面向相似性搜索的时间序列表示方法述评

刘世元　江　浩
(华中科技大学机械科学与工程学院,武汉 430074)
E-mail:jiang_hao1979@163.com

摘　要　时间序列作为一种数据形式,广泛存在于各种商业、医学、工程、自然科学和社会科学等数据库中。近年来,时间序列的相似性搜索问题正得到越来越多的重视。该问题可描述为给定某个的时间序列,要求从一个大型时间序列数据库中找出与之最相似的序列。该问题的有效求解涉及到两个关键难点,即相似性度量的定义和搜索算法的时间复杂度,而这两者都依赖于时间序列的近似表示方法。因此,通过详细评述面向相似性搜索的各种时间序列近似表示方法,对这些方法进行分析和比较,总结了这些方法的优点和不足,并对进一步的研究方向作出了预测。

关键词　时间序列　相似性　数据挖掘　维规约

文章编号 1002-8331-(2004)27-0053-07　**文献标识码** A　**中图分类号** TP182

A Review on Time Series Representation for Similarity-based Pattern Search

Liu Shiyuan　Jiang Hao
(School of Mechanical of Science and Engineering, Huazhong University of Science and Technology, Wuhan 430074)

Abstract: As a data form, time series exists broadly in many applications in the databases of business, medicine, engineering, natural sciences and social sciences. Recently, the similarity-based pattern search of time series has received increasing attention. The problem can be described as: searching the sequence most similar to a given time series from a large time series database. The efficient solution to such a problem involves two difficulties, i.e., the definition of similarity measurement and the time complexity of searching algorithm, both of which rely on the approximate representation of time series. Thus, this paper reviews several methods of time series representation for similarity-based pattern search, compares and shows their advantages as well as limitations, and speculates the possible directions in the field.

Keywords: time series, similarity, data mining, dimensionality reduction

1　引言

时间序列是指按时间顺序排列的一组数据,作为数据库中的一种数据形式,它广泛存在于各种大型的商业、医学、工程、和社会科学等数据库中,形成规模庞大的时间序列数据库。这些巨量时间序列数据库真实地记录了应用系统在各个时刻的所有重要信息,如果能发展某种高效的数据处理方法,发现其中各时间序列之间的相互关系,即以某种度量来表征两个时间序列之间的相似性,并以此为基础实现多个时间序列数据的相似性搜索、聚类和模式发现等,必将大大提高这些时间序列数据库的实用价值。因此,近年来时间序列数据的相似性搜索与模式发现研究正得到越来越多的重视。

时间序列数据的相似性搜索问题最早由 IBM 公司的 Agrawal 等人于 1993 年提出,该问题描述为"给定某个时间序列,要求从一个大型时间序列数据库中找出与之最相似的序列"[1]。对两个时间序列进行比较,最简单的算法通常是比较两序列的时间多项式,即通过对线性或二次多项式的比较,找到两者之间的偏移量,然后对比所采用的相似性度量,就可以判断该两序列是否相似及其相似程度。要在数据库中完成相似性的搜索或查询,则需要将查询序列和数据库中的一系列的序列进行比较[2]。对于一个给定长度为 m 的序列和长度为 n 的数据库,其时间计算复杂度为 $O(mn)$,甚至是 $O(mn^2)$。显然,这种程度的时间复杂度在实际应用中是无法接受的。此外,时间序列数据通常还具有数值范围非有穷甚至离散、采样速度不恒定、干扰噪声形式多样等问题。所以,对时间序列采用近似表示,进行有效的维规约(dimensionality reduction)以消除数据冗余是十分必要的。从理论上来看,基于统计特性描述(如一阶统计量和高阶统计量)或参数建模(如 AR 建模和 ARMA 建模)的传统时间序列分析方法有可能用来解决相似性问题,但实际上并不能得到很好的结果,其主要困难在于相似性度量的定义和算法的时间复杂度,而这两者都依赖于时间序列的近似表示方法。因此,寻求某种鲁棒性强且计算复杂度低的时间序列近似表示方法,一直是解决相似性搜索问题的关键。

迄今为止,时间序列相似性搜索问题已经提出了 10 年左右的时间,在这段时间内,先后出现了许多面向相似性搜索的

基金项目:国家自然科学基金项目(编号:50205009)资助

作者简介:刘世元(1970-),副教授,主要从事信号分析,人工智能,知识发现与数据挖掘,设备故障诊断与监测等方面的研究。江浩(1979-),硕士研究生,从事智能诊断,信号处理与数据挖掘等方面的研究。

时间序列近似表示方法，如 Agrawal 采用的离散傅立叶变换(DFT,Discrete Fourier Transform)[1]、Chan 等人采用的 Haar 小波变换方法[3-7]、Last 等人提出的关键特征(如斜率和信噪比)法[8]、Guralnik 等人提出的字符表方法、Korn 等人提出的奇异值分解(SVD,Singular Value Decomposition)[9]法、Keogh 等人先后提出的分段累积近似法(PAA,Piecewise Aggregate Approximation)[10,11]、分段线性表示(PLR,Piecewise Linear Representation)[12-14]和适应性分段常数近似法 (APCA,Adaptive Piecewise Constant Approximation)[15]等分段方法，以及 Perng 等人提出的界标模型(Landmark Model)[16]等。这些表示方法各有所长，为时间序列相似性研究提供了诸多可以借鉴与参考的方向。

2 离散傅立叶变换

离散傅立叶变换(DFT,Discrete Fourier Transform)是最早被运用于时间序列的相似性提取方法[1]。在对时间序列数据的相似分析中，大多数人采用欧氏几何距离作为相似性计算的依据，因此所选用的方法多采用保持欧氏距离的正交变换法。离散傅立叶变换是一种十分常用的独立于数据的变换。一方面由于在时间域中两个信号的距离与频率域中的欧氏距离相等；另一方面因为 DFT 开头的几个系数表现十分突出，可以集中信号的极大部分的能量，因此可以通过保留 DFT 头几个系数来实现数据压缩，成功地计算出实际距离的下界。

自从 DFT 被 Agrawal 最早应用于时序数据相似性搜索后，又有其他一些论文相继提出了 DFT 的许多扩展和改进方法[17-20]，但核心思想并没有什么变化。DFT 算法的时间复杂度是 $O(n\log n)$，相比于点对点的比较的 $O(mn)$甚至是 $O(mn^2)$已经有了很大的提高。离散傅立叶变换的基本算法如下：

给定信号$\vec{x}=[x_t]$，$t=0,\ldots,n-1$，其 n 点的离散傅立叶变换可以定义为 n 个复数 X_f，$f=0,\ldots,n-1$，组成的序列$\vec{X}$：

$$X_f=\frac{1}{\sqrt{n}}\sum_{t=0}^{n-1}x_t\exp\left(-\frac{j2\pi ft}{n}\right)\quad f=0,1,\ldots,n-1 \tag{1}$$

其中，j 为虚单位，$j=\sqrt{-1}$。同时信号$\vec{x}$可以通过逆变换恢复如下：

$$x_t=\frac{1}{\sqrt{n}}\sum_{f=0}^{n-1}X_f\exp\left(\frac{j2\pi ft}{n}\right)\quad t=0,1,\ldots,n-1 \tag{2}$$

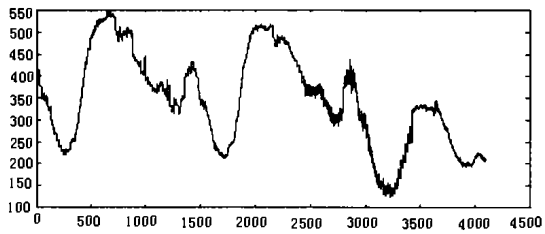


图 1 待处理的原始信号

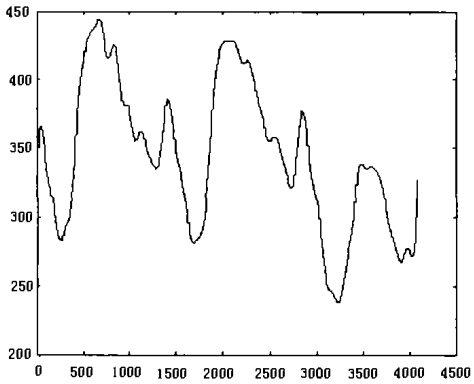
对于如图 1 所描述的一个原始信号，包含 4096 个数据点，对其进行快速傅立叶变换，保留变换结果的头 32 位和 64 位变换系数所恢复的图形分别如图 2(a)和(b)所示。

对比图 1 和图 2 所示结果，可以看出，一方面，DFT 所保留的系数越多，恢复图形中所保留的局部特征也就越多；另一方面，DFT 在数据截取的过程中，舍弃了信号的高频成分，平滑了信号的局部极大值和极小值，因而造成了信息的遗漏。

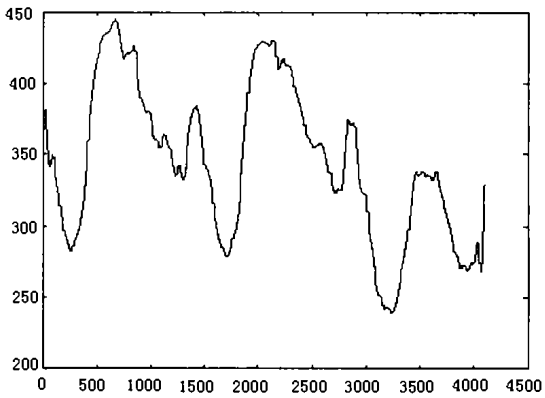
欧氏几何距离在经过 DFT 之后依然得到保持，所以可以用欧氏距离作为时间序列的相似性度量，即通过计算两序列差的平方和的平方根作为这两个时间序列的距离函数。序列$\vec{x}$和序列$\vec{y}$之间的距离具体表示如下：

$$D(\vec{x},\vec{y})=\left(\sum_{t=0}^{n-1}|x_t-y_t|^2\right)^{1/2}\equiv(E(\vec{x}-\vec{y}))^{1/2} \tag{3}$$

如果计算的结果小于一个由用户所定义门槛值 ε，则可以认为这两个时间序列是相似的。欧氏距离是一种较优越的距离估计的方法，尤其是在信号受到高斯噪声干扰的时候[1]。



(a)保留 32 位 DFT 系数



(b)保留 64 位 DFT 系数

图 2 经过 DFT 变换后压缩得到的图形

由于 DFT 变换具有保持欧氏几何距离、计算简便且能够把信号大部分能量集中到很少的几个系数当中等优点，所以用它来表示时间序列数据可以达到一定的要求。头几个系数集中的能量越多，方法也就越有效。但是 DFT 却平滑了原序列中局部极大值和局部极小值，导致了许多重要信息的丢失。此外 DFT 还对序列的平稳性有较高要求，对非平稳序列并不适用。分段 DFT 可以用来缓和这一矛盾，但是分段的方法同样也引入了一些新的问题。例如：分段过大导致判断力度的下降，分段过小又有低频建模的缺陷。因此在实际应用中，DFT 的方法尚存在较大的局限性。

3 离散小波变换

离散小波变换(DWT,Discrete Wavelet Transform)[3-7,21,22]和离散傅立叶变换同样是一种线性信号处理技术。当用于数据向量 D 时，将它转换成数值上不同但长度相同的小波系数的向量 D'。小波变换的数据压缩的实现同样是由数据裁减实现的，即通过仅存放一小部分最强的小波系数来保留近似的压缩数据。DWT 是一种较好的有损压缩，对于给定的数据向量，如果

DWT 和 DFT 保留相同数目的系数，DWT 能提供比原数据更精确的近似，更重要得是小波空间的局部性相当好，有助于保留局部细节。

DWT 在很多场合的应用中要比 DFT 更加有效[21,22]。例如：DWT 拥有时频局部特性，可以同时考虑时域和频域的局部特性，而不像 DFT 只考虑频域特性。

DWT 在很大程度上和 DFT 处理方法很类似，其基本函数由递归函数定义[23]：

$$\psi_{j,k}(t)=2^{j/2}\psi(2^j t-k) \tag{4}$$

其中，2^j 是 t 的缩放比例（j 是缩放比例以 2 为底的对数），$2^{-j}k$ 是在时域内的转换，且 $2^{j/2}$ 保持了 L^2（平方可积函数空间）在不同缩放比例的小波标准。因此任何 $L^2(R)$ 空间内的函数可表示为如下级数：

$$f(t)=\sum_{j,k} a_{j,k}2^{j/2}\psi(2^j t-k) \tag{5}$$

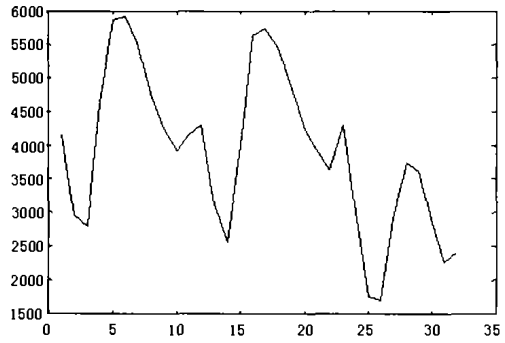
结合方程(4)，可化为：

$$f(t)=\sum_{j,k} a_{j,k}\psi_{j,k}(t) \tag{6}$$

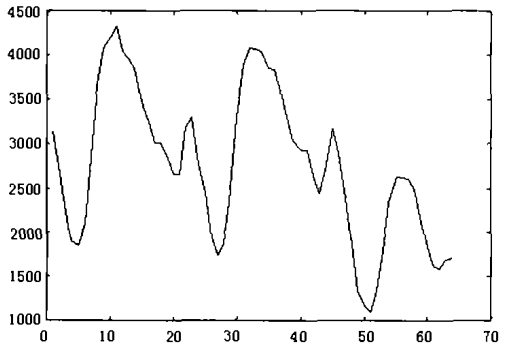
二维系数 $a_{j,k}$ 组成的系列被称为 $f(t)$ 的离散小波变换。$a_{t,j}$ 的计算方法可由内积表示为：

$$f(t)=\sum_{j,k}\langle\psi_{j,k}(t),f(t)\rangle\psi_{j,k}(t) \tag{7}$$

在实际应用中，总是希望能够用尽量小的存储空间实现更快的计算速度。于是 Harr 小波变换[23,24]最早被引入到时间序列的相似性研究中来。该方法得到了系数子集的良好近似，可以在保持欧氏几何距离的同时更加简便快捷地计算结果。对于图 1 所描述的原始信号，经过 Harr 小波变换分别保留 32 位与 64 位的图形如图 3 所示。



(a)保留 32 位 DWT 系数



(b)保留 64 位 DWT 系数

图 3　经过 DWT 变换后压缩得到的图形

仅从图 2 和图 3 的简单对比上来看，和 DFT 相比较而言，DWT 曲线不像 DFT 那么平滑，主要原因是由于 DFT 是通过在截取前几位变换系数后，再通过逆变换还原得到的曲线，构成曲线的元素依然是初始的 4096 个。然后经过几次 DWT 以后，每次保留的全局数据都是变换前数据个数的一半，因此用作描述曲线的元素个数只有保留下来的 16、32、64 个。因此曲线变化比较尖锐。实际上，离散小波变换可以更好地保留数据的局部特性。

小波变换在针对时间序列相似性研究中相比 DFT 并没有太大的优势。Yi-Leh Wu 就曾经验证过[22]DWT 并没有减少相对镜像误差也没有在相似性查询中提高查询的准确性。他认为在时间序列数据库的相似性搜索中，基于 DFT 和基于 DWT 的不同技术相比并没有太大的差异[22]。而且 DWT 无法处理任意长度的序列。而在实际应用中，始终分析一种长度的序列或是在索引中建立各种长度序列的架构显然也是不现实的，因此 DWT 在使用中还是存在很大的缺陷。

Keogh 提出过一种将 DFT 和 DWT 进行结合的方法[23]，Kawagoe 也曾经做过将 DFT 和 DWT 变换结合的方法[24]，这种结合集中了 DFT 和 DWT 各自的优点，可以在更快的时间内给出比单独使用 DFT 和 DWT 方法更好的结果。但是在不同方法的混合上，目前的研究还只是停留在启发式的阶段，在实际应用之前，还需要解决在执行中的具体困难。如果可行的话，这种混合的方法应该会优于单种方法的使用。

4　奇异值分解法

Korn 等人提出的奇异值分解(SVD，Singular Value Decomposition)[9]法在统计学里也称主分量分析(PCA，Principal Component Analysis)或 Kullback-Leibler 分解，是一种基于统计概率分布的投影方法。这种方法搜索 c 个最能代表数据的 k 维正交向量($c\leq k$)，使原来的数据被投影到较小空间，实现数据压缩。而且 SVD 的方法也曾经成功地运用在图像和其他多媒体目标的索引中。

SVD 和其他方法相比一个很重要的不同在于：其他变换的方法是局部的，它们在一时刻检查一个数据并且对数据进行变换。这些变换的完成和其他数据之间是相互独立的。相比之下，SVD 方法是全局的，整个数据库都被检测，并且生成几个最有可能代表原序列数据的相互正交的向量。变换的全局性既是 SVD 的优点同时也是这种方法的缺陷所在。

因此，在以下几种情况下 SVD 是比较理想的方法。例如：在对一系列数据进行 SVD 变换之后，如果希望重建这些数据，SVD 则是一种能够将重建误差控制在最小的一种线性变换。因此可以认为 SVD 可以较好地完成相似性索引的任务[7]。然而 SVD 作为索引架构的同时又存在一些缺陷，最严重的问题就在于它的复杂度。计算 SVD 变换的经典算法的时间和空间复杂度分别为 $O(mn^2)$ 和 $O(mn)$[7]。此外，每一次对数据库的插入和删除都需要对整个索引进行重新计算，因此，尽管近来对 SVD 的快速计算的算法研究有不少进展，但其计算量依然非常大，何况其中有很多快速算法也不可避免地引入了前面提到的不能接受的错误舍弃的可能性。

5　分段分析方法

为了进一步改进时间序列的表示方法，使之能够达到快速、准确、动态灵活的要求，Keogh 等人先后提出了分段累积近

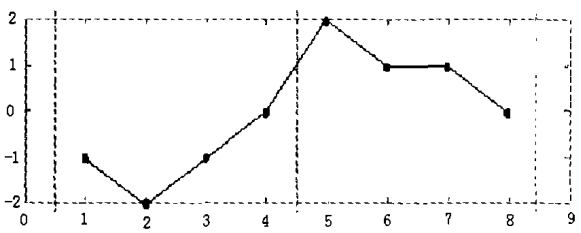
似法(PAA,Piecewise Aggregate Approximation)[10,11]、分段线性表示(PLR,Piecewise Linear Representation)[12-15]和适应性分段常数近似法(APCA,Adaptive Piecewise Constant Approximation)[16]等分段方法。这些方法首先将时间序列分为若干段,然后对每段取出平均值。同时改进了欧氏距离对于每一个点采取同等重视程度的方法,采用有权重的欧氏距离表示方法(weighted Euclidean distance),使方法的准确性和快速性得到了提高。并且具有可以处理任意长度的时间序列,允许持续时间的插入和删除操作,支持有比例的欧氏距离方法和短于所建立索引长度的查询等优点。分段处理思想对时间序列数据的表示有着比较重要的意义,在同以往的处理方法相比较中,体现了比较明显的优势。

假定有时间序列为 $X=x_1,\cdots,x_n$,另有一系列时间序列构成的时间序列数据库 $Y=\{Y_1,\cdots Y_K\}$。不失一般性,假定 Y 中每个序列的长度为 n,令 N 为变换空间的维数,且 $1\leq N\leq n$。假定 N 是 n 的一个因数(这一假定不是方法的使用条件,只是为了使符号简单化)。

长度为 n 的时间序列用 N 维空间向量$\bar{X}=\bar{x}_1,\ldots,\bar{x}_N$来表示。$\bar{X}$的第 i 个元素可以用以下方程来计算[10]:

$$\bar{x}_i=\frac{N}{n}\sum_{j=\frac{n}{N}(i-1)+1}^{\frac{n}{N}i}x_j \tag{8}$$

图 4 给出了分段数据维规约的示意图。简单而言,为了将数据从 n 维向量降为 N 维向量,数据先被分成了相等长度的 N 小段,然后分别计算出每一小段的平均值,这些数值组成的向量就是原来时间序列降低维数后的表示。当 $N=n$ 时,变换后的表示方法和初始表示形式是一样的。若 $N=1$,则所变换后的结果即为原数据序列的简单算数平均值。通常情况下,这种变换得到了原序列的分段常数近似表示,因此这种方法被 Keogh 称为分段累积近似(PAA,Piecewise Aggregate Approximation)。



$X=(-1,-2,-1,0,2,1,1,0)$ $n=|X|=8$

$\bar{X}=(\text{mean}(-1,-1,-1,0),\text{mean}(2,1,1,0))$ $N=|\bar{X}|=2$

$\bar{X}=(-1,1)$

图 4 分段累积近似数据维规约示意图

将分段累积近似的算法运用于图 1 所示信号并采取保留 64 个数据后的结果如图 5 所示。由图 5 的结果可以看出 PAA 算法和 DFT 一样也平滑了时间序列的局部特征。而且若是原始信号的变化频率越高,变化幅度越大,这种平滑作用也越突出,信息的遗漏和错误也越多。

在相似性度量方面,为了确保不会出现错误舍弃的现象,Keogh 在[7]中提出了一种定义在索引空间中的距离度量 DR,具有以下特性:$DR(\bar{X},\bar{Y})\leq D(X,Y)$。这一距离度量可以表示为:

$$DR(\bar{X},\bar{Y})\equiv\sqrt{\frac{n}{N}}\sqrt{\sum_{i=1}^{N}(\bar{x}_i-\bar{y}_i)^2} \tag{9}$$

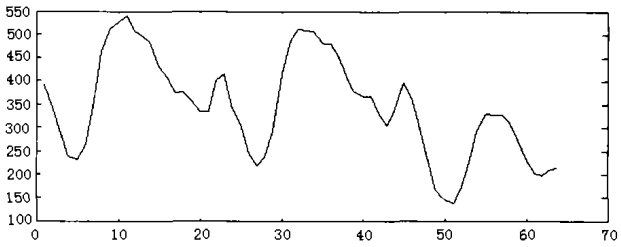


图 5 64 分段 PAA 恢复图示

由于可以避免出现前文所提到的错误舍弃的情况,并且具有准确度高、处理对象灵活等优点,PAA 也不失为一种比较好的处理方法。但是,为了方便表示法的索引,PAA 的每一部分都取了相同的长度。于是,Keogh 后来提出了适用于各部分长度任意的处理方法[8],即适应性分段常数近似法(APCA,Adaptive Piecewise Constant Approximation)。这种方法对每个部分都需要记录两个数字,分别记录该部分内所有数据点的平均值和这一部分的长度。

对于给定的时间序列 $\mathbf{C}=\{c_1,\ldots,c_n\}$,APCA 变换的方法可以表示如下:

$$C=\{\langle cv_1,cr_1\rangle,\ldots,\langle cv_M,cr_M\rangle\},cr_0=0 \tag{10}$$

其中,cv_i 是第 i 个部分数据的平均值,而 cr_i 则是第 i 部分的端点。

对于时间序列 $\mathbf{C}$,在经过 APCA 变换后为 C,并且有查询序列 Q。很明显,由于变换后的 C 所包含的信息量要少于 $\mathbf{C}$,所以在 Q 和 C 之间无法准确定义一个等价于欧式距离的距离度量。为此,Keogh 等人定义了两种 Q 与 C 之间的距离度量作为 $D(Q,\mathbf{C})$的近似[8]。一种是严格的欧式距离的近似 $D_{AE}(Q,C)$,另一种则是非严格的欧氏距离的近似 $D_{LB}(Q,C)$,这两种距离表示的具体定义如下:

$$D_{AE}(Q,C)=\sqrt{\sum_{i=1}^{M}\sum_{k=1}^{cr_i-cr_{i-1}}(cv_i-q_{k+cr_{i-1}})^2} \tag{11}$$

$$D_{LB}(Q',C)=\sqrt{\sum_{i=1}^{M}(cr_i-cr_{i-1})(qv_i-cv_i)^2} \tag{12}$$

其中:

$$Q'=\{\langle qv_1,qr_1\rangle,\ldots,\langle qv_M,qr_M\rangle\},qr_i=cr_i \tag{13}$$

且 $qv_i=\dfrac{\sum_{t=cr_{i-1}+1}^{cr_i}q_t}{cr_i-cr_{i-1}}$

图 6 和图 7 给出了这两种距离和欧氏几何距离相比较的直观表示图。

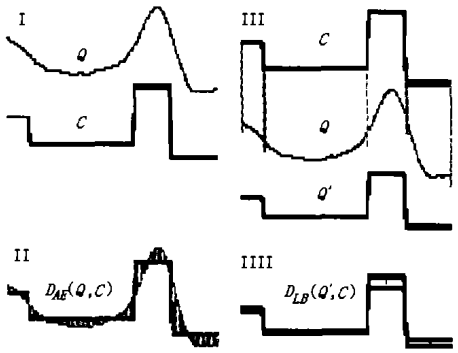


图 6 D_{AE} 和 D_{LB} 示意图

分段线性分割作为一种常用的表示方法有着许多的特点和优势。早在 1974 年,Pavlidis 和 Horowitz 就曾经指出,分段线

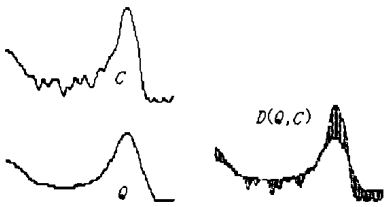


图7 欧氏距离的直观示意

性分割是一种很好的数据压缩和消除噪声的方法，Shatkay 和 Zdonik 则在 1996 年描述过线性(或高次多项式)分割在模糊查询中的应用。Keogh 和 Smyth 又进一步论证了利用线性分割在模式匹配中运用的可能性[12~14]。于是，分段线性表示(PLR，Piecewise Linear Representation)也就可能成为时间序列特征提取的重要方法。可以表述如下：

一个时间序列，在 k 点采样，用诸如 A 这样的大写字母表示。A 的线性分割包含 K 个分段，可以用粗体大写字母 A 表示，其中 A 是长度为 K 的向量组成的 5 维数组，如下式表示：

$$A \equiv \{AXL, AXR, AYL, AYR, AW\} \tag{14}$$

其中，序列 A 的第 i 部分则由 (AXL_i, AYL_i) 和 (AXR_i, AYR_i) 之间的直线表示，AW_i 即表示这一部分所占的比重。

在时间序列被分割成几个部分之后，起初设定的比重均为 1。如果比重发生了任何改变，比重就要重新标准化以使得每部分比重与其相应部分的长度乘积之和等于整个序列的长度，即以下方程总成立：

$$\Sigma_{i=1}^{K} W_i \times (AXR_i - AXL_i) = AXR_k - AXL_1 \tag{15}$$

比重的重新标准化是有着非常重要的意义的，它说明了不管用了多少个分段来表示，时间序列与每个给定长度序列相关的比重总和是不变的。因此，比重这个参数反应了该部分在整个时间序列中相对的重要性。在进行两个时间序列之间的相互比较的时候，因为每一部分的端点通常是不连续的，可将每个序列的每个端点都投影到另一个序列上，并且测量出投影线的距离，并将这一距离作为相似性度量。并且这种度量具有消除噪声、消除变换偏移量、控制幅度缩放比例等优点[12,14]。

在一般数据库的处理中，PLR 的近似方法均不会差于 DFT，但是在高频率领域内，分段线性分割由于采用许多短的分段造成了表示方法和近似程度十分粗糙，因此压缩率很低。DFT 在高频谱领域内相比 PLR 有着独特的优势[14]。

6 界标模型

界标模型(Landmark Model)由 Perng 等人最先提出，是一种集相似性模型和数据模型为一体的方法[16]。该方法借鉴了研究人员对人类和动物在模式识别过程中的发现。认为如果两个由一些转折点和连接这些折点的曲线所组成的图表拥有相似的转折点，就可以认为这两幅图表是相似的[16]。

Perng 等人将时间序列中一些最为重要的点定义为界标(Landmark)，其研究的主导思想就是使用界标来替代需要处理的原始数据序列。在不同的应用场合，界标的定义也有所不同，可以是简单属性(如局部极大值或极小值、拐点等)，也可以是复杂结构。如果将曲线 n 阶导数为 0 的点称为曲线的 n 阶界标[16]，则局部极大、极小值点则是曲线的一阶界标，拐点则为二阶界标。在相似性比较中，高阶界标并没有太大的作用，它仅仅对时间序列的变化有微小的影响，可以通过舍弃来实现数据的压缩。仅从定义上分析就可以发现，界标模型并不像 DFT 和 DWT 一样，将时间序列的局部极大和极小值点平滑处理了，而是基本上保留了这些对于数据特征有着重要意义的点。对图 1 所示信号进行处理结果如图 8 所示。可以看出，所获取的信号几乎保留了原信号所有局部特征，但是经过处理以后的时间序列仍然有 2335 个数据点，较之处理之前的 4096 个数据而言，压缩率还不到一半，在降维的有效性这一点上，这种方法还不如 DFT、DWT 又或是 PAA。

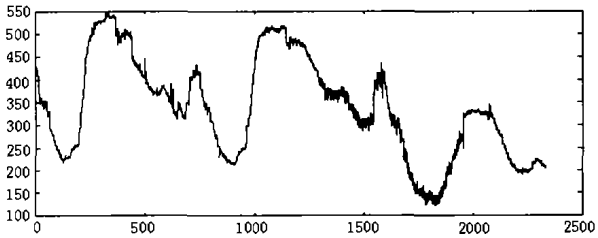


图8 Landmark 提取结果示意图

为解决这一问题，它以最小距离/百分比规则(MDPP，Minimal Distance/Percentage Principle)作为平滑方法来消除噪声干扰。但是界标模型相比其他方法存在着一些问题，一方面要对数据进行二次处理，增加计算的复杂度。而且由于最后得到的压缩结果长度无法预计，在相似性度量上有专门的要求，在实际应用中还存在一些问题有待解决。

相似性模型中的变换方式越多，相似性模型的功能也就越大。在大多数研究中，人们一般采用 2~3 种变换。而 Perng 等人提出了一致平移变换(SH)、一致幅值缩放(UAS)、一致时间缩放(UTS)、一致时间幅度缩放(UBS)、时间归整(或非一致时间缩放)(TW)和非一致幅度缩放(NAS)六种变换方式。除这六种基本变换外，还可以通过结合基本变换来实现复杂变换。这些变换的引入并非单纯为了执行这些变换，更大程度上是为了拓展相似性的语义。界标序列的相似性可表示如下：

给定两个界标序列 $L=\langle L_1,\ldots,L_n\rangle$ 和 $L=\langle L'_1,\ldots,L'_n\rangle$，其中 $L_i=(x_i,y_i)$ 且 $L'_i=(x'_i,y'_i)$，则第 k 个界标之间的距离可定义为：

$$\Delta_k(L,L') = (\delta_k^{time}(L,L'), \delta_k^{amp}(L,L')) \tag{16}$$

$$\delta_k^{time}(L,L') = \left\{ \begin{array}{ll} \dfrac{|(x_k - x_{k-1}) - (x'_k - x'_{k-1})|}{(|x_k - x_k| - |x'_k - x'_{k-1}|)/2}, & \text{如果 } 1 \le k \le n \\ 0, & \text{其他} \end{array} \right\}$$

$$\delta_k^{amp}(L,L') = \left\{ \begin{array}{ll} 0, & \text{如果 } y_k = y'_k \\ \dfrac{|y_k - y'_k|}{(|y_k| - |y'_k|)/2}, & \text{其他} \end{array} \right\}$$

由此，可得出两序列之间的距离为：

$$\Delta(L,L') = (\|\delta^{time}(L,L')\|, \|\delta^{amp}(L,L')\|) = (\delta^{time}, \delta^{amp}) \tag{17}$$

其中，$\|\cdot\|$ 为向量的范数。

另外，界标模型在数据表示上也很有特点。对于给定的界标序列 $L_1,\ldots L_n$，其中 $L_i=(x_i,y_i)$，用一个小的特征集合 $F=\{y,h,v,hr,vr,vhr,pv\}$[3] 为例，其中：

$h_i = x_i - x_{i-1}, v_i = y_i - y_{i-1}, hr_i = h_{i+1}/h_i$

$vr_i = v_{i+1}/v_i, vhr_i = v_i/h_i, pv = v_i/y_i$

以上所有的特征均由界标的坐标产生，但是每个特征都具有不同的性质。有些特征在一些变换后是保持不变的，在各种变换下这些不变的情况可以用表 1 来描述。

通过表 1 所示各种变化可以看出，并不是每种变换的组合都是有意义的。因为当时间序列分段足够长的时候，有可能出

现过度变换使相似关系成为了完全的关系(complete relation),也就是说每一个序列都是彼此相似的。例如:如果时间序列足够长,那么该时间序列经过 TW 和 NAS 变换变成任何形状,这样也就引出了一个在变换方法结合上的一些规定,尽管 Perng 等人也给出了一些规定,但是也给实际操作带来了不变。

表 1 界标模型变换后各特征是否变化的情况

	y	h	v	hv	vr	vhr	Pv
SH		*	*	*	*	*	
UAS		*		*	*		*
UTS	*		*	*	*		*
UBS				*	*	*	*
TW	*		*		*		*
NAS		*		*			

7 总结与展望

从以上对各种时间序列表示方法的描述可以看出一个大致的发展方向及趋势,即时间序列的相似性问题不再单纯是两个时间序列之间的相互关系。从最初的时间序列相似性的点对点的比较,到后来的对时序数据进行相应的 DFT 和 DWT 变换,再有对时间序列的各种分段后的处理,或是界标模型又甚至是将几种变换方法的结合使用,时间序列表示方法的选择已经具有了相当的主观性。它可以是主观选定的一种线性变换,如 DFT、DWT 等;也可以是非线性变换(如 Dynamic Time Warping);甚至可以主观地选取一些认为可以替代原序列的序列特征和关键点,如分段处理、关键特征、斜率和界标等,然后配合以良好的相似性度量函数。

虽然所采用的时间序列表示算法的选取是具有一定的主观性的,但是并不是任何一种变换或者方法都适用于时间序列的表示。作为一种好的表示方法,应该具备以下重要特征:

(1)准确性。不管采取任何一种方法或者变换,都必须尽可能减少在变换过程中的信息遗漏,在能够描述序列变化趋势的同时尽可能准确地描述信号的局部特征。

(2)快速性。考虑到面对大型时间序列数据库的相似性搜索问题,对表示算法的时间复杂度有着较高要求。

(3)一致性。由于时间序列表示算法的目的是要进行相似性比较,因此要求变换后序列和变换前原始序列在相似性度量上具有一致性,不应改变同样时间序列之间的相互关系。

(4)降低原序列维数。造成时间序列相似性比较复杂有一个很重要的原因就是时间序列的高维特征。出于对计算的时间复杂度和存储的空间复杂度两方面的考虑,表示算法应当有良好的降维特性。

在具体选择表示算法的时候,尽量实现算法复杂度、降维的有效性、局部特征和全局特征提取等各方面的协调。另外,考虑到表示方法的直观性,借鉴人类在波形模式识别中所具有的由粗到精、由全局到局部的感知能力及其特点,将时间序列数据看作一个波形,依照感知过程中的重要程度逐渐找出波形中的若干重要的点,同时考虑到分段分析在时间序列局部细节描述的长处,采用"自顶向下"的方法将整个波形逐渐分段线性化,从而建立一种基于感知的时间序列多分辨近似表示模型并实现其快速算法,将是一种可行的思路。(收稿日期:2004 年 2 月)

参考文献

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

2.R Agrawal,K I Lin,H S Sawhney et al.Fast similarity search in the presence of noise,scaling,and translation in times-series databases[J].VLDB Journal,1995:16~23

3.K P Chan,A W Fu.Efficient time series matching by wavelets[C].In:Proceedings of the 15th IEEE International Conference on Data Engineering,1999:126~133

4.Z Struzik,A Siebes.The Haar wavelet transform in the time series similarity paradigm[C].In:Proceedings of the 3rd European Conference on Principles and Practice of Knowledge Discovery in Databases,1999:12~22

5.F K Chan,A W Fu,C Yu.Haar wavelets for efficient similarity search of time-series:with and without time warping[J].Knowledge and Data Engineering,IEEE Transactions on,2003;15(3):686~705

6.I Popivanov,R J Miller.Similarity search over time-series data using wavelets[C].In:Data Engineering,2002,Proceedings,18th International Conference on,2002:212~221

7.B Audit,E Bacry,J F Muzy et al.Wavelet-based estimators of scaling behavior[J].Information Theory,IEEE Transactions on,2002;48(11):2938~2954

8.M Last,Y Klein,A Kandel.Knowledge discovery in time series databases[J].IEEE Transactions on Systems,Man and Cybernetics,2001;31(B1):160~169

9.P Korn,N Sidiropoulos,C Faloutsos et al.Fast nearest-neighbor search in medical image databases[C].In:Proceedings of 22th International Conference on Very Large Data Bases,Bombay,India,1996:215~226

10.E Keogh,K Chakrabarti,M Pazzani et al.Dimensionality reduction for fast similarity search in large time series databases[C].In:Proceedings of the ACM SIGMOD International Conference on Management of Data,2001:151~162

11.E Keogh,M Pazzani.Scaling up dynamic time warping for datamining Applications[C].In:Proceeding of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,2000:285~289

12.E Keogh,M Pazzani.An enhanced representation of time series which allows fast and accurate classification[C].In:clustering and relevance feedback,Proceedings of the 4th International Conference of Knowledge Discovery and Data Mining,AAAI Press,1998:239~241

13.E Keogh,P Smyth.A probabilistic approach to fast pattern matching in time series databases[C].In:Proceedings of the 3rd International Conference of Knowledge Discovery and Data Mining,AAAI Press,1997:24~20

14.E Keogh,M Pazzani.An indexing scheme for fast similarity search in large time series databases[C].In:Scientific and Statistical Database Management,Eleventh International Conference on,1999:56~67

15.E Keogh,K Chakrabarti,S Mehrotra et al.Locally adaptive dimensionality reduction for indexing large time series databases[C].In:ACM SIGMOD 2001,Santa Barbara,California,2001

16.C S Perng,H Wang,S Zhang et al.Landmarks:A new model for similarity-based pattern querying in time series databases[C].In:Data Engineering,Proceedings,16th International Conference on,2000:33~42

17.C Faloutsos,M Ranganathan,Y Manolopoulos.Fast subsequence matching in time-series databases[C].In:Proceedings of the 1994 ACM SIGMOD International Conference on Management of Data,1994:419~429

18.D Refiei,A Mendelzon.Similarity-Based queries for time series data[C].

In:Proc ACM SIGMOD Conf,1997:13~25
19.B D Ripley.Pattern recognition and neural networks[M].Cambridge University Press,1996
20.D Refiei.On similarity-based queries for time series data[C].In:Data Engineering,Proceedings,15th International Conference on,1999:410~417
21.K P Chan,A W Fu.Efficient time series matching by wavelets[C].In:Proceedings of the 15th IEEE International Conference on Data Engineering,1999:126~133
22.Y L Wu,D Agrawal,A E Abbadi.A comparison of DFT and DWT based similarity search in time-series databases[C].In:Proceedings of the Ninth International Conference on Information Knowledge Management CIKM 2000,2000
23.E Keogh,S Chu,M Pazzani.Ensemble-Index:A new approach to index large databases[C].In:SIGKDD '01,San Francisco,CA,2001
24.K Kawagoe,T Ueda.A similarity search method of time series data with combination of Fourier and wavelet transforms[C].In:Temporal Representation and Reasoning,TIME 2002,Proceedings Ninth International Symposium on,2002:86~92
25.G Guimaraes.Temporal knowledge discovery for multivariate time series with enhanced self-organizing Maps neural networks[C].In:IJCNN 2000,Proceedings of the IEEE-INNS-ENNS International Joint Conference on,2000;(6):165~170

(上接21页)

的分馆信息。服务器端通过内部程序解析并检验其访问权限后,调用自己相应的接口函数,并通过其接口函数与数据库相联系并获取返回值,将返回值通过XML-RPC封装后再返回给办证业务子系统[2,3]。其他各个业务子系统也都通过此种方式与应用服务器进行通信。

2.4 XML-RPC应用与一卡通系统的技术难点

由于该系统涉及数据库繁多,办证子系统底层为access数据库、读者用机子系统用到SQLSERVER2000数据库、借阅业务子系底层为oracle数据库等等。所以数据统一性及数据安全十分关键。该系统的解决方案如下。

(1)数据统一性方面

系统主要数据流为读者用户的卡号、密码、姓名、年龄、令牌号、权限和金额数据和操作信息。系统服务器数据库为IBM的DB2数据库、mysql数据库和ldap认证数据库。其用户的基本信息(卡号、密码、姓名、年龄、令牌号、权限)放在DB2和ldap数据库中一式两份,用户消费的金额数据放在DB2数据库中,各种操作信息放在mysql数据库中。其上三个数据库中数据以用户卡号相联系。各个业务系统的用户信息都通过XML-RPC协议从服务器的数据库中获得,金额和用户操作信息通过XML-RPC协议传送到服务器端。服务器端建立操作字典。各业务子系统通过XML-RPC协议调用的接口函数写入操作信息。虽然各个业务系统的数据库与服务端数据库可能完全不同,但各个业务系统中的数据经过XML数据格式的封装可以写入到与其数据库结构完全不同的服务器数据库中。因为XML数据格式是一致的,各个业务子系统的数据通过应用服务器的处理使得和服务器数据库上的数据保持了一致性,从而各个业务子系统的数据在总体上是一致的。

(2)数据的安全性方面

由于一卡通系统涉及大量金额操作,其系统宗旨也是抛弃现金交易,进行电子交易,所以数据的安全性十分重要[1]。应用服务器对各个业务系统分配令牌,令牌与IP地址相关,令牌定时失效。通过令牌,应用服务器能对各个业务子系统进行IP地址和权限控制,使得不在IP地址范围内的主机无法与服务器相连,使得无权限的操作无法得到执行,此外各个业务子系统也不可能不经过应用服务器与服务器数据库进行直接数据通信。该系统对XML-RPC传输中的一些关键信息如:ip地址信息、权限、密码、金额信息进行MD5码加密。

```
<?xml version="1.0" encoding="UTF-8"?>
<Price Sheet>
<user name="江思能"><process name="上网"><Price>
<EncryptedElement name="上网费"algorithm="DES/CBC"contentType="text/xml"encoding="MD5">vJqNpDrQT1vmCVb</EncryptedElement>
<ValidTime>2001-1-1 </ValidTime></Price></user>
..... ..... .....
..... ..... .....
</Price Sheet>
```

通过MD5加密使用户的消费金额信息有了很好的安全性。

3 应用展望

系统通过国家图书馆、天津泰达图书馆、安徽省图书馆的实际使用,验证了系统的可靠性和技术的先进性。该系统不仅仅应用于图书馆的一卡通系统,系统结构可用于小区、校园,甚至银行一卡通系统中。该系统使先进的服务器设计架构与一卡通相结合,使一卡通系统更具有通用性。此外该系统架构和传输数据模式还可用于其它的多分布式系统。

4 结束语

“一卡走天下”是时代发展的需要,但现有的许多一卡通系统存在着许多设计上和技术上的局限性,限制了其发展,而大多数一卡通系统设计主要从卡片的使用和硬件设备上进行讨论。该文抛弃了传统的一卡通服务器的设计方式,结合当今时兴的XML技术提出了一种新型的一卡通设计框架和一种基于XML-RPC协议的数据传输理念。提出了一种新型的一卡通服务器应用程序设计框架,此架构能解决各个业务系统和服务器数据不一致、数据安全性差、数据传输不便的数据瓶颈问题,并对这种新型的数据传输技术进行探讨。使一卡通系统不在为其业务系统繁多、数据量的混乱所烦恼,并使其总体上拥有更好的性能。(收稿日期:2004年6月)

参考文献

1.李维.Dephi5.0分布式多层应用系统篇[M].北京:机械工业出版社,2000
2.Matjax B J.J2EE EAI编程指南[M].北京:电子工业出版社,2000
3.Subcahmanyam A,Cedric B,John D.J2EE编程指南[M].北京:电子工业出版社,2000
4.陈天华.基于IC卡的交通管理系统[J].工业仪表与自动化装置
5.天喻城市一卡通应用案例.武汉天喻信息产业有限公司,2001
6.Simon St Laurent,Edd Dumbill,Joe Johnston.Programming Web Services with XML-RPC(O'Reilly Internet Series)[J].IBM System Journal,2001;35(3~4):311~325
7.David A Chappell,Tyler Jewell.Java Web Services[J].IBM System Journal,2002;33(1~2):303~434