基于傅立叶变换的一种时间序列相似搜索算法

陈晓航　彭　宏　谢运祥
（华南理工大学计算机工程与科学系，广州 510640）

摘　要　该文提出了基于傅立叶变换的一种新的时间序列相似搜索算法。该算法利用高效的索引方法，达到快速的匹配，解决了多序列的子序列匹配问题。大量算例验证了该算法的通用性和有效性，它可以应用到求解各种时间序列相关的实际问题。

关键词　时间序列　离散傅立叶变换　R+树

文章编号 1002-8331-(2002)18-0202-02　**文献标识码** A　**中图分类号** TP274

A DFT-Based Algorithm for Similarity Search in Time Series Database

Chen Xiaohang　Peng Hong　Xie Yunxiang
(Dept.of Comp.Eng.& Scien.,South China Univ.of Tech.,Guangzhou 510640)

Abstract: This paper presents a DFT-based algorithm for similarity search in time series database.Using an efficient index method,people can get a fast matching,and have solved the subsequence matching of multi-subsequence.A number of examples show that the method presented is universally useful and effective.It can be applied for solving various practical problem on time series.

Keywords: Time Series,DFT,R+ Tree

1　引言

时序数据库在商业行为以及科学研究的决策支持方面起着非常重要的作用。时间序列可以用来表示股票价格的变化、商品的销售信息、天气的变化情况、生物信息、音乐信息等。因此，对海量数据进行时间序列的研究和应用已成为数据挖掘领域的一个新课题。

目前，在新型的数据库应用中，对数据挖掘都要求具备相似查找的功能，例如，在数据库中查询价格变化相似的股票、查询销售模式相似的公司等等。因此，设计出一种快速的时间序列相似搜索算法可以大大地提高数据库系统的查询功能，可以从海量数据中快速发现知识。时间序列的相似性查询包括全序列匹配和子序列匹配。在[2]中，Agrawal 等人提出了一种序列数据库中序列匹配的快速方法，比较好地解决了全序列的匹配问题。为了解决子序列的匹配问题，该文提出了一种新的基于傅立叶变换的时间序列相似搜索算法。定义子序列的匹配问题为：$X=(X_f)f=0,1,2,....t-1$ 表示长度为 t 的一个时间序列。给出 $N+1$ 个任意长度的序列 $X,Y_1..Y_n$，以及参数 ε，子序列的匹配就是在找出所有的 X 的子序列 X' 与 Y_i 的子序列 Y_i'，使 X' 与 Y_i' 的距离小于 ε。其中 X 与 Y 的距离：

$$D(X,Y)=\frac{\sum_{f=0}^{n-1}|X_f-Y_f|^2}{\sum_{f=0}^{n-1}(X_f^2+Y_f^2)}$$

若 $D(X,Y)<\varepsilon^2$ 则被认为 X 与 Y 是匹配的。在解决子序列匹配的问题方面，Faloutsos 提出了一个经典的算法：将序列提取特征，采用 ST-Index 算法进行索引，用 PreFixSearch 或 MultiPiece 方法进行匹配搜索。该算法的缺点是匹配搜索产生较多的错误信号，无法满足消除噪音的要求。Agrawl 等人提出了一个考虑了噪音、幅度等问题的子序列匹配算法，该算法的缺点是产生点数据库过大，从而造成了索引树过高，不利于匹配过程的查找。针对以上方法的缺点，笔者提出一个新的快速 F-R-T 算法，解决了多序列的子序列匹配问题。

2　新的快速 F-R-T 算法

该算法分为建立索引、获取匹配对、组成匹配序列 3 步。

2.1　建立索引

先定义一个查找长度 $\omega(\omega>1)$，ω 的选定由具体的应用决定。把长度为 ω 的滑动窗口放在序列的每一个起始位置，就可以获得一段长度为 ω 的子序列，对这段序列进行离散傅立叶变换，可以提取前几个参数作为特征（假设为 f 个参数，一般 f 取 3 即可）[2]，将这个长度为 ω 的子序列变成 f 维特征空间上的点。在每一个可能位置上都安置滑动窗口，然后移动滑动窗口，可以得到 $Len(S)-\omega+1$ 个 f 维特征空间上的点。显然，由特征空间的点组成的数据库远远大于序列数据库，为了避免对整个序列的检索提高查找速度，Christos 等提出一种贪婪算法。图 1 显示了两个轨迹 $C1$ 和 $C2$，坐标 $F1$ 和 $F2$ 是特征空间中的两个特征。

把模分成子轨迹，每一段子轨迹用最小边界矩形（Minimun Bound Rectange）代替。采用 R+树[4]来存储这些 MBR，用 R+树进行检索，找到与查询序列相交的所有 MBR，通过这些 MBR读

基金项目：国家自然科学基金（编号：50007001）；广东省自然科学基金（编号：990582）；广州市科委基金（编号：2000-J-006-01）

作者简介：陈晓航（1978-），男，硕士研究生，主要研究方向：数据挖掘技术。彭宏（1956-），男，教授，博导，主要研究方向：计算智能，数据挖掘技术。谢运祥（1966-），男，副教授，博士，主要研究方向：电气控制技术。

取子序列。在图 1 中,可以看到两个轨迹,第一个 $C1$ 被分为3个子轨迹,而第二个 $C2$ 被分为 5 个子轨迹。目标是将一个时间序列映射成特征空间中的矩形集合。

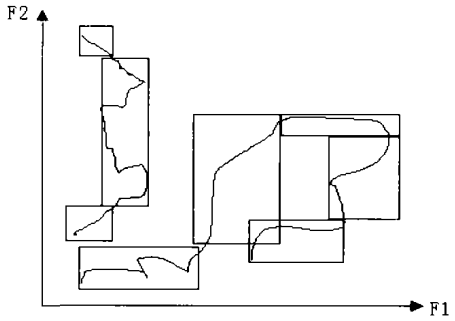

图 1

给出进行 MRB 的贪婪算法如下:

```
Procedure SetUpMRB()
PutIntoFirstMBR ();//将序列的第一个点放到第一个 MBR 中,产生第一个子轨迹
For 特征空间中的没一个点 p
 If(IncreaseCost(CurrentMBR,p)==True)then
  CurrentMBR= NewMBR();//开始另一个子轨迹
 Else
  IncludeToMBR(p,CurrentMBR);
End Procedure
```

边界代价函数为:$MC=\prod_{i=1}^{n}(L_i+0.5)/k$。其中 $L1,L2.....Ln$ 为 k 个点组成的 MBR 边,对被查询的子序列 $Y_1..Yn$ 用 SetupMRB 建立 MRB,用 R+树对这些 MRB 进行索引,查询将在这个 R+树上进行。

2.2 获得匹配对

将序列 X 采用滑动窗口的方法,获得 $Len(X)-\omega+1$ 个 f 维特征空间上的点。目标是找出所有得到的点中与这些点相匹配的点对。由于序列 X 中的一个 f 维特征空间的点 Q 与相匹配的点在以点 Q 为中心,半径为 ε 的球体内,通过 R+树索引,可以找出与这个球体相交的所有 MRB,检查里面的子序列窗口,判断该窗口是否与 Q 点匹配,如果是,则可以用以下匹配算法进行匹配。

```
/* 匹配算法 */
Procedure PointMatch(Point Q,Rtree *Root)
 S=Sphere(Q,ε);//建立球体
 while MRB∩S
  for every point q in MRB
   if(CheckMatch(q,Q)==True)
    Label(q,Q);//如果匹配则做标记.
   End if
  End for
 End while
End Procedure
```

经上述过程后,可以获得所有匹配的窗口对,这些窗口对的排序结果就是要得到的结果。

2.3 组成匹配序列

在获得匹配对后,可以将这些匹配对看作一个点,因此,组成匹配序列的过程被认为是在无环图中寻找最长的路径。假设匹配对为子序列 S、T,利用获得的点,可以组成如下的图:

(1)每一个匹配的点为图的一个点。

(2)如果两个匹配对 $M=(Si,Ti)$,$N=(Sj,Tj)$满足以下条件,则它们之间可以连接:M 中的 Si、Ti 的起点必须小于 N 中 Sj、Tj 的起点。即 $First(Si)<First(Sj)$,$First(Ti)<First(Tj)$。$First()$函数返回子序列在序列中的位置。

对图中两个起点相同的边 $P1$,$P2$, 如果 $P1$ 的长度小于 $P2$,那么对图中的任何一条边 Arc 来说,$P1$+Arc 的长度也小于 $P2$+Arc。通过用反拓扑的方法来遍历整个图,从而获得最长的路径,即得到序列中匹配长度最大的匹配子序列。同时还可以得到其它满足匹配长度要求的子序列。

3 实验结果

在一个有 8000 个点的股票价格数据库上,用 C++实现 F-R-T 算法。设 ε=0.1,即如果两个点 X,Y 的差别小于 0.1,那么认为它们是相似的,ω=128,表 1 显示了新算法和 Agrawal 算法在时间上的比较:

表 1

序列长度	F-R-T 算法所用时间(秒)	Agrawal 算法所用时间(秒)
1024	85	78
2048	263	292
4096	1021	1417

其中 Y 序列的个数为 4 个,由此可见,该算法在解决长序列匹配的问题上,时间复杂度小。

4 结论

该文解决了多序列的子序列匹配问题,提出了一种基于傅立叶变换的时间序列相似性搜索的新算法,并且应用到股票价格分析上。实验结果表明,该算法比 Agrawal 算法时间复杂度小,更具有通用性和有效性,它可以应用到求解各种时间序列相关的实际问题。(收稿日期:2001 年 12 月)

参考文献

1.段立娟等.时间数据库中的相似序列的挖掘[J].计算机科学,2000;27:39~44

2.Agrawal R et al.Efficient similarity search in sequence database[C].In:FODO Conf Evanston Illinois,1993:69~84

3.Faloutsos C et al.Fast subsequence matching in time series database [C].In:SIGMOD Conference,1994:419~429

4.Sellis C et al.The R+ tree:A dynamic index for multi-dimensional objects.VLDB,1987:507~518

5.Agrawal R et al.Fast similarity search in the presence of moise scaling, and translation in time-series database.VLDB,1995:305~316

6.Carrie B.Graphs and networks[M].Claredon Press,Oxford,1978