第26卷 第4期 河北理工学院学报 Vol. 26 No. 4
2004年11月 Journal of Hebei Institute of Technology Nov. 2004

文章编号:1007-2829(2004)0-0040-03

变换域时间序列相似性搜索

魏 莲

(桂林工学院 电子与计算机系,广西 桂林 541004)

关键词:时间序列;小波变换;字符串;模式匹配

摘 要:在时间序列的研究中,经常需要计算二个序列的相似程度。由于序列变化的多样性和复杂性,结果通常不能很好地满足要求。采用变换法则对时间序列进行从时域到频域的转换,再将转换后的数据按照一定的规则变换成字符序列;利用求最长公共子序列的方法计算二个序列的匹配度,实现时间序列的相似性搜索。

中图分类号:TP 311 文献标识码:A

0 引言

时间序列是指按时间顺序排列的观测值的集合,目前,在许多行业数据库中都存在着大量的时间序列数据,对时间序列的相似性搜索就是在这一类数据中发现与给定序列模式相似的序列,或是发现二个序列的相似度有多大。如地质勘探中的地层对比、股票数据的趋势分析等。在时间序列的相似性研究中,有基于距离表示的相似性算法,有基于形态表示的相似性算法。序列由于变化频率的不同,或由于本身变化的复杂性,直接比较很难得到好的效果,而研究其频谱关系要比直接研究序列本身要简单方便得多。

本文通过将序列进行一定的变换,通过阈值截取的方法将数值序列转换为字符序列,采用求最大公共子串的方法对序列的相似性进行定义。

1 小波变换法

小波分析是一种时-频信号分析法,它是信号函数与小波函数的内积。在小波变换中,小波函数的选取至关重要,小波函数具有以下性质:

$$\int_R \psi(x)\,dx = 0$$

$\psi(x)$在有限区间外很快趋于零,这一特点使得小波具有"窗口"的作用,因此,小波变换具有较好的局部性。

正是因为小波变换具有在时域和频域局部化的特点,它为信号的分析与处理提供了强有力的新手段。小波变换通过平移小波获得信号的时间信息,通过缩放小波的宽度来获得信号的频率特性,通过平移和缩放得到一系列的小波系数,这些系数代表了小波和局部信号之间的相互关系,利用这些系数所反映的信号之间相互关系,可以对信号进行更深层的和不同角度的分析研究。

如图1所示,原始时间序列经小波变换,转换成小波系数。

采用小波变换将原始序列分解成不同的频率区换的特点是,在空间域将信号分解为不同的层次,在分解运算的同时形成了频率域中的多层次分解。

对于时间序列的匹配来说,在很多情况下,需要考虑序列数值的大小,更多关心的是序列的形态、变化的方向,小波变换后的序列能帮助发现序列隐藏的潜在趋势。

收稿日期:2004-06-28

作者简介:魏莲(1970-),女,桂林工学院电子与计算机系讲师。

3　基于字符串的模式匹配算法

时间序列问题中,对序列的分析通常是以时间为轴进行分析的,但由于序列变换的多样性,某一序列相对其他序列可能存在部分序列段相差甚远,而其余部分相似程度很高的情况,或是序列中存在时间段部分缺失的情况,致使二者不能很好地匹配。

有一种字符串匹配算法可以很好地用于解决该问题。

给定两个字符序列 X 和 Y,其长度分别为 n 和 m,怎样衡量二者有多像? 也就是说,要确定二者携带的信息中有多少是相同的? 为此,引出子序列与公共子序列的概念。

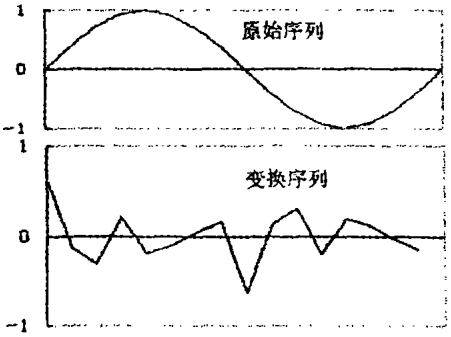

图1　时间序列变换分析

一个给定序列的子序列是指在该序列中删去若干元素后得到的序列[3]。

对于序列 X 和 Y,当另一序列 Z 既是 X 的子序列又是 Y 的子序列时,称 Z 是序列 X 和 Y 的公共子序列。两个序列的公共子序列可能会有多个,具有最大长度的公共子序列为二者的最长公共子序列。

求取最长公共子序列对于序列匹配具有相当重要的意义。当两个序列具有足够长的非重叠的相似子序列时,则认为此两序列是相似的。

例如,若序列 $X=\{a,b,c,b,d,a,b\}$,$Y=\{b,d,c,a,b,a\}$,则序列 $\{b,c,b,a\}$ 是 X 和 Y 的一个公共子序列,且为 X 和 Y 的最长公共子序列,因为 X 和 Y 没有长度大于4的公共子序列。

求最长公共子序列的方法对于时间序列的相似性搜索具有较好的适用性。它允许两个序列存在局部不匹配的情况,也允许一个序列相对另一个序列存在部分缺失,只要这种缺失的数量在允许的范围,则认为序列是相似的,如图2所示。

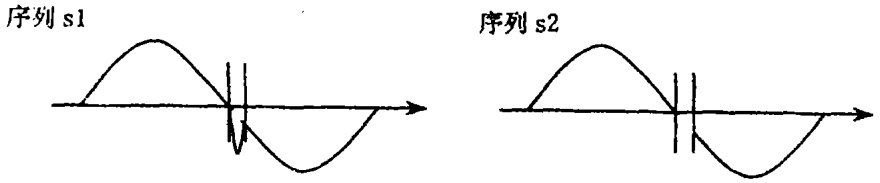

图2　序列存在部分缺失的情况

序列相似性定义:在图2中,序列 $S1$ 和 $S2$ 存在长度为 n 的最长公共子序列,如果 n 相对于序列 $S1$、$S2$ 的长度满足某一给定的条件,则认为 $S1$ 和 $S2$ 是相似的。

将最长公共子序列算法用于时间序列相似性搜索中。对于待匹配的数字序列,可以通过阈值截取将它们转换为字符序列。

3　实例分析

采用小波变换法将时序数据转换到频率域,对数据变换角度进行分析,小波变换将数据分解为频率域上不同层次的特征数据,分析这些数据可以帮助发现序列中隐藏的一些规律,其相似性搜索步骤如图3所示。

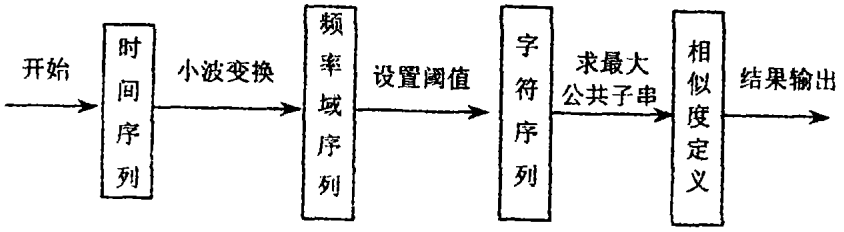

图3　基于变换的时间序列相似性搜索流程图

本文对图2中的 $s1$ 和 $s2$ 序列采用小波变换方法进行从时间域到频率域的变换,变换后的序列分别对

应图4中的(a)和(b)。

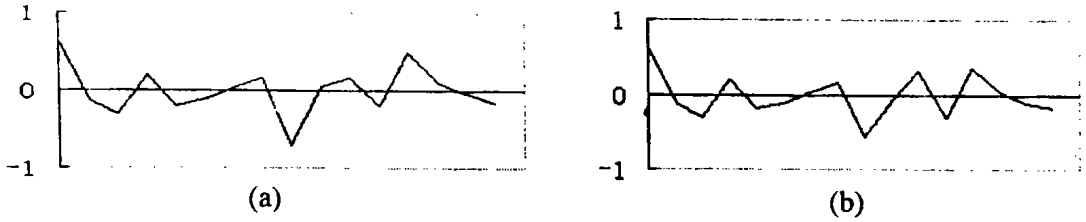

图4 变换时间序列图

对于图4中的变换序列,设域值$\beta=\frac{1}{n}\sum_i \alpha_i/\sqrt{2}$[$-\beta$,0]区间的数字用'$a$'表示,[$-0$,$\beta$]区间的数字用'$b$'表示,[$\beta$,$\infty$]区间的数字用"$c$"表示,($-\infty$,$-\beta$)共间的数字用"$d$"表示,则$sl$可表示为字符序列"*cbbabbaadaabaabb*',$s2$可表示为字符序列'*cbbabbaadbabaabb*'。二者的最长公共子序列的长度为15,定义序列匹配度$\xi=\frac{n}{(a+b)/2}$,可求得$\xi=0.94$。

4 结论

本文采用变换域方法对时间序列进行相似性搜索。通过小波变换实现序列从时域到频域的变换,将变换后的序列转换成字符序列,利用最长公共子序列的长度来判断两个序列是否相似。

由于小波的窗口作用和字符匹配的特点,算法对于序列中存在的时间段缺失、部分序列不匹配或序列变换频率不一致的情况具有较好的匹配效果。但窗口内小波变换产生的振荡现象暂时还无法消除。在序列匹配度的定义中,充分考虑到子序列对于两个原始序列的相似性。对于同一问题不同层次的分析需求,可以通过设立多级阈值或调节小波变换尺度来满足不同精度的要求。由于阈值的设立与序列最终变换的值相关,一定程度上消除了不同序列刻度不一致的情况,将序列幅值的影响因素减到最小。

参考文献:

[1] 李世雄. 小波变换及其应用[J]. 高等数学研究,2002(3).

[2] 杨敏,王志坚. 时间序列相似性搜索算法研究[J]. 山东师范大学学报,2001(12).

[3] 王晓东. 算法设计与分析[M]. 北京:清华大学出版社,2003.

Time series similarity search based on transformation

WEI Lian

(Dep. of Electronics and Computer, Guilin Institute of Technology, Guilin Guangxi 541004, China)

Key words: time - series; wavelet transform; string; pattrern matching

Abstract: In the time - series searching, similarity of the two series is always involved. Because of the diversities and complication of the series, the results cannot meet the needs well. This paper proposes a method by translating the series from time domain to frequency domain, then to a string sequence. By calculating the longest common subsequence, finally, similarity searching in time - series data sets is realized.