基于小波概要的区间差分 skyline 研究

程文聪[1] 邹 鹏[2] 贾 焰[1]

(国防科技大学计算机学院 长沙 410073)[1] (装备指挥技术学院 北京 101416)[2]

摘 要 在很多应用中需要分析大量的时序数据,而相对于其它数据具有支配优势的时序数据片段往往会引起特别的关注。基于量值度量,现有的区间 skyline 查询可以返回给定时间区间内所有没有被其他数据支配的时序数据,这种查询有时不能满足应用的需求,且可能存在"淹没"现象。为此提出了区间差分 skyline 的概念,针对数据增长率属性进行分析,以解决现有区间量值 skyline 的不足。目前很多时序数据呈现为数据流的形式,由于资源的限制往往只会维护一个反映数据概况的概要结构,在此背景下提出了基于常用的小波概要支持不同粒度区间差分 skyline 查询的基本算法,继而在保证准确性的基础上提出了改进后的快速算法。在真实股票价格数据集上的实验验证了所提方法的有效性。

关键词 时序数据,区间差分 skyline,小波概要

中图法分类号 TP311 **文献标识码** A

Research on Interval Differential Skyline Based on Wavelet Synopsis

CHENG Wen-cong[1] ZOU Peng[2] JIA Yan[1]

(School of Computer Science, National University of Defense Technology, Changsha 410073, China)[1]

(Academy of Equipment Command & Technology, Beijing 101416, China)[2]

Abstract In many applications, we need to analyze a large number of time series. Segments of time series demonstrating dominating advantages over others are often of particular interest. Based on volume measure, the current interval skyline query returns the time series which are not dominated by any other time series in the interval. Some times this kind of query can not satisfy application requirements, and the "submerge" phenomenon may exist. So we proposed the concept of the interval differential skyline which focusing on the attribute of increasing rate of data to fix the shortage of the former kind of interval skyline query. Currently most of the time series are generated as data streams. Due to the limitation of the resource, people only maintain synopses which describe the main data characters. In this background we proposed the algorithm to implement the interval differential skyline query in different granularities based on the common used wavelet synopsis and then we improved the efficiency of the naïve algorithm on the basis of keeping the accuracy of the results. Extensive experiments on the real stock price data set demonstrate the effectiveness of the proposed methods.

Keywords Time series, Interval differential skyline, Wavelet synopsis

随着监控能力和自动化水平的不断提高,越来越多的领域中产生了大量随时间不断变化的时序数据。这些时序数据表现为数值的形式且在每个时间点上(按特定的时间粒度)均存在一个值,如零售业中各种商品按月销售量的记录、股市中各种股票按天市值价格的记录、网络中部署的安全设施检测到的各种安全事件按小时数量的记录等。

为了能从多个时序数据中找出在指定时间区间内较为活跃的时序数据(这些数据往往是人们感兴趣的数据),近期出现了关于区间 skyline(interval skyline)的相关研究[1]。该研究的目标是从多个时序数据中发现所有在指定时间区间内没有被其他任何数据在量值上支配的时序数据(为了区别于本文的研究,称其为基于量值度量的区间 skyline)。图 1 是其研究目标的一个实例说明。如果一个电力供应商需要分析服务范围内不同地区的电力消耗情况,各个地区在不同时间点上的电力消耗形成了多个时序数据,基于量值度量的区间 skyline 查询可以解决如下的问题:在 5 月 2—9 日的时间区间之内,哪些地区有过最大的电力消耗?由于地区 3 在给定的时间区间内电力消耗总量最大,而地区 2 在 5 月 6 日(处于查询区间之内)有最大的电力消耗,因此基于量值度量的区间 skyline 查询结果为地区 2 和地区 3。

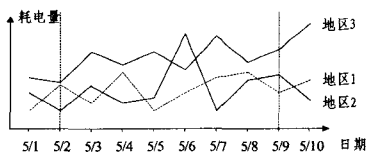

图 1 基于量值度量的区间 skyline 示例

基于量值度量的区间 skyline 能够满足一些在多时序数据上进行选择分析的需要,但存在以下两个问题:(1)在很多

到稿日期:2009-12-23 返修日期:2010-03-09 本文受国家 863 项目(2007AA01Z474,2007AA010502,2006AA01Z451)资助。

程文聪(1981—),男,博士生,主要研究方向为网络安全及时序数据分析,E-mail:emailtocheng@yahoo.com.cn;邹 鹏(1957—),教授,博士生导师,主要研究方向为网络信息安全和分布式计算等;贾 焰(1961—),女,教授,博士生导师,主要研究方向为网络信息安全、数据库和数据挖掘。

涉及多个时序数据分析的应用中，时序数据的变化率是更值得关注的一种重要度量。如在话题分析中，除了涉及话题网页的点击量值属性外，话题的热度（即网页点击的突发变化）也是一个评估话题价值的重要指标[2]；在选择股票的分析中，人们更关心哪些股票具有最大的价格涨幅，而股票市值本身却相对不那么重要；在网络安全分析中，各种病毒、木马、扫描事件在一定时期的增长率是威胁程度的重要度量，而基于量值度量的区间 skyline 不能满足这种分析需求；(2)使用量值作为从多个时序数据中选择感兴趣数据的标准时，容易发生某个（或少数几个）量值较大的时序数据长期"淹没"其他数据的情况，使得其他时序数据即使发生较大的异常变化（这种情况下的变化部分往往蕴藏了重要的领域信息），也不会成为查询结果，从而使区间 skyline 查询在很大程度上失去实际意义。如图 2 所示，如果地区 3 的耗电量一直较大，但没有出现很大的变化，则基于量值度量的区间 skyline 查询返回的地区 3 并没有很大的信息量和决策参考价值，而其它地区（如地区 2）在 6 日时的变化较大，明显具有继续深入分析的价值，但却被地区 3 所"淹没"，这种情况在很多应用中出现的概率很大。

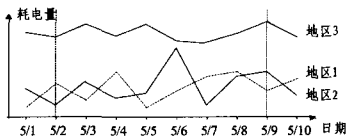

图 2　基于量值度量的区间 skyline 不适用的示例

基于上述原因，本文提出基于差分度量的时序数据区间 skyline，即将区间 skyline 的度量由单个时间点上的量值变为前后两个时间点量值的差分值（下面分别将其简称为区间量值 skyline 和区间差分 skyline）。区间差分 skyline 还有如下良好的性质而使其在分析多时序数据中的活跃数据时更具意义：如果时序数据集合 S 中的某个时序数据 s 在区间$[i:j]$的任意一个子区间上具有最大线性回归斜率（由最小二乘法计算），则 s 一定属于$[i:j]$上定义的区间差分 skyline 集合。该性质作为定理 1 将在后面加以证明。

在大量的应用中，时序数据的产生往往表现为数据流的形式。由于资源的限制，处理这种流数据时一般不会保存所有的原始数据，而只是维护一个在总体上反映时序数据特征的概要结构，因此不能采用直接根据原始序列求取区间差分 skyline 的方法，而只能通过离散逆小波变换求得原始序列后再进行计算，这种做法的时间开销较大。本文讨论了如何在时序数据流中常用的小波概要支持下处理指定时间粒度上的区间差分 skyline 查询的问题，并利用小波变换细节参数的差分属性提出了一种快速完成区间差分 skyline 查询的方法。

1　理论基础

1.1　skyline 和区间 skyline

skyline 的概念首先由 Börzsönyi 等[3]引入，可以在难以确定评分函数的情况下为多目标决策问题提供支持。skyline 问题在数据库和数据分析研究领域受到了广泛的关注，产生了大量有价值的研究结果，如其中著名的结论[4]：任意在各个属性上单调变化的评分函数所产生的 top-1 查询结果必属于其 skyline 集合。而相对地，每个 skyline 集合中的点也必为至少一种评分函数所确定的 top-1 查询结果。skyline 的相关形式化定义如下。

定义 1（支配关系）　对于具有多维可度量属性的点 p 和点 p'，如果 p 在所有维度上均优于或等于 p'，并且在至少一个维度上严格优于 p'，则称 p 支配 p'，记作 $p \succ p'$。

定义 2（skyline）　给定一个点集 D，D 中的点具有多维可度量属性，其 skyline 集合中包含了所有没有被其他点所支配的点，用于得到 skyline 集合的查询称为 skyline 查询。

图 3 展示了 skyline 的一个经典示例，其中包含了多个代表酒店的点，每个点具有两个可度量属性：到海滩的距离（纵坐标）以及价格（横坐标）。skyline 查询会返回所有不比其它酒店差（即距海滩较近、价格较便宜）的酒店（图 3 中加黑点）。当旅客想要在这些酒店中选择出价格便宜和到海滩距离近的酒店时，旅客只需要考虑 skyline 集合，而不在 skyline 集合中的酒店不需要考虑，因为总是可以在 skyline 集合中找到一个酒店，在价格上更便宜，并且到海边的距离更近。在理论上已证明，对任意单调评价函数 H，使其达到最优值的点必定属于 skyline 集合，因此 skyline 在多目标决策、偏好查询等领域具有重要价值。

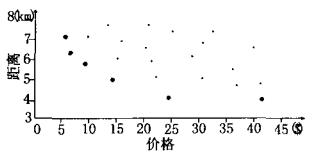

图 3　展示 skyline 概念的酒店选择示例

关于区间 skyline（interval skyline）的研究将 skyline 的概念引入了多时序数据分析领域：将每一个时序数据都视为一个考察点，某给定时间区间内的各个时间点（在一定时间粒度的定义下）作为该考察点的多个维度，继而在由众多时序数据所组成的考察点集中根据 skyline 的相关概念和方法从中选择没有被其他数据支配的那些时序数据，作为查询结果返回。文献[1]在近期对其进行了阐述和研究，其中使用量值属性作为在各个时间点维度上的度量。其相关定义如下。

定义 3（区间量值 skyline）　给定一个由多个时序数据组成的集合 S 和一个时间区间$[i:j]$，将$[i:j]$内的各个时间点（在一定时间粒度的定义下）作为时序数据的维度，量值属性作为维度的度量从而定义支配关系，则区间量值 skyline 集合包含了所有没有被 S 中任何其它数据支配的时序数据，表示为：

$$vol\_Sky[i:j]=\{s\in S \mid \nexists s'\in S, s' \succ^{vol}_{[i:j]} s\}$$

1.2　小波和小波概要

离散小波变换（Discrete Wavelet Transform，DWT）是时序数据分析领域中一种重要的方法，通过函数层次分解描述原始数据，函数中的参数包括一个总体近似参数和一系列细节参数。与其他数据分解方法（如离散傅里叶变换、分段线性近似等）相比，小波变换是唯一一种可以同时支持在时域和频域中进行多粒度分解的方法，且概念简单、易于实现，因而在时序数据压缩等领域得到大量的应用。

Haar 小波变换是最早出现也是最常用的一种小波变换，通过在不同时间粒度上平均化数据序列中相邻两个数据值的方式来计算小波参数，最后得到的小波参数是总体平均值（近似参数）和各个时间粒度上的差分值（细节参数）。表 1 是一个 Haar 小波变换的例子。设原始数据为{8,4,5,3,9,5,3,7}。第一层上的成对平均值为{(8+4)/2=6,(5+3)/2=4,(9+5)/2=7,(3+7)/2=5}。该层上与各个均值对应的差分

值为{8−6=2,5−4=1,9−7=2,3−5=−2}。基于这一新粒度的均值数据,使用同样的方法可以获得更上层的数据。最终的小波参数包括了总体均值(近似参数)和每一个不同粒度上的差分值(细节参数):{5.5,−0.5,1,1,2,1,2,−2}。

表1 Haar小波变换示例

	均值	细节参数
原始数据	{8,4,5,3,9,5,3,7}	—
第一级粒度	{6,4,7,5}	{2,1,2,−2}
第二级粒度	{5,6}	{1,1}
第三级粒度	{5.5}	{−0.5}

误差树[5]是一种常用的用于理解和研究Haar小波变换性质的数据结构。表1中的小波变换过程可表示为如图4所示的误差树。树中的非叶节点对应各个小波参数,其中根节点对应所有数据的总体均值,其余非叶节点均对应该节点左右子树所涉及数据的均值之差的一半(即对应细节参数),叶节点对应原始数据(叶节点不实际保存),树中的边权值赋为(左子树为正,右子树为负)。根据时序数据的误差树可以还原出原始数据值(即IDWT过程,Inverse Discrete Wavelet Transform),每一个叶节点所表示的原始数据值可由从根节点出发至该叶节点的父节点为止的路径中节点值与其下方对应边权值的乘积之和得到。如图4中,对于原始时序数据中第3个数据5,有+5.5+(−0.5)−1+1=5。

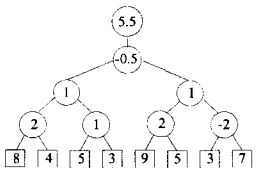

图4 表1中小波变换对应的误差树

本文使用文献[6]中提出的在多数据流场景下对误差树的标注方法:每一个非叶节点使用其所在层次及在该层中的位置序号作为脚标进行标识,即非叶节点表明这是对应于时序数据s_u的误差树中位于第l层第p个位置的节点;对应近似参数的节点位置序号定义为−1;时序数据s_u的第i个原始数据标识为$d_{u,i}$,如图5所示。

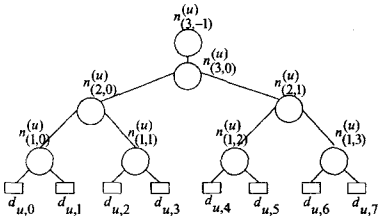

图5 多流场景下的误差树标注方法

由于数据流的数据规模往往较大,因此资源受限一般不会保存数据流中的原始数据,而是维护一个概要结构用于记录数据流的大致特征,小波概要是数据流处理中常用的一种概要结构[7-9]。数据流上的小波概要利用了小波变换的一个重要性质:利用远小于原始数据规模N的B个重要的小波参数就能重构出原始数据的一个较好的近似。将一个时序数据流小波参数中B个最重要的参数保留下来即为该数据流上的小波概要C_B。小波变换由于其多粒度性和易于在线更新的特点为数据流场景下的小波概要构造提供了必要条件。

近年来有多项工作研究如何在数据流上维持小波概要和利用小波概要对时序数据或多时序数据进行分析和处理。在数据流上维护小波概要的一个重要问题就是如何选择B个重要的小波参数作为小波概要。常用的度量小波参数重要性的标准包括误差平方和[10]、最大绝对误差和最大相对误差[11]等,主要目标是最小化由小波概要还原后的数据与原始数据的某种误差。由于该问题与本文讨论的问题正交,因此不在此进行深入讨论。文献[6]在研究背景上与本文较为接近,同样是在小波概要的支持下对多个时序数据进行分析,该工作中讨论了如何在小波概要的支持下对多个时序数据进行kNN和相似序列查询,并利用小波变换的多粒度性质提出一种快速查询算法。

2 区间差分skyline

2.1 定义及性质

由于区间量值skyline存在着如引言中所述的问题,因此提出区间差分skyline的概念,其定义如下。

定义4(区间差分skyline) 给定一个由多个时序数据组成的集合S和一个时间区间$[i:j]$,将$[i:j]$内的各个时间点(在一定时间粒度的定义下)作为时序数据的维度,以当前时间点相对于前一时间点的量值增长量属性作为维度的度量,从而定义支配关系,则区间差分skyline集合包含了所有没有被S中任何其它数据支配的时序数据,表示为

$$dif\_Sky[i:j]=\{s\in S\mid \nexists s'\in S, s'\succ^{dif}_{[i:j]} s\}$$

定义4所描述的区间差分skyline由于具有如定理1所描述的性质,因而在关注于趋势变化的多时序数据分析和偏好决策中更具参考价值。

定理1 给定一个由多个时序数据组成的集合S和一个时间区间$[i:j]$,如果S中的某个时序数据s在$[i:j]$的任意一个子区间上具有最大线性拟合斜率(由最小二乘法计算),则$s\in dif_Sky[i:j]$。

为了证明定理1,需要引入以下引理。

引理1 由最小二乘法确定的线性拟合函数$\hat{s}(t)=\hat{\theta}+\hat{\eta}t$的参数可由如下方式获取:

$$\hat{\eta}=\sum_{t=t_b}^{t_e}\left(\frac{t-\bar{t}}{SVS}\right)(s(t)-\bar{s})=\sum_{t=t_b}^{t_e}\left(\frac{t-\bar{t}}{SVS}\right)s(t) \tag{1}$$

$$\hat{\theta}=\bar{s}-\hat{\eta}\bar{t} \tag{2}$$

式中,

$$SVS=\sum_{t=t_b}^{t_e}(t-\bar{t})^2=\sum_{t=t_b}^{t_e}(t-\bar{t})t$$

$$\bar{s}=\left(\sum_{t=t_b}^{t_e}s(t)\right)/(t_e-t_b+1)$$

$$\bar{t}=\left(\sum_{t=t_b}^{t_e}t\right)/(t_e-t_b+1)=(t_b+t_e)/2$$

证明略。

证明(定理1的证明):使用反证法进行证明。

若定理1不成立,则有以下两个条件同时成立:

(1)$s\in S$是$[i:j]$的某个长度为l的子区间$[k_1:k_l]$内具有最大斜率的时间序列,其中$i\leqslant k_1<\cdots<k_l\leqslant j$,即

$\forall s'\in S, s'\neq s$,在$[k_1:k_l]$内有$\hat{\eta}_s>\hat{\eta}_{s'}$。

(2)$s\notin dif\_Sky[i:j]$,即

$\exists s'\in, s'\neq s, s'\succ^{dif}_{[i:j]} s$

由条件(2)可知,对于$\forall y, i\leqslant y\leqslant j$,有$(s'[y]-s'[y-1])(s[y]-s[y-1])$,且$\exists c, i\leqslant c\leqslant j$使得$(s'[c]-s'[c-1])>$

$(s[c]-s[c-1])$，则 $\forall x, 1\leqslant x\leqslant l, (s'[k_x]-s'[k_{x-1}])\geqslant(s[k_x]-s[k_{x-1}])$，即 $\forall x, 2\leqslant x\leqslant l, (s'[k_x]-s[k_x])\geqslant(s'[k_{x-1}]-s[k_{x-1}])$。

设 $(s'[k_x]-s[k_x])-(s'[k_1]-s[k_1])=r_k$，则 $r_k\geqslant 0, r_k\geqslant r_{k-1}$，由式(1)可知在 $[k_1:k_l]$ 内有：

$$\begin{aligned}\hat{\eta}_{s'}-\hat{\eta}_s&=\sum_{t=k_1}^{k_l}\left(\frac{t-\bar{t}}{SVS}\right)(s'(t)-s(t))\\&=\sum_{t=k_1}^{k_l}\left(\frac{t-\bar{t}}{SVS}\right)(s'(k_1)-s(k_1))+\sum_{t=k_1}^{k_l}\left(\frac{t-\bar{t}}{SVS}\right)r_k\\&=0+\sum_{t=k_1}^{k_l}\left(\frac{t-\bar{t}}{SVS}\right)r_k=\sum_{t=k_1}^{k_l}\left(\frac{t-\bar{t}}{SVS}\right)r_k\end{aligned}\tag{3}$$

由于在 $[k_2:k_l]$ 内 $r_k\geqslant r_{k-1}$，故序列 r_k 的斜率 $\hat{\eta}_{r_k}\geqslant 0$，故式(3)中 $\hat{\eta}_{s'}-\hat{\eta}_s\geqslant 0$，因此 $\hat{\eta}_{s'}\geqslant\hat{\eta}_s$，这与假设条件 1 矛盾，故定理 1 成立。

证毕。

2.2 利用小波概要计算区间差分 skyline

一种直接利用小波概要计算区间差分 skyline 的方法是通过 IDWT 还原原始时序数据，再按所要求的时间粒度对这些原始数据进行聚集计算，得到指定粒度上的数据均值，最后在获取差分序列后计算区间差分 skyline。

为了支持区间查询，引入定理 2。通过定理 2 可以找出在原始数据级给定区间 $[i:j]$ 后对应的误差树中不同层次(对应了不同粒度)上的节点范围，也可以找出误差树中的一个具体节点所对应的原始数据范围区间。

定理 2 误差树中叶节点 $d_{u,i}\sim d_{u,j}$ 对应的 l 层节点范围为

$$n^{(u)}_{(l,\lfloor i/2^l\rfloor)}\sim n^{(u)}_{(l,\lfloor j/2^l\rfloor)}$$

以节点 $n^{(u)}_{(l,p)}(p\neq -1)$ 作为根节点的子树对应叶节点的范围为 $d_{u,2^l*p}\sim d_{u,2^l*(p+1)-1}$。

证明可根据误差树的二叉树性质获得，限于篇幅省略。

由于存在如下的定理 3[6]，使得不必通过完整的 IDWT 过程获取原始数据再聚集到指定粒度上，而可直接根据指定粒度所对应层次之上的部分误差树获得该粒度上各个对应数据段的均值。

定理 3 给定一个误差树 T，其中的节点 $n^{(u)}_{(l,p)}$ 所对应的数据段均值 $a^{(u)}_{(l,p)}$ 可由如下公式获得：

$$a^{(u)}_{(l,p)}=\sum_{n^{(u)}_{(lx,px)}\in path(root,n^{(u)}_{(l,p)})} n^{(u)}_{(lx,px)}\times g(lx,px)\tag{4}$$

式中，

$$g(lx,px)=\begin{cases}1, & px=-1 \text{ or } lchild(n^{(u)}_{(l,p)})=n^{(u)}_{(lx,px)}\\ -1, & \text{otherwise}\end{cases}$$

$path(root,n^{(u)}_{(l,p)})$ 表示从误差树的根节点出发到 $n^{(u)}_{(l,p)}$ 的父节点为止的路径。

由于利用定理 3 获取某一粒度上均值的方法(在本文中称之为 PIDWT，Partial IDWT)与 IDWT 方法类似，仅仅是计算终止条件产生了变化，而且利用定理 3 计算某一粒度上差分序列的计算复杂度不高于直接利用 IDWT 的方法(当考察目标为原始数据层时，等价于直接利用 IDWT 方法)，因此本文将利用 PIDWT 计算区间差分 skyline 的方法作为基本算法与后面所提改进后的快速算法进行比较。相应的算法 Sky-PIDWT(Skyline-Partial IDWT)如图 6 所示。

算法 1 Sky-PIDWT

输入：多个时序数据的小波概要的集合 WC_set；
　　查询区域 $[i:j]$；感兴趣的时间尺度 l
输出：$dif\_skyline$
1. for each WC_u *in* WC_set //提取每一个时序数据的小波参数
2. $avg_u=$PIDWT(WC_u, l, i, j)；
3. //还原为 l 层涉及 $[i:j]$ 内数据的均值序列
4. $dif_u=dif(avg_u)$；//由均值序列计算差分序列
5. $dif\leftarrow dif_u$；//差分序列置入序列集合中
6. end;
7. dif_skyline=vol_skyline(dif,i,j)；
8. //使用区间量值 skyline 的方法在新数据上计算区间差分 skyline
9. report ($dif\_skyline$)；

图 6 利用 PIDWT 计算区间差分 skyline 的算法

算法 1 的计算复杂度为 $O(N_l\times high t_{root}\times m+N_l\times m)$，其中 m 为所考察的时序数据数量，N_l 为考察时间区间对应误差树第 l 层的数据规模，$high t_{root}$ 为从根节点到考察层次 l 的距离。该计算复杂度由计算差分序列的复杂度和根据差分序列来计算区间差分 skyline 的复杂度两部分组成。前一部分是计算复杂度的主要部分。

时序数据的小波细节参数存放了在各个不同时间粒度定义下不同时间点的数据量值差，因此根据这个性质可以通过时序数据流小波概要中的细节参数直接计算一定时间粒度定义下的区间差分 skyline，而不需要由小波概要根据 IDWT 还原原始时序数据或根据定理 2 的 PIDWT 还原某一粒度上的均值序列再进行后续计算。

由于有如下定理 4，因此可以利用小波细节参数计算一定时间粒度上的差分序列，继而得到所需的区间差分 skyline，从而提高查询的效率。首先给出相邻节点间连通路径概念的定义。

定义 5(相邻节点间的连通路径) 设误差树中两相邻节点的标识分别为 $n^{(u)}_{(l,p)}$ 和 $n^{(u)}_{(l,p+1)}$，同时涉及这两个节点的最小子树标识为 sub_T，该子树的根节点标识为 $n^{(u)}_{(l+z,r)}$，则由 $n^{(u)}_{(l,p+1)}$ 的父节点开始经由 $n^{(u)}_{(l+r,z)}$ 到达 $n^{(u)}_{(l,p)}$ 的父节点为止的路径称为这两个节点间的连通路径。

如图 7 所示，以误差树中叶节点 $d_{u,1}$ 和 $d_{u,6}$ 为例，其与前一相邻节点的最小子树分别以图中的两个虚线三角区域标识，而连通路径中的节点分别为 $\{n^{(u)}_{(1,0)}\}$ 和 $\{n^{(u)}_{(1,3)}, n^{(u)}_{(2,1)}, n^{(u)}_{(1,2)}\}$。

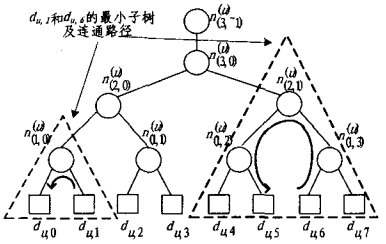

图 7 相邻节点间的最小子树和连通路径示例

定理 4 在给定粒度下某时间段与其前一时间段的均值差分值可由误差树中这两个时间段对应的相邻节点间的连通路径 P 中的节点值得到。

证明：设某两个相邻时间段对应的小波参数在误差树中对应节点的标识分别为 $n^{(u)}_{(l,p)}$ 和 $n^{(u)}_{(l,p+1)}$，对应均值为 $a^{(u)}_{(l,p)}$ 和 $a^{(u)}_{(l,p+1)}$，连通路径为 P，下面证明其均值差：

$$a\_dif=a^{(u)}_{(l,p+1)}-a^{(u)}_{(l,p)}=2n^{(u)}_{(l+z,r)}+\Sigma n^{(u)}_{P/r}\tag{5}$$

式中，$n_{P/r}^{(u)}$表示路径 P 中除最小子树 sub_T 的根节点 $n_{(l+z,r)}^{(u)}$ 外所经过的节点值。使用归纳法证明：

a)当时 $z=1$，$n_{(l,p)}^{(u)}$和 $n_{(l,p+1)}^{(u)}$的父节点即为涉及这两个节点的最小子树 sub_T 的根节点 $n_{(l+1,r)}^{(u)}$，由小波参数的定义可知均值差分值：

$$a\_dif=a_{(l,p+1)}^{(u)}-a_{(l,p)}^{(u)}=-2n_{(l+1,r)}^{(u)} \tag{6}$$

故当 $z=1$ 时定理 4 成立。

b)当 $z>1$ 时，从上层向下层开始递推证明过程。若节点位于 $l+1$ 层时定理成立，设 $n_{(l,p)}^{(u)}$和 $n_{(l,p+1)}^{(u)}$的父节点分别为 $n_{(l+1,m)}^{(u)}$和 $n_{(l+1,m+1)}^{(u)}$，则有：

$$a\_dif_{l+1}=a_{(l+1,m+1)}^{(u)}-a_{(l+1,m)}^{(u)}=-2n_{(l+z,r)}^{(u)}+\Sigma n_{P/r;n1;n2}^{(u)} \tag{7}$$

式中，$n_{P/r;n1;n2}^{(u)}$表示原路径 P 中除 sub_T 的根节点 $n_{(l+z,r)}^{(u)}$及 $n_{(l+1,m)}^{(u)}$，$n_{(l+1,m+1)}^{(u)}$外的节点值。由小波细节参数的计算方法可知($n_{(l,p)}^{(u)}$和 $n_{(l,p+1)}^{(u)}$分别为其父节点的左子树和右子树)：

$$\begin{aligned}&a_{(l,p+1)}^{(u)}-a_{(l+1,m+1)}^{(u)}=n_{(l+1,m+1)}^{(u)}\\&a_{(l,p)}^{(u)}-a_{(l+1,m)}^{(u)}=-n_{(l+1,m)}^{(u)}\end{aligned} \tag{8}$$

因此当节点位于 l 层时，有

$$\begin{aligned}a\_dif_l&=a_{(l,p+1)}^{(u)}-a_{(l,p)}^{(u)}\\&=(a_{(l+1,m+1)}^{(u)}-a_{(l+1,m)}^{(u)})+(n_{(l+1,m+1)}^{(u)}+a_{(l+1,m)}^{(u)})\\&=-2n_{(l+z,r)}^{(u)}+\Sigma n_{P/r;n1;n2}^{(u)}+n_{(l+1,m+1)}^{(u)}+n_{(l+1,m)}^{(u)}\\&=-2n_{(l+z,r)}^{(u)}+\Sigma n_{P/r}^{(u)}\end{aligned} \tag{9}$$

故定理 4 的递推过程成立。

由 a)，b)可知定理 4 成立。

证毕。

根据定理 4 快速计算对应误差树中 l 层粒度的差分区间 skyline 算法 Sky-ETP(Skyline-Error Tree Path)如图 8 所示。

图 8 中算法 2 的复杂度为 $O(length_p\times m+N_l\times m)$，与算法 1 一样由计算差分序列的复杂度和根据差分序列进而计算区间差分 skyline 的复杂度两部分组成，其中 m 为所考察的时序数据数量，N_l 为考察时间区间对应误差树第 l 层的数据规模，$length_p$ 为误差树中 N_l 个 l 层相邻节点间的总连通路径长度。前项中计算差分序列的复杂度是算法 2 计算复杂度的主要部分。定理 5 给出了误差树中处于 l 层、数据规模为 N_l 的相邻节点间总连通路径长度 $length_p$ 的上限。

算法 2 Sky-ETP

输入：多个时序数据的小波概要的集合 WC_set；

查询区域$[i,j]$；感兴趣的时间尺度 l

输出：$dif\_skyline$

1. for each WC_u in WC_set //提取每一个时序数据的小波参数
2. for each wc_p in $WC_u(l,i,j)$
3. //选取 l 层涉及$[i,j]$的每一个小波参数值及相应误差树节点
4. $P\leftarrow$path(wc_{p-1},wc_p); //获取需要的误差树中连通路径
5. $n_{root}=root(wc_{p-1},wc_p)$; //获取最小子树根节点
6. $dif_u(p)=-2n_{root}+\Sigma n_{P/r}$//计算差分值
7. end for;
8. $dif\leftarrow dif_u$; //差分序列置入序列集合中
9. end for;
10. $dif\_skyline=vol\_skyline(dif,i,j)$;
11. //使用区间量值 skyline 的方法在新数据上计算区间差分 skyline
12. report (dif_skyline)

图 8 利用小波细节参数计算区间差分 skyline 的算法

定理 5 设一个时序数据在粒度 l 上考察的数据规模为 N_l，则对应的误差树在相应层次上相邻节点间的总连通路径长度上限为 $3N_l+3+\lceil\log_2(N_l+1)\rceil^2$。

证明：误差树中在某一层次上的一个节点对应了在相应时间粒度上的一段数据，因此误差树在该粒度所对应的层次 l 上将考察 N_l 个节点。由于误差树是一个二叉树，因此在层次 l 上最多有$\lfloor(N_l+1)/2^n\rfloor+1$ 的节点与其左相邻节点在 $l+n$ 层上有共同的先辈节点，且路径长度为 $2n-1$。如当 $N_l=7$ 时，有$(7+1)/2=4$ 个节点与其左相邻节点有相同的父节点(即在 $l+1$ 层上有相同的先辈节点)，路径长度为 1；当 $N_l=6$ 时，有$\lfloor(6+1)/2^2\rfloor+1=2$ 个节点与其左相邻节点在 $l+2$ 层上有相同的先辈节点。因此对于误差树 l 层中的 N_l 个考察节点，相应的总连通路径长度最大为：

$$\begin{aligned}&\sum_{n=1}^{\lceil\log_2(N_l+1)\rceil}(\lfloor\frac{1}{2^n}(N_l+1)\rfloor+1)\times(2n-1)\\&<\sum_{n=1}^{\lceil\log_2(N_l+1)\rceil}(\frac{1}{2^n}(n_l+1)+1)\times(2n-1)\\&=(N_l+1)\sum_{n=1}^{\lceil\log_2(N_l+1)\rceil}(\frac{n}{2^{n-1}}-\frac{1}{2^n})+\sum_{n=1}^{\lceil\log_2(N_l+1)\rceil}(2n-1)\end{aligned} \tag{10}$$

令

$$N_s=\lceil\log_2(N_l+1)\rceil \tag{11}$$

将式(11)代入式(10)，则总路径长度最大为：

$$\begin{aligned}&(N_l+1)\sum_{n=1}^{N_s}(\frac{n}{2^{n-1}}-\frac{1}{2^n})+\sum_{n=1}^{N_s}(2n-1)\\&=(N_l+1)(\sum_{n=1}^{N_s}\frac{n}{2^{n-1}}-(1-2^{-N_s}))+N_s^2\end{aligned} \tag{12}$$

令

$$S_n=\sum_{n=1}^{N_s}\frac{n}{2^{n-1}} \tag{13}$$

则有

$$\frac{1}{2}S_n=\sum_{n=1}^{N_s}\frac{n}{2^n} \tag{14}$$

式(13)和式(14)相减，得到：

$$\begin{aligned}\frac{1}{2}S_n&=1+\frac{1}{2}+\frac{1}{4}+\cdots+\frac{1}{2^{N_s-1}}-\frac{N_s}{2^{N_s}}\\&=\sum_{n=1}^{N_s}\frac{1}{2^n-1}-\frac{N_s}{2^{N_s}}=2-\frac{1}{2^{N_s-1}}-\frac{N_s}{2^{N_s}}\end{aligned}$$

则

$$S_n=4-2^{-N_s}(4+2N_s) \tag{15}$$

由式(12)和式(15)可知总路径长度最大为：

$$(N_l+1)(3-2^{-N_s}(3+2N_s))+N_s^2<3N_l+3+\lceil\log_2(N_l+1)\rceil^2 \tag{16}$$

故 $3N_l+3+\lceil\log_2(N_l+1)\rceil^2$ 为总路径长度的一个上限。

证毕。

由于一般情况下，有

$$N_l\times height_{root}>3N_l+3+\lceil\log_2(N_l+1)\rceil^2$$

因此算法 2 中计算差分序列部分的效率优于算法 1 的相应部分。当 m 较大时，算法 2 相比算法 1 可以节省较多的计算时间。实验部分的结果证明了这一结论。

3 实验

为了验证所提概念和方法的有效性，采用来自于 NYSE，AMEX 和 NASDAQ 的部分历史股票数据集(取自 http://www.crsp.com/)进行实验分析。实际使用的数据集中包含了 500 支股票的价格数据，每一支股票数据均为从 2003 年中

至2007年底共1024天的股票收市价格。实验环境为Inter Core2 Duo CPU T9550 (2.66GHz)、2GB内存、Window XP。算法在Matlab7.0下实现。

为了直观地展示区间量值skyline和区间差分skyline的差异，选择了数据集中30支股票的部分数据在原始数据粒度上进行比较展示，如图9所示。图9(a)标注出了其中的区间量值skyline，由于有一支股票价格一直处于较高的水准，因此"淹没"了其他数据，在此情况下区间量值skyline的参考价值非常有限；而在图9(b)所展示的区间差分skyline集合中，共包含了7支股票的数据，直观来看，这些数据在测试数据集中表现得较为活跃。

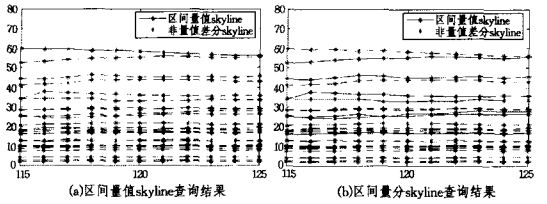

图9　区间量值skyline与区间差分skyline查询结果比较

图10展示了测试数据集中包含的时序数据数量不同时(区间长度固定为50；区间起点选择为第100个时间点；考察粒度为原始数据级)区间差分skyline的查询结果。可以从中发现，当考察的时序数据数量增加时，区间差分skyline的数量并没有同比增加，反而可能出现下降的情况，该查询在此数据集上没有出现因所考察的时序数据数量增加而使查询结果增长到无法控制的情况。

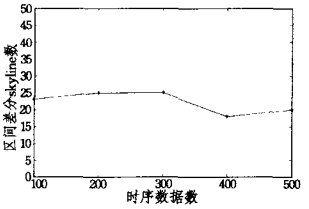

图10　时序数据数量与区间差分skyline集合规模的关系

图11展示了考察区间大小不同时(使用全部500个时序数据；区间起点选择为第100个时间点；考察粒度为原始数据级)查询出的区间差分skyline集合的规模。可以发现，随着考察区间的增大，区间差分skyline的规模也在增加，但增加速度远小于考察区间增大的速度。

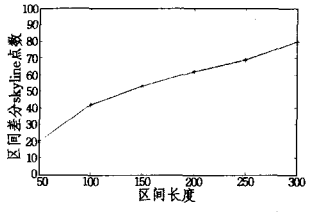

图11　区间长度与区间差分skyline集合规模的关系

图12展示了不同粒度时(使用全部500个时序数据；区间长度设定为200；区间起点选择为第100个时间点)区间差分skyline集合的规模(当$l=0$时为原始数据的粒度)。可以看出，随着时间粒度的增加，区间差分skyline集合的规模逐步下降。由于人们一般有对于短期区间关注于细粒度、对于长期区间关注于较大粒度的查询习惯，因此当处理长区间的查询需求时，可以适当提高处理粒度，从而获得更精简的查询结果。

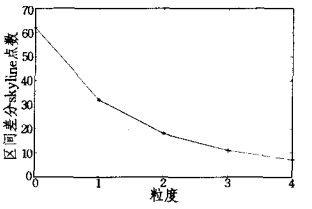

图12　考察粒度与区间差分skyline集合的规模关系

图13展示了不同规模的小波概要对检测结果准确性的影响(区间长度固定为200；区间起点选择为第100个时间点；考察粒度为原始数据级)，图13(a)为漏检率，图13(b)为误检率。这种影响是由小波概要自身的还原误差所造成的。本文所提方法本身并不会产生误差，它是一种建立在小波概要基础上的精确算法。

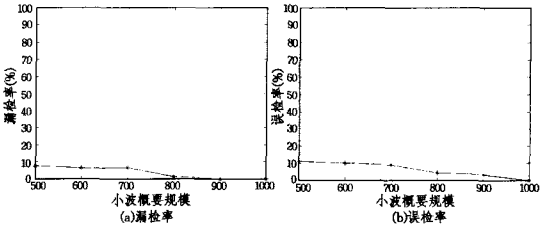

图13　不同规模小波概要对检测结果准确性的影响

图14展示了利用Sky-PIDWT方法和利用Sky-ETP方法分别计算不同长度的区间差分skyline所消耗的时间(所分析的数据集合为全部500个时序数据；区间起点选择为第100个时间点；考察粒度为原始数据级)。可以看出Sky-ETP方法的效率较高，且当区间长度增加时，Sky-ETP方法的耗时增长较为缓慢。

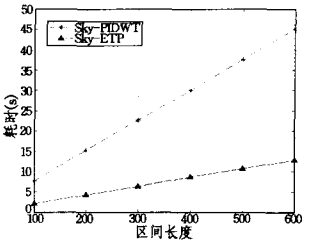

图14　区间长度不同时Sky-PIDWT与Sky-ETP效率比较

图15展示了利用Sky-PIDWT方法和利用Sky-ETP方法在不同层次中分析同样数量的节点时所消耗的时间(所分析的数据集合为全部500个时序数据；区间起点选择为第100个时间点；各层分析的节点数量均为50)。从中可以看出，Sky-PIDWT方法由于其计算复杂度与所考察层次到根节点的距离有关，因此计算同样数量的节点时，在低粒度上的计算耗时较多。而Sky-ETP方法的计算复杂度与层次无关，有稳定且较高的效率表现。

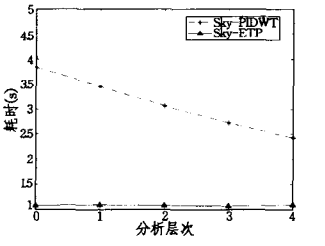

图15　考察层次不同时Sky-PIDWT与Sky-ETP效率比较

(下转第202页)

的评估值为 $r_2(M)$，确定度为 0.9875，显然方案 s_1 为最优开发方案。

结束语 云模型是一种体现随机性与模糊性关联的不确定性智能模型，它能够实现定性概念与定量值间的转换，是目前解决不确定性问题较为先进的研究方法，文章借鉴了云模型的研究思想，提出了基于自然语言的模糊多属性云决策方法，利用云模型描述属性评估集和等级评估集，并在此基础上设计了云归一化算法和云集结算法，有效地表达了人为主观认识的模糊性及被考察现象本身存在的客观随机性，较好地实现了自然语言的评价量化。最终案例分析的结果表明了该方法的简单性和可行性。

参 考 文 献

[1] Francisco M. An adaptive consensus support model for group decision-making problems in a multigranular fuzzy linguistic context[J]. IEEE Transactions on Fuzzy Systems, 2009, 17(2): 279-290

[2] Alonso S. Group decision making with incomplete fuzzy linguistic preference relations[J]. International Journal of Intelligent Systems, 2009, 24(2): 201-222

[3] Dong Yucheng. Linguistic multiperson decision making based on the use of multiple preference relations[J]. Fuzzy Sets and Systems, 2009, 160(5): 603-623

[4] Hsu Tsuen-Ho, Tsai Tsung-Nan, Chiang Pei-Ling. Selection of the optimum promotion mix by integrating a fuzzy linguistic decision model with genetic algorithms[J]. Information Science, 2009, 179(1/2): 41-52

[5] Zhang Zaifang. Fuzzy group decision-making for multi-format and multi-granularity linguistic judgments in quality function deployment[J]. Expert Systems with Applications, 2009, 36(5): 9150-9158

[6] Sandra B-A. Linguistic markers of decision processes in aproblem solving task[J]. Cognitive Systems Research, 2009, 10(2): 102-123

[7] Liu Yang. A group decision-making method uncertain linguistic information[J]. Journal of Northeastern University: Natural Science, 2009, 30(4): 601-604

[8] Alonso S. Group decision making wit incomplete fuzzy linguistic preference relations [J]. International Journal of Intelligent Sysstems, 2009, 24(2): 201-222

[9] Garcia-Lapresta J L. Defining the Borda count in a linguistic decision making context[J]. Information Sciences, 2009, 179(14): 2309-2316

[10] 李华莹，罗自强，李德毅. 基于云模型的汽车款式知识表示[J]. 舰船电子工程，2006，26(6)：1-4

[11] 李德毅，刘常昱. 不确定性人工智能[J]. 软件学报，2004，15(11)：1583-1594

[12] 杜鹢，李德毅. 基于云的概念划分及其在关联采掘上的应用[J]. 软件学报，2001，12(2)：196-203

[13] 刘常昱，李德毅. 正态云模型的统计分析[J]. 信息与控制，2005，34(2)：236-248

(上接第 165 页)

从图 14 和图 15 可以看出，Sky-ETP 方法较之 Sky-PID-WT 方法的效率有了较大幅度的提高。这与之前对两种方法的时间复杂性分析的结果相同，从实验角度说明了分析结果的正确性。

结束语 随着信息技术的发展和自动化水平的提高，很多应用领域都产生了大量的时序数据，数据分析的对象也从单一时序数据扩展到了多个时序数据，同时很多时序数据呈现为数据流的形式。由于数据规模较大且有实时性的要求，因此只会维护一个反映其主要特征的概要结构而不会保留原始数据。本文讨论了针对多时序数据选择分析的区间 skyline 问题，针对原有研究的不足提出了区间差分 skyline 的概念，对其应用场景和性质进行了分析和研究。继而讨论了如何基于时序数据流常用的小波概要支持区间差分 skyline 查询的问题，并利用 Haar 小波变换中细节参数的差分属性提出了一种快速计算区间差分 skyline 的方法。在真实股票数据集上的实验表明了本文所提方法的有效性。

当前，在数据流概要结构支持下对多时序数据进行分析和处理的研究还处于起步阶段。未来的工作包括研究如何在数据流概要结构的支持下快速处理各种多时序数据分析中涉及到的问题，例如研究在概要结构的支持下如何对多时序数据的聚类、kNN 查询、异常分析和检测等问题进行高效处理，并将在网络安全、话题发现、金融分析等更多应用领域中分析所提方法的适用场景和有效性。

参 考 文 献

[1] Jiang B, Pei J. Online Interval Skyline Queries on Time Series [C]// Proc. of the IEEE Int'l Conf. on Data Engineering. 2009: 1036-1047

[2] He Q, Chang K, Lim E. Using Burstiness to Improve Clustering of Topics in News Streams[C]// Proc. of the 7th IEEE Int'l Conf. on Data Mining. 2007: 493-498

[3] Börzsönyi S, Kossmann D, Stocker K. The Skyline Operator[C]// Proc. of the 17th Int'l Conf. on Data Engineering. 2001: 421-430

[4] Papadias D, Tao Y, Fu G, et al. An optimal and progressive algorithm for skyline queries[C]// Proc. of the ACM SIGMOD Int'l Conf. on Management of Data. 2003: 467-478

[5] Matias Y, Vitter J S, Wang M. Wavelet-based histograms for selectivity estimation[J]. SIGMOD Rec, 1998, 27(2): 448-459

[6] Hung H, Chen M. Efficient range-constrained similarity search on wavelet synopses over multiple streams[C]// Proc. of the 15th ACM Knt'l Conf. on information and Knowledge Management. Arlington, Virginia, USA, 2006: 327-336

[7] Guha S, Harb B. Wavelet synopsis for data streams: minimizing non-euclidean error[C]// Proc. of the 11th ACM SIGKDD Int'l Conf. on Knowledge Discovery in Data Mining. Chicago, Illinois, USA, 2005: 88-97

[8] Garofalakis M, Gibbons P B. Wavelet synopses with error guarantees[C]// Proc. of the ACM SIGMOD Int'l Conf. on Management of Data. Madison, Wisconsin, 2002: 476-487

[9] 陈华辉，施伯乐. 数据流上具有数据遗忘特性的小波概要[J]. 计算机研究与发展，2009，46(2)：268-279

[10] Gilbert A C, Kotidis Y, Muthukrishnan S, et al. One-pass wavelet decompositions of data streams[J]. IEEE Transactions on Knowledge and Data Engineering, 2003, 15(3): 541-554

[11] Garofalakis M, Kumar A. Deterministic wavelet thresholding for maximum-error metrics[C]// Proc. of the 23rd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. Paris, France, 2004: 166-176