论和未来工作

本文提出的基于小波的伪周期检测方法 MPD,其检测误差平均为 0.008,同基于 sketch 的检测方法相比提高了 99.9968%.基于部分片段检测周期的方法 PMPD 可以大幅提高 MPD 效率的同时,周期检测误差仅有微小增加.同基于 sketch 相比,PMPD 时间效率提高了 96.78%,而且通过三次插值,可以进一步将周期检测误差平均降低了 2.2624%.未来工作中,我们将进一步研究伸缩伪周期(scaled periodic)和平移伪周期(shift periodic)检测方法以及周期变化检测.

References:

[1] Chen AL, Tang CJ, Yuan CA, Peng J, Hu JJ. An anti-noise algorithm for mining asynchronous coincidence pattern in multi-streams. Journal of Software, 2006,17(8):1753−1763 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/17/1753.htm [doi: 10.1360/jos171753]

[2] Indyk P, Koudas N. Identifying representative trends in massive time series data sets using sketches. In: Abbadi AE, ed. Proc. of the 26th Int'l Conf. on Very Large Data Bases. San Fransisco: Morgan Kaufmann Publishers, 2000. 363−372.

[3] Tang L, Cui B, Li HY, Miao GS, Yang DQ, Zhou XB. Effective variation management for pseudo periodical streams. In: Chan CY, ed. Proc. of the ACM SIGMOD Int'l Conf. on Management of Data. New York: ACM, 2007. 257−268.

[4] Kanth KVR, Agrawal D, Singh A. Dimensionality reduction for similarity searching in dynamic databases. In: Tiwary A, ed. Proc. of the ACM SIGMOD Int'l Conf. on Management of Data. New York: ACM Press, 1998. 166−176.

[5] Rafiei D, Mendelzon AO. Efficient retrieval of similar time sequences using DFT. In: Tanaka K, ed. Proc. of the 5th Int'l Conf. on Foundations of Data Organizations and Algorithms (FODO). Netherlands: Kluwer Press, 1998. 249−257.

[6] Chan KP, Fu AWC. Efficient time series matching by wavelets. In: Kitsuregawa M, ed. Proc. of the ICDE Int'l Conf. on Data Engineering. Sydeny: IEEE Press, 1999. 126−133.

[7] Chen QX, Chen L, Lian X, Liu YB, Yu JX. Indexable PLA for efficient similarity search. In: Koch C, ed. Proc. of the Int'l Conf. on Very Large Data Bases. New York: ACM Press, 2007. 435−446.

[8] Yi BK, Faloutsos C. Fast time sequence indexing for arbitrary *Ap* norms. In: Abbadi AE, ed. Proc. of the 26th Int'l Conf. on Very Large Data Bases. San Fransisco: Morgan Kaufmann Publishers, 2000. 385−394.

[9] Vitter JS. Random sampling with a reservoir. ACM Trans. on Mathematical Software, 1985,11(1):37−57. [doi: 10.1145/3147.3165]

[10] Datar M, Gionis A, Indyk P, Motwani R. Maintaining stream statistics over sliding windows. SIAM Journal on Computing, 2002, 31(6):1794−1813. [doi: 10.1137/S0097539701398363]

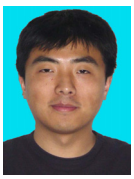
[11] Matias Y, Vitter JS, Wang M. Wavelet-Based histograms for selectivity estimation. ACM SIGMOD Record, 1998,27(2):448−459. [doi: 10.1145/276305.276344]

[12] Sakurai Y, Papadimitriou S, Faloutsos C. BRAID: Stream mining through group lag correlations. In: Özcan F, ed. Proc. of the ACM SIGMOD Int'l Conf. on Management of Data. New York: ACM Press, 2005. 599−610.

[13] Hoeffding W. Probability inequalities for sums of bounded random variables. Journal of the American Statistical Association, 1963,58(1):13−30.

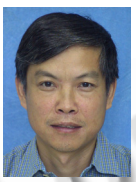
附中文参考文献:

[1] 陈安龙,唐常杰,元昌安,彭京,胡建军.挖掘多数据流的异步偶合模式的抗噪声算法.软件学报,2006,17(8):1753−1763. http://www.jos.org.cn/1000-9825/17/1753.htm [doi: 10.1360/jos171753]



李晓光(1973－),男,辽宁沈阳人,博士,副教授,CCF 会员,主要研究领域为数据挖掘,XML 数据库,信息检索.

宋宝燕(1965－),女,博士,教授,CCF 高级会员,主要研究领域为数据流管理,数据挖掘.



于戈(1962－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为数据库理论与技术.

王大玲(1962－),女,博士,教授,CCF 高级会员,主要研究领域为数据库,数据挖掘.