学校编码：10384　　分类号_____密级 _____
学　　号：200228019　　UDC_____

厦　门　大　学

硕　士　学　位　论　文

利用基于移动均值的索引实现时间序列的相似查找

Fast Similarity Search in Timeseries Databases with a New Indexing Mechanism Based on Moving Average

林　子　雨

指导教师姓名：　薛　永　生　教授
专　业　名　称：　计　算　机　应　用
论文提交日期：　2005　年　4　月
论文答辩日期：　2005　年　6　月
学位授予日期：　2005　年　　月

答辩委员会主席：_____
评　　阅　　人：_____

2005 年 4 月

厦门大学学位论文原创性声明

兹呈交的学位论文，是本人在导师指导下独立完成的研究成果。本人在论文写作中参考的其他个人或集体的研究成果，均在文中以明确方式标明。本人依法享有和承担由此论文而产生的权利和责任。

声明人（签名）：

年　　月　　日

摘 要

在我们的日常生活和工作中，时序数据是广泛存在的,通过对这些数据的系统性的分析，从中发现其系统内部的模式、知识和规则从而对系统的演变趋势进行准确的预测无疑具有重大的理论价值和应用价值。由此，基于时间序列的数据挖掘正日益受到越来越多研究人员的关注，也必将成为数据挖掘领域中一个新的研究热点。

在海量的时间序列数据中检索特定的内容，我们不仅要借助于高效的数据压缩技术，从而尽可能地减少数据存储所需要的空间，而且还要建立有效的索引机制，并使其具备动态建立、动态更新和快速查找的重要特性。本文通过对时间序列特性的研究，发现了并证明了相似时间序列的移动均值之间所具备的特殊关系，基于这一关系，提出了基于移动均值的缩距比关系定理，其“裁减”功能十分强大，可以快速淘汰那些不符合条件的时间序列，以达到快速查找满足条件的时间序列的目的。

在缩距比关系定理的基础上，本文又进一步提出了一个基于移动均值的索引来解决在大型时间序列数据库中进行相似查找的问题。该索引机制具备了有效索引机制的基本特征，支持索引的动态建立和更新，并具有较小的空间开销和很好的查找性能，并且能够保证在检索过程中不会漏掉符合条件的时间序列。最后，在一个股票交易数据集上进行了诸多实验，并与目前已提出的具有良好性能的其他索引机制进行了在同等实验条件下的性能比较，实验结果证明了基于移动均值索引机制所具有的良好性能。

关键词：MABI，数据挖掘，时间序列，移动均值，欧氏距离，分段聚集近似，缩距定理，缩距比关系定理

Abstract

Timeseries data exist everywhere in our daily life. Through the comprehensive analysis of these data, we can derive some valuable patterns, knowledge and rules from the system, which can be used to accurately predict the trend of the evolution of certain system.

Locating what we want from large amounts of data, we not only have to turn to data compression technique of high efficiency, so as to greatly reducing the space cost, but need an indexing mechanism with good performance as well. Such indexing mechanism must give support to the features such as dynamic building, updating and fast searching. Through the research of the characteristics of timeseries, we find certain relationship existing between similar timeseries, based on which we here propose the DRR relation theorem. This theorem is most powerful in the aspect of pruning capability, which means that most of the unqualified candidates can be pruned during the searching process, and thus the performance enhancement is achieved through this way.

Furthermore, based on the DRR relation theorem, we also present a new indexing mechanism to deal with the problem of fast similarity search in very large timeseries databases. This indexing mechanism have the fundamental features of a desirable index system such as good searching performance, low cost of space, dynamic building and updating, and by no means missing those qualified candidates. Finally the paper reports some experiment results conducted on a stock price data set, and shows the good performance of MABI method.

Keywords： MABI, moving average, time-series databases, Euclidean distance, piecewise aggregate approximation, distance reducing theorem, DRR relation theorem

目 录

第一章 基于时间序列的数据挖掘 1
1.1 数据挖掘概述 1
1.1.1 什么是数据挖掘 1
1.1.2 在何种数据上进行数据挖掘 3
1.1.3 数据挖掘的功能 4
1.2 基于时间序列的数据挖掘介绍 5
1.2.1 什么是时间序列 5
1.2.2 基于时间序列的数据挖掘的研究 7
1.2.2.1 趋势分析 9
1.2.2.2 时序分析中的相似搜索 11
1.2.2.3 序列模式挖掘 12
1.2.2.4 周期分析 14
1.2.3 时间序列的存储和索引问题的研究现状 15
1.3 本文的主要工作 17
第二章 时间序列的存储和索引机制 18
2.1 降维技术 18
2.1.1 离散傅立叶变换 18
2.1.2 分段线性表示 20
2.1.3 离散小波变换 22
2.1.4 奇异值分解 24
2.1.5 分段聚集近似 24
2.2 降维技术的有效性和完整性 26
2.3 时间序列的匹配方式 27
2.4 有效索引机制具备的特征 28
2.5 索引机制相关研究成果 29
2.5.1 R 树 29
2.5.2 扩展多维动态索引文件 30
2.5.3 TIP-索引 31
2.5.4 STB-索引 32
2.5.5 其他索引方法 34
第三章 移动均值和相关定理 35
3.1 移动均值 35
3.1.1 移动均值的描述 35
3.1.2 移动均值的定义 35
3.2 距离度量方法 37

3.2.1 欧氏距离和曼哈坦距离 37
3.2.2 时间序列相似的定义 38
3.3 缩距定理 39
3.3.1 缩距定理的描述 39
3.3.2 缩距定理的证明 40
3.3.3 缩距定理的推论 43
3.4 缩距比和缩距比关系定理 43
3.4.1 缩距比的定义 43
3.4.2 缩距比关系定理 44
3.5 线性序列相似定理 45
第四章 MABI 索引机制 47
4.1 MABI 索引的描述 47
4.2 MABI 索引的结构 47
4.3 MABI 索引树生成过程和算法 49
4.4 在 MABI 索引中的查找过程和算法 49
4.5 MABI 索引方法中的三次筛选 52
4.6 定理在 MABI 索引中的应用 53
4.7 性能分析 55
4.7.1 实验描述 55
4.7.2 缩距比的分布统计 56
4.7.3 缩距比关系定理的淘汰能力测试结果 57
4.7.4 淘汰率曲线 58
4.7.5 MABI 与其他索引方法的性能比较 59
4.8 MABI 索引方法的有效性 60
4.9 MABI 索引方法的不足 61
第五章 结束语 62
参考文献 63
个人研究成果 67
致谢 69

Contents

Chapter 1 Data mining based on time series 1
1.1 Data mining 1
1.1.1 What is data mining 1
1.1.2 Types of data used in data mining 3
1.1.3 Function of data mining 4
1.2 Introduction on data mining based on timeseries 5
1.2.1 What is timeseries 5
1.2.2 Research on data mining based on timeseries 7
1.2.2.1 Trend analysis 9
1.2.2.2 Similarity search in timeseries analysis 11
1.2.2.3 Sequence pattern mining 12
1.2.2.4 Cycle analysis 14
1.2.3 Research on the storing and indexing of time series 15
1.3 Our work 17
Chapter 2 Storing and indexing mechanism of time series 18
2.1 Dimension reduction 18
2.1.1 Discrete Fourier transformation 18
2.1.2 Piecewise linear representation 20
2.1.3 Discrete wavelet transformation 22
2.1.4 Singular value decomposition 24
2.1.5 Piecewise aggregate approximation 24
2.2 Effectivity and integrity of dimension reduction 26
2.3 Matching methods of timeseries 27
2.4 Characteristics of effective indexing mechanisms 28
2.5 Relating research accomplishments of indexing mechanism 29
2.5.1 R tree 29
2.5.2 Extended multidimensional dynamic index file 30
2.5.3 TIP-indexing 31
2.5.4 STB-indexing 32
2.5.5 Other indexing mechanisms 34
Chapter 3 Moving average and relating theorems 35
3.1 Moving average 35
3.1.1 Description of moving average 35
3.1.2 Definition of moving average 35
3.2 Distance measuring methods 37

3.2.1 Euclidean distance and Manhattan distance 37
3.2.2 Definition of timeseries similarity 38
3.3 Distance reducing theorem 39
3.3.1 Description of distance reducing theorem 39
3.3.2 Proof of distance reducing theorem 40
3.3.3 Deduction of distance reducing theorem 43
3.4 Distance reducing rate and relating theorem 43
3.4.1 Definition of distance reducing rate 43
3.4.2 Distance reducing rate relation theorem 44
3.5 Sequence similarity theorem 45
Chapter 4 MABI indexing mechanism 47
4.1 Description of MABI index 47
4.2 Structure of MABI index 47
4.3 Construction of MABI indexing tree 49
4.4 Querying in MABI index 49
4.5 Three steps of sifting in MABI index 52
4.6 Application of theorems in MABI index 53
4.7 Performance analysis 55
4.7.1 Description of experiments 55
4.7.2 Distribution of distance reducing rates 56
4.7.3 Test results of prunning capability 57
4.7.4 Prunning rate curve 58
4.7.5 Performance comparision 59
4.8 Valadity of MABI index 60
4.9 Deficiency of MABIindex 61
Chapter 5 Conclusion 62
References 63
Personal research accomplishments 67
Acknowledgement 69

第一章　基于时间序列的数据挖掘

1.1　数据挖掘概述

1.1.1　什么是数据挖掘

简单地说，数据挖掘是从大量数据中提取或挖掘知识。该术语实际上有点用词不当。注意，从砂子或矿石挖掘黄金称作“黄金挖掘”，而不是“砂石挖掘”。这样，数据挖掘应当更正确地命名为“从数据中挖掘知识”，不幸的是这有点长。“知识挖掘”是一个短术语，可能不能反映从大量数据中挖掘。毕竟，挖掘是一个很生动的术语，它抓住了从大量的、未加工的材料中发现少量金块这一过程的特点。这样这种用词不当携带了“数据”和“挖掘”，成了流行的选择。还有一些术语，具有和数据挖掘类似但稍有不同的含义，如数据库中知识挖掘、知识提取、数据/模式分析、数据考古和数据捕捞。

许多人把数据挖掘视为另一个常用的术语“数据库中的知识发现”或KDD的同义词。而另一些人只是把数据挖掘视为数据库中知识发现过程的一个基本步骤。知识发现过程由以下步骤组成：

（1）数据清理（消除噪声或不一致数据）

（2）数据集成（多种数据可以组合在一起）

（3）数据选择（从数据库中检索与分析任务相关的数据）

（4）数据变换 (数据变换或统一成适合挖掘的形式，如通过汇总或聚集操作)

（5）数据挖掘（基本步骤，使用智能方法提取数据模式）

（6）模式评估（根据某种兴趣度度量，识别表示知识的真正有趣模式）

（7）知识表示（使用可视化或知识表示技术，向用户提供挖掘的知识）

数据挖掘步骤可以与用户或知识库交互。有趣的模式提供给用户，或

作为新的知识存放在数据库中。注意，根据这种观点，数据挖掘只是整个过程中的一步，尽管是最重要的一步，因为它发现隐藏的模式。在产业界、媒体和数据库研究界，“数据挖掘”比较长的术语“数据库中知识发现”更流行。

典型的数据库挖掘系统具有以下主要成份：

（1）数据库、数据仓库或其他信息库：这是一个或一组数据库、数据仓库、电子表格或其他类型的信息库。可以在数据上进行数据清理或集成。

（2）数据库或数据仓库服务器：根据用户的数据挖掘请求，数据库或数据仓库服务器负责提取相关数据。

（3）知识库：这是领域知识，用于指导搜索，或评估结果模式的兴趣度。这种知识可能包括概念分层，用于将属性或属性值组织成不同的抽象层。用户确信方面的知识也可以包含在内。可以使用这种知识，根据非期望性评估模式的兴趣度。领域知识的其他例子有兴趣度限制和元数据。

（4）数据挖掘引擎：这是数据挖掘系统基本的部分，由一组功能模块组成，用于特征化、关联、分类、聚类分析以及演变和偏差分析。

（5）模式评估模块：通常，此成份是用兴趣度度量，并与数据挖掘模块交互，以便将搜索聚焦在有趣的模式上。它可能使用兴趣度阈值过滤发现的模式。模式评估模块也可以与挖掘模块集成在一起，这依赖于所用的数据挖掘方法的实现。对于有效的数据挖掘，建议尽可能深地将模式评估推进到挖掘过程中，以便将搜索限制在有兴趣的模式上。

（6）图形用户界面：本模块在用户和数据挖掘系统之间通信，允许用户与系统交互，指定数据挖掘查询或任务，提供信息、帮助搜索聚焦，根据数据挖掘的中间结果进行探索式数据挖掘。

数据挖掘技术涉及多学科技术的集成，包括数据库技术、统计学、机器学习、高性能计算、模式识别、神经网络、数据可视化、信息检索、图像与信号处理和空间数据分析。

1.1.2　在何种数据上进行数据挖掘

原则上讲，数据挖掘可以在任何类型的信息存储上进行。这包括关系数据库、数据仓库、事务数据库、高级数据库系统、展开文件和 WWW。高级数据库系统包括面向对象和对象-关系数据库；面向特殊应用的数据库，如空间数据库、时间序列数据库、文本数据库和多媒体数据库。挖掘的挑战和技术可能因存储系统而异。

（1）关系数据库：关系数据库是表的集合，每个表都被赋予一个唯一的名字。每个表包含一组属性（列或字段），并通常存放大量元组（记录或行）。关系中的每个元组代表一个被唯一的关键字标识的对象，并被一组属性值描述。语义数据模型，如实体-联系数据模型，将数据库作为一组实体和它们之间的联系进行建模。

（2）数据仓库：数据仓库是一个面向主题的、集成的、时变的、非易失的数据集合。数据仓库系统允许将各种应用系统集成在一起，为统一的历史数据分析提供坚实的平台，对信息处理提供支持。通常，数据仓库用多维数据库结构建模。其中，每一维对用于模式中的一个或一组属性，每个单元存放某个聚集度量值。数据仓库的实际物理结构可以是关系数据库或多维数据立方体。

（3）事务数据库：一般地说，事务数据库由一个文件组成，其中每个记录代表一个事务。通常，一个事务包含一个唯一的事务标识号和一个组成事务的项的列表。事务数据库可能有一些与之相关联的附加表，包含关于销售的其他信息，如事务的日期、顾客的 ID 号、销售者 ID 号、销售分店，等等。

（4）高级数据库系统和高级数据库应用：新的数据库应用包括处理空间数据（如地图）、工程设计数据（如建筑设计、系统部件、集成电路）、超文本和多媒体数据（包括文本、影像、图像和声音数据）、时间相关的数据（如历史数据或股票交易数据）和 WWW（通过 Internet 可以使巨大的、

广泛分布的信息存储）。这些应用需要有效的数据结构和可伸缩的方法，处理复杂的对象结构、变长记录、半结构化或无结构的数据以及文本和多媒体数据，并具有复杂结构和动态变化的数据库模式。为了响应这些需求，开发了高级数据库系统和面向特殊应用的数据库系统。这些包括面向对象和对象-关系数据库系统、空间数据库系统、时间和时间序列数据库系统、文本和多媒体数据库系统、异种和遗产数据库系统、基于 WWW 的全球信息系统。

1.1.3 数据挖掘的功能

数据挖掘功能在于用指定数据挖掘任务中要找的模式类型。数据挖掘任务一般分为两种类型：描述和预测。描述性挖掘任务刻画数据库中数据的一般特性。预测性数据挖掘任务在当前数据上进行推断，以进行预测。数据挖掘功能以及它们可以发现的模式类型如下：

（1）概念/类描述：特征化和区分。数据特征化是目标类数据的一般特征的汇总。数据区分则是将目标类对象的一般特性与一个或多个对比类对象的一般特性比较。

（2）关联分析：发现关联规则，这些关联规则展示属性-值频繁地在给定数据集中一起出现的条件。关联分析广泛地用于购物篮分析或事务数据分析。

（3）分类与预测：分类是这样的过程，它找出描述并区别数据类或概念的模型（或函数），以便能够使用模型预测类标记未知的对象。在某些应用中，人们可能希望预测某些空缺的或不知道的数据值，而不是类标记。当被预测的值是数值数据时，通常称为预测。

（4）聚类分析：与分类与预测不同，聚类分析数据对象，而不考虑已知的类标记。一般情况下，训练数据中不提供类标记，因为不知道从何开始。聚类，可以用于产生这种标记。

（5）孤立点分析：数据库中可能包含一些数据对象，它们与数据的一般行为或模型不一致。这些数据对象是孤立点。孤立点可以使用统计试验检测。

（6）演变分析：数据演变分析描述行为随时间变化的对象的规律或趋势，并对其建模。

1.2　基于时间序列的数据挖掘介绍

1.2.1 什么是时间序列

时间序列是在时间轴方向上记录的一段有限的实数值序列（如图 1-1 所示）。在不同的场合下，可能使用不同的名称。有时，我们称它为对象，有时，又称它为序列，而在小波理论研究中，我们又称其为信号。在本文的论述过程中，可能会不加以区别地使用这些名称，但必须明确一点，它们都表示同一个事物，只不过在不同的上下文中以及当我们讨论数据某个方面的性质时，使用某个名称会更贴切。

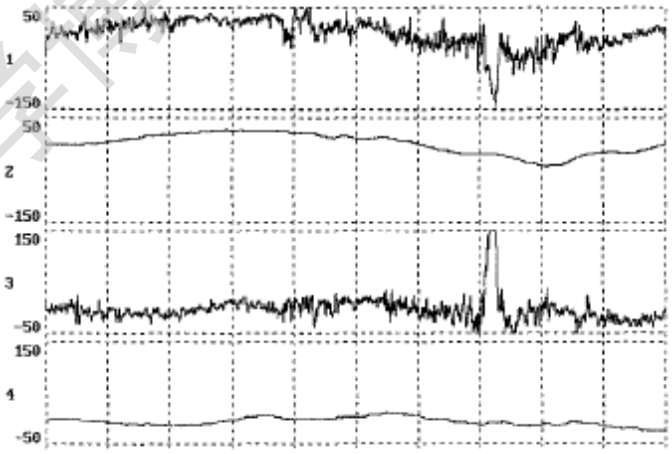

图 1-1 时间序列的例子

在日常生活中，在不同的领域中都会产生大量的时间序列数据，我们

可以简称为“时序数据”。通过收集、记录和整理这些数据，并配以先进的数据挖掘工具，我们就能够从时间序列中找到很多对现实生活极具价值的一些新东西，进而用来指导我们的工作和生活。目前，在商业领域中，对时间序列研究成果应用较为成功的行业包括医疗、金融、气象等，比如，医生可以通过对脑电图的分析进行病理诊断，股票分析家可以利用股票的历史数据预测股票的未来行情，气象部门也可以通过历年积累的数据进行预报工作。可以这么说，时间序列数据库就象一座价值不可估量的金矿，等待我们人类用智慧去开采它们。

既然谈到时间序列，很自然就会联想到时间序列分析。时间序列分析是根据系统观测得到的时间序列数据，通过曲线拟合和参数估计来建立数学模型的理论和方法。它一般采用曲线拟合和参数估计方法（如非线性最小二乘法）进行。时间序列分析常用在国民经济宏观控制、区域综合发展规划、企业经营管理、市场潜量预测、气象预报、水文预报、地震前兆预报、农作物病虫灾害预报、环境污染控制、生态平衡、天文学和海洋学等方面。

对时间序列分析，我们需要对时间序列进行建模。时间序列建模基本步骤是：①用观测、调查、统计、抽样等方法取得被观测系统时间序列动态数据。②根据动态数据作相关图，进行相关分析，求自相关函数。相关图能显示出变化的趋势和周期，并能发现跳点和拐点。跳点是指与其他数据不一致的观测值。如果跳点是正确的观测值,在建模时应考虑进去，如果是反常现象，则应把跳点调整到期望值。拐点则是指时间序列从上升趋势突然变为下降趋势的点。如果存在拐点，则在建模时必须用不同的模型去分段拟合该时间序列，例如采用门限回归模型。③辨识合适的随机模型,进行曲线拟合,即用通用随机模型去拟合时间序列的观测数据。对于短的或简单的时间序列，可用趋势模型和季节模型加上误差来进行拟合。对于平稳时间序列，可用通用 ARMA 模型（自回归滑动平均模型）及其特殊情况的