第24卷 第12期 北京理工大学学报 Vol.24 No.12
2004年12月 Transactions of Beijing Institute of Technology Dec. 2004

文章编号:1001-0645(2004)12-1070-05

基于小波变换和反馈的时间序列相似模式搜索算法

秦吉胜[1], 王淑静[2], 宋瀚涛[1]

(1. 北京理工大学 信息科学技术学院计算机科学工程系,北京 100081; 2. 中国航空结算中心,北京 100028)

摘 要: 为得到有价值的相似时间序列,分析小波变换及其在寻找相似时间序列上的优越性;指出了现有的基于小波变换的相似时间序列搜索算法的2个缺点,提出了基于小波变换和加权反馈的时间序列相似模式匹配算法和基于验证深度的验证方法;通过实验证明了算法的有效性和实用性.

关键词: 时间序列; 相似模式; 小波变换; 反馈; 验证深度

中图分类号: TP 182 **文献标识码**: A

Algorithm for Finding Similar Patterns over Time-Series Data Based on Wavelets and Feedback

QIN Ji-sheng[1], WANG Shu-jing[2], SONG Han-tao[1]

(1. Department of Computer Science and Engineering, School of Information Science and Technology, Beijing Institute of Technology, Beijing 100081, China; 2. China Aviation Accounting Center, Beijing 100028, China)

Abstract: To obtain valuable similar time-series, the advantage of the wavelet transform and its performance in finding similar patterns over time-series data are analyzed. The disadvantages of existing algorithms based on wavelet transform are summarized. An algorithm for finding similar patterns over time-series data based on wavelets and feedback is put forward, besides the method for validating results based on the depth of verification. The validity and practicability of the algorithm are proved through experiments.

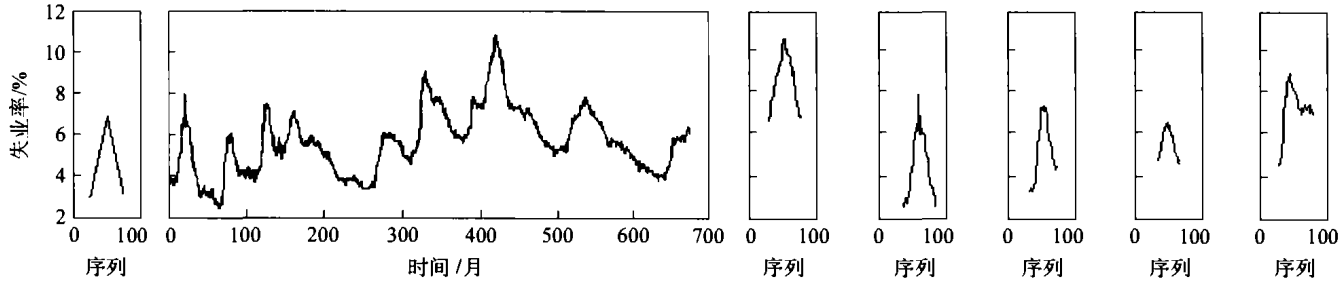
Key words: time-series; similar pattern; wavelet transform; feedback; verification-depth

时间序列数据广泛存在于商业等领域,它描述了某个变量随时间的变化. 当前对时间序列的相似匹配算法的研究主要有:Agrawal 等提出使用 DFT(离散傅里叶变换)进行特征提取,并使用傅里叶变换的前几个系数近似原序列,同时提出了一个索引机制 F-Index[1];Faloutsos 等把 F-Index 推广应用到子序列的匹配问题[2];Rafiei 和 Mendelzon 建议使用 DFT 的对称性提高距离度量的精度[3];Chan 和 Fu 使用 Haar 小波变换进行相似搜索,提高了算法的性能[4];StruZik 和 Siebes 也使用一种和 Haar 小波变换类似的方法 PAA(piecewise aggregate approximation)用于进行相似搜索[5,6];Yi 和 Faloutsos 提出一种基于分段思想的相似变换方法[7]. 但基于小波变换的时间序列匹配算法存在2个问题.

① 用户无法有效地参与到挖掘过程中. 算法开始时,需要准备一段标准参考序列,即在长时间序列中寻找的目标序列,但是用户往往在开始时并不明确所要查询的序列,最初的需求往往是一个形状,而不是一个确定的数字序列,如搜寻目标是:寻找与图

收稿日期: 2004-02-25
基金项目: 国家部委预研项目(G1998030414)
作者简介: 秦吉胜(1976-),男,博士生,E-mail:qinjisheng@hotmail.com; 宋瀚涛(1940-),男,教授,博士生导师.

1a 所示形状相似的子序列，那么，图 1b 所示的时间序列在不同尺度的小波变换域上，可能搜索到许多相似子序列，如图 1c～ 图 1g，已提出的小波算法到此会把所有结果输出，算法结束. 但是，并不需要那些无用的子序列，假如只对其中的 3 个子序列分别(如图 1c，1d，1e)感兴趣，所要寻找的序列的特征与它们相似，那么挖掘还将继续. 本文将控制理论中的反馈原理引入到序列匹配中，为这些感兴趣的序列和标准参考序列赋予一定的权值，组成新的参考序列，再次进行相似搜索，以达到令用户满意的效果. 作者还给出了一个基于小波变换和加权反馈的时间序列相似模式匹配算法.



(a) 参考序列　(b) 待搜索时间序列　(c) 序列 1　(d) 序列 2　(e) 序列 3　(f) 序列 4　(g) 序列 5

图 1　美国 1948-01-01～2003-09-01 失业率数据及其分析图

Fig. 1　U. S. Unemployment rate from 1948-01-01 to 2003-09-01 and analysis

② 对某个尺度的小波变换空间进行序列匹配后得到匹配序列，需要对其结果进行验证，在已知的文献中验证都是在原序列上进行的，这种方法是不科学的，在此作者提出了验证深度的概念及其实现.

1　离散小波变换[8]

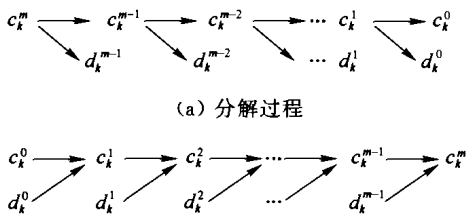
小波变换是从经典傅里叶变换发展而来，它通过满足 $\int_R \varphi(x)\mathrm{d}x=0$ 的基本小波函数 $\psi(x)$ 的伸缩和平移构成一族小波函数系表示和逼近一个函数. 对任意平方可积函数 $f(x)$，其离散小波变换为

$$W_\varphi f(j,k)=\int_{-\infty}^{\infty} f(x)\overline{\varphi_{j,k}(x)}\mathrm{d}x=$$

$$\langle f,\varphi_{j,k}\rangle,\quad f(x)\in L^2(R).$$

在数字化实现中，离散小波变换是利用 Mallat 分解与重构算法完成的. 一维信号的小波分解与重构过程可由 2 组滤波器级联滤波而产生. 设 $\{C_k^m\}$ 为输入序列，$\{C_k^{m-i}\}$ 为经 i 次低通滤波得到的输出，$\{d_k^{m-i}\}$ 为第 i 次高通滤波得到的输出. 一维信号的小波分解和重构过程如图 2 所示. 每次对低频分量(即近似信号)进行分解. 这样得到的 C_k^{m-i} 是在尺度 i 上的近似信号，d_k^{m-i} 为尺度 i 上的细节信号.

利用离散小波变换进行时间序列相似模式搜索是因为小波变换具有以下性质和优点[8,9]：① 局部特征. 小波变换有无限基函数，可以捕捉到数据的局部特性，而傅里叶变换只能捕捉到数据的整体特征. ② 降维功能. 小波变换后的序列长度是变换前的 1/2，而且在降维过程中会剔除原始序列中的高频噪声信号. ③ 多分辨分析. 小波变换是分等级的，对于不同的应用，可以方便地调整，随着尺度的增加，形状越来越清晰. ④ 效率高. 小波变换算法的执行速度非常快，时间复杂度为 $O(n)$(n 为序列长度)，而傅里叶变换的时间复杂度是 $O(n^2)$，快速傅里叶变换(FFT)的时间复杂度是 $O(n\lg n)$.



(a) 分解过程

(b) 重构过程

图 2　一维信号的小波分解与重构过程

Fig. 2　Wavelet decompose and re-construct procedure of 1-D signal

2　基于小波变换和加权反馈的时间序列相似模式匹配算法

文献[10]中给出了一个基于 DFT 变换和反馈的时间序列相似模式匹配算法. 与 DFT 变换相比，小波变换具有更高的效率和性能，并且 DFT 方法在确定权值时，需要把信号从频域转换到时域复原信号，以使用户确定其兴趣度. 而小波变换后的序列直接表示了原序列的形状，用户可以直接确定其兴趣度，不需要进行逆变换.

2.1　概念和定义

时间序列相似模式匹配就是在给定序列 S 中

查找与目标序列 T 相似的子序列 F，查找到的 F 有多个，T 和 F 之间的相似性度量使用欧几里德距离.

定义 1 欧几里德距离. 设 $T=(T_1,T_2,\cdots,T_n)$，$F=(F_1,F_2,\cdots,F_n)$，则 T 与 F 间的欧几里德距离定义为

$$d(T,F)=\left|\sum_{i=1}^{n}(T_i-F_i)^2\right|^{1/2}.$$

给定一个阈值 ε，如果 T 和 F 之间的距离 $d(T,F)\leqslant\varepsilon$，则 T 与 F 相似. 一般情况下，数据库中的时间序列 S 和目标序列 T 都很长，计算需要较长时间，因而首先需要用小波变换进行降维处理，得到降维后的序列 S' 和 T'，然后计算 S' 和 T' 之间的距离 $d'=d(T',F')$. 根据 Parseval 定理[8]，如果采用的是正交小波，那么小波域序列之间的欧氏距离将不超过原始序列之间的欧氏距离，即 $d'\leqslant d$，因此，如果 $d\leqslant\varepsilon$，则必然有 $d'\leqslant\varepsilon$. 这一性质保证了在小波域使用范围查询得到的序列包含了所有的正确结果，但是可能存在小波域检索到的却不能满足条件的序列.

定义 2 范围查询. 假设原时间序列的长度为 n，小波域序列的长度为 l，一般 $l\ll n$. 范围查询的定义为：对一个查询点 $T=(T_1,T_2,\cdots,T_n)$，检索所有与 T 距离不超过 ε 的点集.

在进行 T 点的 ε 范围查询时，首先在小波域内基于 R*-tree 索引进行 T' 的 ε 范围查询，然后把在小波域检索到的初步检索序列进行验证就可以去掉不符合条件的序列. Parseval 定理使得该方法可以适用于任何一种正交小波[1].

2.2 算法

基于小波变换和加权反馈的时间序列相似模式匹配算法的思想是：首先，对给定长度为 N 的序列 S，按照目标序列 T 的长度 n 每间隔 1 个数据点取 1 个长度为 n 的子序列，这样得到 $N-n+1$ 个子序列，并对目标序列 T 和 $N-n+1$ 个子序列进行 k 尺度小波变换，将这些小波系数采用 R*-tree 进行索引；然后，在小波域采用范围查询得到 T' 的 ε 范围相似子序列并进行深度验证，把初步结果展示给用户. 用户对感兴趣的相似子序列赋予一定的权值，并把这些相似子序列和原目标序列 T' 按权值叠加，得到新的目标序列 T'. 再次对 T' 进行 ε 范围查询和反馈，直到得到满意的结果. 算法的流程描述如下.

输入：待查询的序列 S，目标序列 T，小波变换尺度 k，验证深度 m，查询范围 ε.

输出：相似子序列集合 F.

① 设置相似子序列集合 F 为空.

② 对 S 进行 n 长度子序列划分，得到子序列数组 H，对 H 和 T 进行 k 尺度小波变换，得到 H' 和 T'，同时对 H 进行 m 尺度小波变换，得到验证序列 P.

③ 将 H' 成员和其 MBR 插入并形成一个 R*-tree.

④ 对于每一个叶子节点 r，计算其与 T' 之间的距离 $d(r,T')$，如果 $d(r,T')\leqslant\varepsilon$，则把该叶子节点插入临时相似子序列集合 Q.

⑤ 对于 Q 中的每一个成员，使用 P，对 m 深度进行验证，去掉在 m 深度中不满足 ε 范围查询的子序列.

⑥ 用户对 Q 中的元素和 T' 分别赋予一定的权值，并叠加得到新的查询目标序列 T'. 假设 Q 中的元素为 $\{S_1,S_2,\cdots,S_p\}$，对应的权值为 $\{w_1,w_2,\cdots,w_p\}$，原始 T' 的权值为 W，则得到新的查询目标序列 T' 的叠加公式为 $T'=(S_1w_1+S_2w_2+\cdots+S_pw_p+T'W)/(w_1+w_2+\cdots+w_p+W)$，跳转到第④步，直到得到满意的结果.

⑦ 返回相似子序列集合 F，其中，小波变换尺度 k 决定了算法执行的维数和算法的效率，k 越大，则维数 $n/2^k$ 越小，算法处理的数据越少，速度越快；同时 k 越大，越显示子序列的整体形状，但会丢失序列的有用信息，因而，k 的数值需根据具体应用选择. 每个子序列权值大小由用户自己决定，这就使用户可以参与到挖掘的过程中，提高了算法的实用性.

3 时间序列相似模式验证算法

Parseval 定理决定了在尺度 k 上进行小波变换和 ε 范围查询得到的相似序列，不会产生遗漏，但可能存在原序列上不满足 ε 范围的序列，所以需要对这些相似子序列进行验证. 传统算法中，是在原序列上对这些子序列进行验证，检查这些子序列和目标序列之间的欧氏距离是否满足 $d(T,F)\leqslant\varepsilon$，这虽然符合问题的要求，但验证方法过于单一. 由于范围 ε 是个经验值，如果数据存在很强的噪声，会因为某 1 个噪声而去掉 1 个可能非常有用的相似序列.

例如，1 个如图 3a 所示标准的正弦信号数据序列 T，$\varepsilon=1$，在尺度 k 上查找到 1 个待验证的相似子序列 F，如图 3b 所示，T 和 F 2 个序列的长度都为

128，因为噪声的原因，使得 $F[80]$ 数据产生了突变 $T[80]=\sin[(80/128)2\pi]=-0.707$，$F[80]=2$，其余位置的数据 $T[i]=F[i]$（$i=0,1,2,79,81,\cdots,127$），可以计算 $d(T,F)=2.707>1=\varepsilon$，因此去掉该子序列. 但是，这个子序列是非常有价值的，去掉它是错误的. 如果对 T 和 F 都做一次小波变换，可以去掉该噪声，得到 T'（图 3c）和 F'（图 3d），可以计算得到 $d(T',F')=0<1=\varepsilon$，就可以把这个子序列找回来. 因此，只在原序列上进行验证是不科学的，所以提出了验证深度的概念.

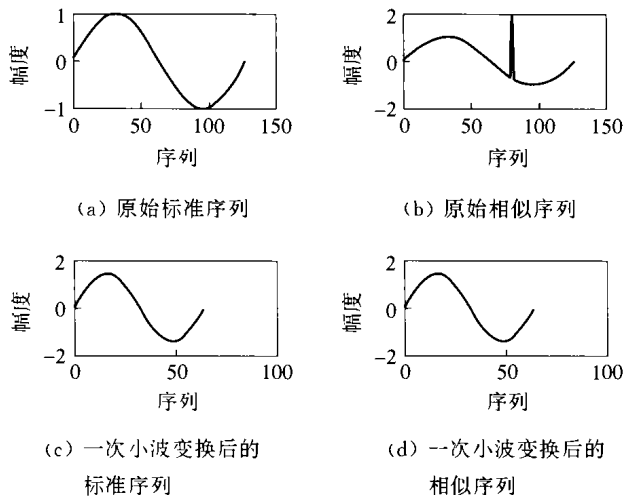


图 3　验证深度必要性说明图

Fig. 3　Explanation of verification-depth is necessary

定义 3　验证深度. 在 k 尺度小波变换下得到的相似序列需要进行验证. 验证可以在原序列上进行（$m=0$），也可以在尺度 m（$d=1,2,3,\cdots,k-1$）的小波变换域上进行，m 被称为验证深度. 验证深度是一个相似性严格程度的度量. 如果验证深度 $m=0$，则表示验证在原序列上进行，那么数据的任何细节都将被计算在内，如果 $m=1$，则去掉了一部分细节，而更关注数据的整体形状. 依次类推，m 越大，说明越关注整体形状，对相似性的要求越宽松；而 m 越小，则说明越关注细节，对相似性的要求越严格.

验证深度可以由用户自行调整，它表达了用户对相似性的苛求度，适当的验证深度可以避免丢失一些有价值的相似序列，增加挖掘的灵活性和实用性.

4　算法评估

关于小波变换的性能及其与 DFT 的比较，已经在文献[9]中做了详细说明，本文不再赘述，下面仅对作者提出的反馈算法和验证深度做必要的测试.

算法使用的测试数据来源于 http://research.stlouisfed.org/fred2/series/UNRATE 提供的美国 1948-01-01 到 2003-09-01 的失业率按月时间序列 S，共 669 个数据点，如图 1b 所示. 对该时间序列，使用基于小波变换和反馈的算法挖掘相似时间序列，构造 1 个标准参考序列 T（如图 1a 所示）. T 表示了 1 个明显的上升和下降过程，代表了 1 个波峰，共有 32 个数据点. 对 S 进行每隔 1 个数据点得到 1 个长度为 32 的子序列抽样，可以得到 638 个子序列集 F. 然后对标准参考序列 T 和 F 进行尺度 2 的小波变换并进行范围查询，可以得到如图 1c～图 1g 的相似序列，对这些序列进行深度为 1 的验证，即对它们的一次小波变换进行范围验证，可以去除图 1g.

在余下的序列中，若对图 1d，1c 的序列比较感兴趣，分别对其赋予一定的权值，并和原参照序列 T 进行加权平均，得到新的参考序列. 利用反馈原理对新的参考序列再次进行范围查询，得到相似序列为图 1d 和图 1c. 对这 2 个序列进行深度为 1 的验证后，得到的最终相似时间序列为图 1d 和图 1c. 对这 2 个结果进行考察发现，图 1d 代表美国 1948 年 3 月到 1951 年 1 月的失业率，其中波峰在 1949 年 10 月，为 7.9%.

图 1c 代表美国 1981 年 8 月到 1984 年 3 月的失业率，其中波峰在 1982 年 11 月和 12 月之间，为 10.8%. 以上查询结果证明了基于小波变换和反馈原理的相似时间序列查询算法的有效性和允许用户参与而产生的结果的有用性；同时进行了一定深度的验证，使得查询结果更关注于趋势的变换，避免去掉有用的时间序列.

参考文献：

[1] Agrawal R, Faloutsos C, Swami A. Efficient similarity search in sequence databases [A]. Proceedings of the 4th International Conference of Foundations of Data Organization and Algorithms (FODO) [C]. London: Springer-Verlag, 1993. 69－84.

[2] Faloutsos C, Ranganathan M, Manolopoulos Y. Fast subsequence matching in time-series databases [A]. Proc of the ACM SIGMOD [C]. New York: ACM Press, 1994. 419－429.

（下转第 1095 面）

地提高 HK001 惯导系统的精度.

参考文献：

[1] 林士谔. 动力调谐陀螺仪[M]. 北京：国防工业出版社，1983.
Lin Shi'e. Dynamically tuned gyroscope[M]. Beijing: National Defence Industry Press, 1983. (in Chinese)

[2] 邓正隆. 惯性导航原理[M]. 哈尔滨：哈尔滨工业大学出版社，1994.
Deng Zhenglong. Theory of inertial navigation[M]. Harbin: Harbin Institute of Technology Press, 1994. (in Chinese)

[3] 崔中兴. 惯性导航系统[M]. 北京：国防工业出版社，1982.
Cui Zhongxing. Inertial navigation system [M]. Beijing : National Defence Industry Press, 1982. (in Chinese)

[4] Britting K R. Inertial navigation systems analysis [M]. New York: Wiley, 1971.

[5] 任坚信，张　鹏，任思聪. 大角速率动调陀螺仪的运动分析与误差研究[J]. 仪表技术与传感器，2002，12：49－51.
Ren Jianxin, Zhang Peng, Ren Sicong. Motion and error analysis of DTG with larger angular velocity[J]. Instrument Technique and Sensor, 2002, 12: 49－51. (in Chinese)

[6] 孟　中，张　涛. 降低动力调谐陀螺输出噪声的方法[J]. 光学精密工程，2002，10(4)：420－424.
Meng Zhong, Zhang Tao. Method used in the suppression of DTG output noise [J]. Optics and Precision Engineering, 2002, 10(4): 420－424. (in Chinese)

[7] 杨梅仓. 动力调谐陀螺的锥形运动及噪声信号的抑制[J]. 中国惯性技术学报，1994，2(2)：28－33.
Yang Meicang. The conic movement and the noise signal suppression of dynamically tuned gyroscope[J]. Journal of Chinese Inertial Technology, 1994, 2(2): 28－33. (in Chinese)

[8] 缪玲娟，陈家斌. 动力调谐陀螺寻北系统及其误差分析[J]. 北京理工大学学报，1997，17(3)：374－379.
Miao Lingjuan, Chen Jiabin. Study on DTG north seeking system and its errors[J]. Journal of Beijing Institute of Technology, 1997, 17(3): 374－379. (in Chinese)

[9] Qi Yutong, Chen Fengyu, Su Haibin. Error analysis of a dynamically tuned gyro strapdown northfinder [J]. Journal of Beijing Institute of Technology, 1999, 8(3): 331－336.

[10] Yang Weiqin, Jiang Hong. Modeling nonstationary time series for gyroscopic drift analysing[J]. Journal of Beijing Institute of Technology, 1995, 4(1): 1－6.

(上接第 1073 面)

[3] Rafiei D, Mendelzon A. Similarity-based queries for time series data[A]. Proc of the ACM SIGMOD Conf [C]. New York: ACM Press, 1997. 13－25.

[4] Chan Kinpong, Fu Ada Wai-chee. Efficient time series matching by wavelets[A]. Proc of the ICDE Conf[C]. Washing: IEEE Computer Society, 1999. 126－133.

[5] Ruzik Z R S, Siebes A P J M. The Haar wavelet transform in the time series similarity paradigm[A]. Principles of Data Mining and Knowledge Discovery [C]. London: Springer-Verlag, 1999. 12－22.

[6] Keogh E, Chakrabarti K, Pazzani M, et al. Dimensionality reduction for fast similarity search in large time series databases [J]. Knowledge and Information Systems, 2001, 3(3): 263－286.

[7] Yi B K, Faloutsos C. Fast time sequence indexing for arbitrary L_p norms[A]. Proc of the VLDB Conf[C]. Cairo Egypt: Morgan Kaufmann, 2000. 385－394.

[8] 冉启文. 小波变换与分数傅里叶变换理论及应用[M]. 哈尔滨：哈尔滨工业大学出版社，2001. 102－108.
Ran Qiwen. Wavelet transform and discrete Fourier transform theory and application [M]. Haerbin: Haerbin Institute of Technology Press, 2001. 102－108. (in Chihese)

[9] Popivanov I, Miller R J. Similarity search over time-series data using wavelets[A]. Proceeding of the 18th International Conference on Data Engineering [C]. Lilleakerveien Norway: ICDE Press, 2002. 212－221.

[10] 郑斌祥，席裕庚，杜秀华. 利用反馈的时序模式挖掘算法研究[J]. 控制与决策，2002，17(5)：527－532.
Zheng Binxiang, Xi Yugeng, Du Xiuhua. Research on similarity mining in time series data sets by feedback[J]. Control and Decision, 2002, 17(5): 527－532. (in Chinese)

[11] Beckmann N, Kriegel H P, Schneider R, et al. The R*-tree: An efficient and robust access method for points and rectangles [A]. Proceedings of ACM SIGMOD[C]. New York: ACM Press, 1990. 322－331.