

时间序列数据挖掘综述^{*}

贾澎涛¹, 何华灿¹, 刘 丽¹, 孙 涛²

(1. 西北工业大学 计算机学院, 西安 710072; 2. 西安科技大学 建筑与土木工程学院, 西安 710054)

摘 要: 在综合分析近年来时间序列数据挖掘相关文献的基础上, 讨论了时间序列数据挖掘的最新进展, 对各种学术观点进行了比较归类, 并预测了其发展趋势。内容涵盖了时间序列数据变换、相似性搜索、预测、分类、聚类、分割、可视化等方面, 为研究者了解最新的时间序列数据挖掘研究动态、新技术及发展趋势提供了参考。

关键词: 时间序列; 数据挖掘; 相似性搜索; 模式发现

中图分类号: TP311.13 文献标志码: A 文章编号: 1001-3695(2007)11-0015-04

Overview of time series data mining

JIA Peng-tao¹, HE Hua-can¹, LIU Li¹, SUN Tao²

(1. School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China; 2. School of Architecture & Civil Engineering, Xi'an University of Science & Technology, Xi'an 710054, China)

Abstract: Based on analyzing many literatures about time series data mining, this paper showed the development, academic opinions and proposed about time series data mining recently. The contents contained data transformation, similarity match, prediction, classification, clustering, segmentation and visualization of time series. The aim was to compare and classify these literatures and put forward reference for scholars who research development, new techniques and trends of time series data mining.

Key words: time series; data mining; similarity search; pattern discovery

数据挖掘是知识发现过程中的一个步骤。它主要是利用某些特定的知识发现算法, 在一定的运算效率限制下, 从数据中挖掘出有价值的知识^[1]。原则上讲, 数据挖掘可以应用于任何类型的信息源。这包括关系数据库、数据仓库、事务数据库、其他高级数据库系统、平面文件(flat files) 和 WWW 上的数据^[2]。在这些数据集之中, 有一类数据集的数据之间存在着时间上的关系, 这类数据被称为时间序列。在对时间序列进行数据挖掘的过程中, 必须考虑数据集之中数据间存在的时间关系, 这类数据挖掘称为时间序列数据挖掘(time series data mining, TSDM)。Keogh^[1] 认为时间序列是普遍存在的, 图像数据、文本数据、影像数据、手写体数据、脑扫描数据等都可看做是时间序列。研究如何有效地从这些复杂的海量时间序列中挖掘潜在的有用知识, 具有重要的理论价值和现实意义。因此 TSDM 已成为数据挖掘研究的一个重要分支。

1 时间序列挖掘

1.1 时间序列的定义

从时间序列的角度来看, 每个数据单元可以被抽象为一个二元组(t, o)。其中: t 为时间变量; o 为数据变量, 反映数据单元的实际意义, 如某种商品的销售金额、股票的价格等。由此, 对于时间序列可以给出如下定义:

定义 1 时间序列 R 是一个有限集{ (t_1, o_1) , (t_2, o_2) , ... ,

(t_n, o_n) } , 满足 $t_i < t_{i+1}$ ($i = 1, 2, \dots, n-1$) 。

由时间序列组成的数据库称为时间序列数据库。针对时间序列数据库的挖掘就是时间序列数据挖掘。时间序列数据挖掘是时间序列数据库中知识挖掘的一个步骤, 它发现时间序列数据中的时态模式或模型^[3]。

1.2 时间序列挖掘的任务

从不同的角度出发, TSDM 的研究者给出了不同的 TSDM 的综述。这些综述的侧重点各不相同^[3~8], 有的侧重时间序列模式的聚类与分类^[3], 有的侧重阐述当前研究中存在的不足^[4], 有的侧重时间序列相似搜索^[6], 有的侧重模式发现与时间序列预测^[7~8]。因此这些研究者对 TSDM 的任务理解也有差异。由于人们对时间序列研究目的不同, 自然会对 TSDM 的研究任务提出不同的观点。综合收集到的资料, 将 TSDM 的研究任务归纳为如下六点: 时间序列相似性搜索; 时间序列聚类; 时间序列分类; 时间序列相关规则提取与模式分析; 海量时间序列可视化; 时间序列预测。

2 时间序列数据挖掘的主要研究内容

TSDM 始于 20 世纪 90 年代开始的数据库相似搜索的研究。R. Agrawal 等人^[9] 于 1993 年发表了第一篇关于时间序列数据库相似搜索的研究论文; C. Faloutsos 等人^[10] 于 1994 年提出了一种子序列相似性查询的方法。这两篇文章奠定了时

收稿日期: 2006-10-18; 修返日期: 2007-01-30 基金项目: 国家自然科学基金资助项目(50474041)

作者简介: 贾澎涛(1977-), 女, 陕西蒲城人, 博士研究生, 主要研究方向为数据挖掘、人工智能原理及应用(jiapengtao@ xust. edu. cn); 何华灿(1938-), 男, 教授, 博导, 主要研究方向为人工智能基础及其应用、泛逻辑学; 刘丽(1978-), 女, 博士研究生, 主要研究方向为模糊控制、泛逻辑控制、人工智能原理及应用; 孙涛(1976-), 男, 硕士, 主要研究方向为管理信息系统、数据挖掘。

间序列相似性研究的基础,引起了广大 TSDM 研究者的注意。文中提出的时间序列数据库相似性搜索算法应遵循的原则已被普遍接受。近十年来, TSDM 成为国际学术界研究的热点之一, 国际著名的学术会议和期刊如 SIGKDD、VLDB、ICDE、ACM SIGMOD 以及 IEEE Trans on Knowledge and Data Engineering 等每年都有不少关于 TSDM 的研究成果报道。综合来看, TSDM 研究的主要内容包括以下七个方面。

2.1 时间序列数据变换

时间序列通常是海量数据, 数据中可能存在大量的噪声, 直接对原始数据进行挖掘效率很低, 甚至不可行。因此就需要在时间序列数据挖掘之前对原始数据进行变换, 使得数据挖掘在更高层次的数据上进行。所谓时间序列数据变换就是将原始时间序列映射到某个特征空间中, 并用它在这个特征空间中的映像来描述原始的时间序列。这样可以实现数据压缩, 减少计算代价。目前已有的时间序列数据表示主要有离散傅里叶变换(discrete Fourier transform, DFT)^[9~12]、奇异值分解(singular value decomposition, SVD)^[1]、离散小波变换(discrete wavelet transform, DWT)^[13~16]、动态时间扭曲(dynamic time warping, DTW)^[17~21]、分段合计近似(piecewise aggregate approximation, PAA)^[1]、分段线性表示(piecewise linear representation, PLR)^[22~24]、分段多项式表示(piecewise polynomial representation, PPR)^[7]等。

2.2 时间序列数据库相似搜索

时间序列相似性搜索是基于内容的查询。由于时间序列自身数值上的连续性与波动性, 造成了时间序列相似性搜索的特有问题。

首先是如何定义相似性。相似性可能不仅仅依赖于人的主观意识, 还依赖于目前分析的任务, 甚至数据本身。目前时间序列相似性的度量主要是基于距离的度量, 包括 Euclidean 距离^[25~27]、非 Euclidean^[28~30]、DTW^[17~21, 28, 31]等。基于 Euclidean 距离的查找比基于 DTW 距离的查找速度快, 但 DTW 的查找结果要优于 Euclidean 距离。可以采用下界函数(lower bounding function) 加速 DTW 的查找速度^[18]。

其次, 时间序列是任意长度的流式序列, 而不是一个离散的元素。时间序列一般都是海量数据, 这就需要非常有效的索引机制来提高索引效率。目前相似性索引结构主要采用空间索引结构。从大的方面分, 空间数据索引技术可分为树结构^[32~34](包括 R 树、R* 树等) 和网格文件两类。

在相似性搜索中, R. Agrawal 和 C. Faloutsos 等人提出的时间序列数据库相似性搜索算法应遵循的原则值得关注。其中最重要的一条就是相似搜索算法应满足非漏报性质^[9, 10], 即理论上算法应能将数据库中全部满足相似条件的序列都检索出来, 没有遗漏。要满足非漏报原则, 就要求相似查询算法所采用的数据表示满足如下条件:

$$D_F(Q, C) \leq D(Q, C) \quad (1)$$

其中: Q 是查询序列; C 是时间序列数据库中的任意序列; D_F 是采用某种高级数据表示后两个序列之间的距离; D 是两个序列之间的真实距离。

研究者对时间序列数据库相似搜索问题的研究取得了大量的成果。目前, 时间序列数据库相似搜索仍然是 TSDM 领域

的研究热点之一。

2.3 时间序列聚类/分类分析

在对时间序列进行分类和聚类时, 需要先根据用户指定的模式长度和时间粒度, 将原始时间序列数据分割为等长度的子序列集合或不重叠的子序列集合(这些子序列一般不等长); 然后再对这些子序列进行聚类和分类分析。

对时间序列进行聚类的算法有基于相似性(或距离)^[35]、基于特征^[36, 37]、基于模型^[38]和基于分割的聚类分析^[7, 39]。其中, 在基于相似性的方法中, 一种常用的距离度量是欧氏距离或者在此基础上的一些改进作为相似性测度。由于欧氏距离测度针对的是确定性向量空间, 而时间序列的长度通常是变化的、对时间变化敏感, 并且不能高效地表示为有限维空间的一个点。传统的聚类分析大多是基于向量的, 它们不能很好地解决时间序列聚类问题。近年来对时间序列的聚类研究更多地使用基于模型的聚类分析, 如基于 HMM 的时间序列聚类^[40]。近年来, 时间序列数据挖掘中的聚类技术发展很快, 取得了很多研究成果。

时间序列分类就是给定一个未知类别的时间序列, 将其划分到某些预定义的类别之中。许多分类算法在时间序列中都有应用, 如决策树、神经网络、贝叶斯分类器等。近年来用分类器融合对时间序列进行分类成为了一个热门研究方向。

2.4 时间序列数据可视化

时间序列可视化挖掘是 TSDM 一个较新的研究领域, 也是一个有广阔应用前景的研究领域。所谓时间序列数据可视化挖掘就是利用图形图像技术、虚拟现实技术以及数据挖掘技术, 将复杂的时间序列以人们易于理解的、直观的和图形化的方式呈现出来。时间序列可视化是一个应用前景广阔的研究方向^[41~44]。目前国外研究较多, 也开发出了相应的可视化工具, 如 time series spirals、theme river、time searcher、vizTree、time series bitmaps 等。国内这方面的研究成果较少。

2.5 时间序列分割与模式发现

模式发现是 TSDM 的重要研究内容之一, 并出现了大量的研究成果。针对不同的应用目的, 人们试图从时间序列数据库中发现的模式也各不相同, 如特定模式、频繁模式、周期模式、感兴趣模式、惊奇模式、异常模式、例外模式等。为了从一个时间序列中抽取模式, 需要某种算法将一个长时间序列分割为若干个相对短的子序列, 以便对这些子序列进行聚类/分类分析^[45, 46]、检测时间序列中的变化点^[47]、对分割后的时间序列建立动态模型^[48, 49]。

时间序列分割主要有两个应用: 系统模型变化检测, 即当产生时间序列的系统的模型(或参数) 发生变化时, 应用分割算法可以检测到这种变化是何时发生的; 应用分割算法创建时间序列的高级数据表示, 以便对时间序列进行索引、聚类和分类^[1]。因此, 研究时间序列分割算法具有重要的理论价值和现实意义, 并已成为时间序列数据挖掘研究的主要研究内容之一。

2.6 时间序列预测

时间序列预测一直是众多科学领域感兴趣的问题。由于研究对象的高度一致性, TSDM 研究中关心的时间序列预测问题与时间序列分析理论中关心的时间序列预测问题在许多方

面是相同的。时间序列预测根据时间序列型数据,由历史的和当前的数据推测未来的数据,也可以认为是以时间为关键属性的关联知识。时间序列预测技术大体可分为:

a) 传统的线性时间序列预测技术。1968 年 Box 和 Jenkins 提出了一套比较完善的时间序列建模理论和分析方法。这些经典的数学方法通过建立随机模型,如自回归模型、自回归滑动平均模型、求和自回归滑动平均模型和季节调整模型等进行时间序列的预测。

b) 非线性时间序列预测技术。这类方法主要采用嵌入空间法或神经网络等方法,特别是混沌时间序列预测^[50]和基于神经网络的时间序列预测^[51,52]。

混沌时间序列预测是建立在 Takens^[53]提出的嵌入定理和相空间重构理论基础上的。其目的是试图在高维相空间中恢复混沌吸引子。时间序列预测的神经网络模型包括模糊神经网络、径向基函数(RBF) 网络、小波神经网络以及积单元神经网络等。

c) 其他技术,如滑动窗口二次自回归模型^[7]、基于云模型的时间序列预测^[54]等。

2.7 TSDM应用研究

时间序列数据广泛存在于现实世界的各个领域,因此 TS-DM的应用十分广泛。典型的应用包括股票预测、机电系统诊断、医学诊断、生物信息学、营销指导、运动图像分析、生产过程监测等。在每一篇时间序列数据挖掘论文中都给出了实际应用,但其中许多文章介绍的思想到底能不能用于实际令人怀疑。研究者应将更多的注意力放在那些真实的、有实际价值的应用上。

3 未来的研究方向

下面罗列出时间序列挖掘研究中尚待解决的问题及未来的研究方向,虽不能概括所有的尚待解决的问题及方向,但可供研究者们参考。它们是:

a) 海量时间序列数据查询。传统的时间序列数据库通常按照数据来源或时间关系进行索引。这样的索引结构在基于相似查询场合,其效率极其低下,特别是对海量时间序列数据库更是不可行的。因此需要建立针对海量时间序列数据的索引结构,以便在其之上对海量数据作相似性查询,从而获得较高的查询效率。

b) 主题发现。与相似性搜索相比,主题发现(motif discovery) 有更广的应用性。但现有的主题发现算法往往得到太多的无意义的主题,因此如果能够给出评价主题的规则,让系统在众多主题中自动不遗漏地提取更有意义的主题是笔者进一步研究的目标。

c) 聚类时间序列流。在进行时间序列聚类时,一般先将时间序列分割成子序列,然后对子序列进行聚类。这样有可能意义不大。能不能直接在时间序列流上进行更有意义的聚类是今后时间序列聚类研究的一个方向。

d) 有效的时间序列分类算法。目前有很多时间序列分类算法,如决策树、神经网络、贝叶斯分类器等,但所有的这些方法都不如基于 DTW 的 1 邻近算法。DTW 是很有效的线性方法,然而 1 邻近需要访问每个属性,效率很低。因此研究更有

效的时间序列分类算法是亟待解决的问题。

e) 海量时间序列可视化。最好的数据挖掘和模式识别工具就是人的眼睛。那么能不能将海量的时间序列中隐含的知识,如规则、孤立点、异常事例等,以人们易于理解的方式可视化,是人们普遍关心的问题。

f) 多时间粒度的数据挖掘与知识发现。在不同时间尺度研究系统或现象的行为,有助于加深人们对该系统或现象的认识。因此,多时间粒度的数据挖掘与知识发现是目前研究的热点问题。

g) 新的 TSDM 算法。在相似性搜索时,给定两个时间序列,如果事先不将子序列标准化,找到所有相似的子序列是很容易的,但这是没有意义的。问题在于不是追求速度,而是以何种方式搜索相似更有意义。

在进行分类聚类时,有时结果意义不大甚至是无意义的。给定两个聚类(或分类) 在一起的时间序列,如果能自动构建对为什么把它们放在一起的解释那将会更有意义。因此,研究新的 TSDM 算法以更有意义的方式进行挖掘,是研究者们致力解决的问题。

参考文献:

[1] EOGH E. Data mining and machine learning in time series database [C] //Proc of the 5th Industrial Conference on Data Mining (IC-DM) . Leipzig: [s. n.], 2005.

[2] HAN Jia-wei, KAMBER M. Data mining: concepts and techniques [M] . New York: Morgan Kaufmann Publishers, 2000: 7-9.

[3] LIN Wei-qiang, ORGUN M A, WILLIAMS G J. An overview of temporal data mining[C] //Proc of the 1st Australian Data Mining Workshop. Canberra: University of Technology. 2002: 83-90.

[4] KEOGH E, KASETTY S. On the need for time series data mining benchmarks: a survey and empirical demonstration[C] //Proc of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2002: 102-111.

[5] HETLAND M L. A survey of recent methods for efficient retrieval of similar time sequences[EB/OL] . (2001) . <http://citeseer.nj.nec.com/hetland01survey.html>.

[6] HOPPNER K. Time series abstraction methods: a survey[C] //Proc of GI Jahrestagung Informatik, Workshop on Knowledge Discovery in Databases. Dortmund: [s. n.], 2002: 777-786.

[7] 李爱国. 时间序列数据分割与时态模式挖掘研究[D]. 西安: 西安交通大学, 2003: 2-4.

[8] LAXMAN S, SASTRY P S. A survey of temporal data mining[J] . S ādhan ā Academy Proceedings in Engineering Sciences, 2006, 31(2) : 173-198.

[9] AGRAWAL R, FALOUTSOS C, SWAMI A. Efficient similarity search in sequence databases[C] //Proc of the 4th International Conference on Foundations of Data Organization and Algorithms. London: Springer-Verlag, 1993: 69-84.

[10] FALOUTSOS C, RANGANATHAN M, MANOLOPOULOS Y. Fast subsequence matching in time-series databases[C] //Proc of the ACM SIGMOD International Conference on Management of Data. Mineapolis: ACM Press, 1994: 419-429.

[11] RAFIEI D, MENDELZON A O. Querying time series data based on similarity[J] . IEEE Trans on Knowledge and Data Engineering, 2000, 12(5) : 675-693.

[12] WANG Chang-zhou, WANG Xiao-yang. Multilevel filtering for high

- dimensional nearest neighbor search[C] //Proc of ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. Dallas: ACM Press, 2000: 37-43.
- [13] HUHTALA Y, K RKK INEN J, TOIVONEN H. Mining for similarities in aligned time series using wavelets[C] //Proc of Data Mining and Knowledge Discovery: Theory, Tools, and Technology. Orlando: [s. n.], 1999: 150-160.
- [14] STRUZIK Z R, SIEBES A. Measuring time series ' similarity through large singular features revealed with wavelet transformation[C] //Proc of the International Workshop on Database and Expert Systems Application. Florence: IEEE Computer Society Press, 1999: 162-166.
- [15] POPIVANOV I, MILLER R J. Similarity search over time-series data using wavelets[C] //Proc of the 18th International Conference on Data Engineering. San Jose: IEEE Computer Society Press, 2002: 212-221.
- [16] 张海勤, 蔡庆生. 基于小波变换的时间序列相似模式匹配[J]. 计算机学报, 2003, 26(3) : 373-377.
- [17] AGRAWAL R, LIN K I, SAWHNEY H S, *et al.* Fast similarity search in the presence of noise, scaling, and translation in time-series databases[C] //Proc of the 21st Conference on Very Large Data Bases. San Francisco: Morgan Kaufmann, 1995: 490-501.
- [18] KEOGH E. Exact indexing of dynamic time warping[C] //Proc of the 28th International Conference on Very Large Databases. Hong Kong: Morgan Kaufmann, 2002: 406-417.
- [19] RATH T M, MANMATHA R. Lower-bounding of dynamic time warping distances for multivariate time series, Technical Report MM-40 [R]. Amherst: Center for Intelligent Information Retrieval Technical Report, University of Massachusetts, 2003.
- [20] CHIU S, KEOGH E, HART D, PAZZANI M. Iterative deepening dynamic time warping for time series[C] //Proc of the 2nd SIAM International Conference on Data Mining. Baltimore: SIAM Press, 2002: 148-156.
- [21] KEOGH E, PAZZANI M J. Derivative dynamic time warping[C] //Proc of the 1st SIAM International Conference on Data Mining. Chicago: SIAM Press, 2001: 209-211.
- [22] KEOGH E, CHAKRABARTI K, PAZZANI M, *et al.* Dimensionality reduction for fast similarity search in large time series databases[J]. Journal of Knowledge and Information Systems, 2001, 3(3) : 263-286.
- [23] KEOGH E, CHAKRABARTI K, PAZZANI M. *et al.* Locally adaptive dimensionality reduction for indexing large time series databases [J]. ACM Transactions on Database Systems, 2002, 27(2) : 188-228.
- [24] GE X, SMYTH P. Deformable markov model templates for time-series pattern matching[C] //Proc of the 6th ACM SIGKDD Int 'l Conference on Knowledge Discovery and Data Mining. Boston: ACM Press, 2000: 81-90.
- [25] FERHATOSMANOGLU H, TUNCEL E, AGRAWAL D, *et al.* Approximate nearest neighbor searching in multimedia databases[C] //Proc of the 17th IEEE Int 'l Conference on Data Engineering. Heidelberg: IEEE Computer Society Press, 2001: 503-511.
- [26] KAHVECI T, SINGH A, GUREL A. An efficient index structure for shift and scale invariant search of multi-attribute time sequences [C] //Proc of the 18th Int 'l Conference on Data Engineering. San Jose: IEEE Computer Society Press, 2002.
- [27] LOH W, KIM S, WHANG K. Index interpolation: an approach to subsequence matching supporting normalization transform in time-series databases[C] //Proc of the 9th ACM CIKM Int 'l Conference on Information and Knowledge Management. McLean: ACM Press, 2000: 480-487.
- [28] WU Y, AGRAWAL D, ABBADI A E. A comparison of DFT and DWT based similarity search in time-series databases[C] //Proc of the 9th ACM CIKM Int 'l Conference on Information and Knowledge Management. McLean: ACM Press, 2000: 488-495.
- [29] CARA -VALENTE J P, LOPEZ-CHAVARRIAS I. Discovering similar patterns in time series[C] //Proc of the 6th ACM SIGKDD Int 'l Conference on Knowledge Discovery and Data Mining. Boston: ACM Press, 2000: 497-505.
- [30] SHIM K, SRIKANG R, AGRAWAL R. High-dimensional similarity joins[J]. IEEE Trans on Knowledge and Data Engineering, 2002, 14(1) : 156-171.
- [31] KIM S, PARK S, CHU W. An index-based approach for similarity search supporting time warping in large sequence databases[C] //Proc of the 17th International Conference on Data Engineering. Heidelberg: IEEE Press, 2001: 607-614.
- [32] BECKMANN N, KRIEGEL H P, SCHNEIDER R, *et al.* The R* -tree: an efficient and robust access method for points and rectangles [C] //Proc of ACM SIGMOD International Conference on Management of Data. Atlantic: ACM Press, 1990: 322-331.
- [33] TAO Y, PAPADIAS D, SUN J. The TPR* -tree: an optimized spatio-temporal access method for predictive queries[C] //Proc of the 29th VLDB Conference. Berlin: Morgan Kaufmann Publishers, 2003: 790-801.
- [34] TAO Y, PAPADIAS D. MN3R-tree: a spatio-temporal access method for timestamp and interval queries[C] //Proc of the 27th VLDB Conference. Roma: Morgan Kaufmann Publishers, 2001: 431-440.
- [35] KALPAKIS K, GADA D, Puttagunta V. Distance measures for effective clustering of ARIMA time-series[C] //Proc of the IEEE International Conference on Data Mining. San Jose: IEEE Computer Society Press, 2001: 273-280.
- [36] M RCHEN F. Time series feature extraction for data mining using DWT and DFT[EB/OL]. (2003-11-05). <http://www.mybytes.de/papers/moerchen03time.pdf>.
- [37] ZHANG Hui, HO Tu-bao, LIN Mao-song. A non-parametric wavelet feature extractor for time-series classification[C] //Proc of the 8th Pacific-Asia Conf Knowledge Discovery and Data Mining. Berlin: Springer, 2004: 595-603.
- [38] VAITHYANATHAN S, DOM B. Model-based hierarchical clustering [C] //Proc of the 16th Conference Uncertainty in Artificial Intelligence. Stanford, California: Morgan Kaufmann, 2000: 599-608.
- [39] KEOGH E, CHU S, HART D, *et al.* An on-line algorithm for segmenting time series[C] //Proc of the IEEE Int 'l Conference on Data Mining. San Jose: IEEE Computer Science Press, 2001: 289-296.
- [40] 段江娇, 薛永生, 林子雨, 等. 一种新的基于隐 Markov 模型的分层时间序列聚类算法[J]. 计算机研究与发展, 2006, 43(1) : 61-67.
- [41] YU J, HUNTER J, REITER E, *et al.* Recognising visual patterns to communicate gas turbine time-series data[C] //Applications and Innovations in Intelligent Systems. London: Springer, 2002: 105-118.
- [42] HOCHHEISER H, SHNEIDERMAN B. Visual queries for finding patterns in time series data[R]. Maryland: University of Maryland, 2002.

较, 其效果自然不是很理想。以上仿真还说明, 在线调度模式下, TD_MCT 算法的调度结果明显优于 TD_OLB 算法。因此, 在线调度一般采用 TD_MCT 为佳。此外, 从仿真结果还可看出, 在 TD_min-min、TD_max-min、TD_sufferage 三种算法的性能评比中, TD_sufferage 总体上是最好的一种算法。

在传统的基于 makespan 的目标调度问题中, min-min 的性能总是比 max-min 算法要好^[2]。但从以上仿真结果可以看出, 在基于信任驱动的调度问题中, 随着 V_m 、 V_q 值的增加, 即随着具有强信任关系任务数目的增加和机器异构性的增加(这正是网格的主要特点之一), TD_max-min 算法的性能逐渐明显优于 TD_min-min。这进一步说明研究基于信任机制的调度算法具有重要的理论与实用价值。比如, 当 $V_q = 1$ (此时强信任关系任务数目占有所有任务数目的 25%, 弱信任关系任务数占总任务数目的 25%, 有 50% 的是无信任关系的任务), 机器异构性低($V_m = 0.1$) 时, TD_min-min 得出的平均信任效益值比 TD_max-min 得出的平均效益稍好; 但是当机器异构性增高时($V_m = 0.8$), TD_min-min 与 TD_max-min 得出的平均效益值基本一致。又比如, 当 $V_q = 2$ (此时强信任关系任务数目占有所有任务数目的 50%, 弱信任关系任务数目占总任务数目的 50%, 无信任关系的任务数目为 0), 机器异构性低($V_m = 0.1$) 时, TD_max-min 得出的平均效益值要高于 TD_min-min 得出的平均信任效益值, 与 TD_sufferage 得出的平均信任效益值不分上下; 但是当机器异构性高时($V_m = 0.8$), TD_max-min 得出的平均效益值高于 TD_min-min 的值。

从以上仿真结果还可以看出, 当机器异构性确定时, 任务的异构性并不影响所有算法的优劣顺序。

4 结束语

将信任机制与任务调度机制有效融合是真实网格环境中大规模分布式资源管理的难题之一。本文提出了基于信任机制的网络任务调度新算法; 同时, 在一致的 ETC 矩阵的条件下, 通过调节机器异构和任务异构性, 变动强、弱信任关系的任务数目, 对所有基于信任机制的算法进行仿真分析和比较研究。结果表明, 基于信任机制的调度算法不仅优于传统基于时

间的调度算法, 而且在传统基于 makespan 的目标调度问题中, min-min 的性能总是优于 max-min 的结论不同, 当强信任关系任务的数量大于弱信任关系及无信任关系的任务数时, TD_max-min 算法明显优于 TD_min-min 算法。这个发现进一步说明研究基于信任机制的网络调度算法的理论和实际意义。

参考文献:

[1] UYYA R, ABRAMSON D, VENUGOPAL S. The grid economy [J]. Proceedings of the IEEE, 2005, 93(3): 698-714.

[2] BRAUN R D, SIEGEL H J, BECK N, *et al.* A comparison study of static mapping heuristics for a class of meta-tasks on heterogeneous computing systems [C] // Proc of the 8th HCW. San Juan: IEEE Computer Society Press, 1999: 15-29.

[3] HE Xiao-shan, SUN Xian-he, LASZEWSKI G von. QoS guided min-min heuristic for grid task scheduling[J]. Journal of Computer Science and Technology, 2003, 18(4): 442-451.

[4] DO AN A, ZG NER F. Scheduling of a meta-task with QoS requirements in heterogeneous computing systems[J]. Journal of Parallel and Distributed Computing, 2006, 66(2): 181-196.

[5] AZZEDIN F, MAHESWARAN M. Integrating trust into grid resource management systems[C] // Proc of International Conference on Parallel Processing. Vancouver: IEEE Computer Society Press, 2002: 47-54.

[6] 王东安, 徐浩, 南凯, 等. 基于推荐的网格计算的信任模型[J]. 计算机应用研究, 2006, 23(2): 96-98.

[7] 郭学理, 杨慧, 何鹏. 一种基于信任度的网格计算资源分配模型[J]. 计算机应用研究, 2005, 22(2): 93-95.

[8] 张伟哲, 刘欣然, 云晓春, 等. 信任驱动的网络任务调度算法[J]. 通信学报, 2006, 27(2): 73-79.

[9] 潘海军, 陆魁军, 吴朝晖. 基于网格系统的信任量化研究[J]. 计算机应用研究, 2005, 22(10): 49-51.

[10] 黄宝边, 曾文华. 网格计算中基于信任机制的动态任务调度[J]. 计算机应用, 2006, 26(1): 65-69.

[11] ALI S, SIEGEL H J, MAHESWARAN M, *et al.* Representing task and machine heterogeneities for heterogeneous computing systems [J]. Tamkang Journal of Science and Engr, 2000, 3(3): 195-207.

(上接第 18 页)

[43] OCHHEISER H, SHNEIDERMAN B. Visual specification of queries for finding patterns in time series data[D]. Maryland: University of Maryland, 2001.

[44] HOCHHEISER H. Interactive graphical querying of time series and linear sequence data sets[D]. Maryland: University of Maryland, 2003.

[45] 李斌, 谭立湘, 章劲松, 等. 面向数据挖掘的时间序列符号化方法研究[J]. 电路与系统学报, 2000, 5(2): 9-14.

[46] 覃征, 李爱国. 时间序列数据的稳健最优分割[J]. 西安交通大学学报, 2003, 37(4): 338-342.

[47] SRIPADA S G, REITER E, HUNTER J, *et al.* Segmenting time series for weather forecasting[C] // Proc of ES. 2002.

[48] HANLON B, FORBES C. Model selection criteria for segmented time series from a bayesian approach to information compression[R]. [S.

l.]: Monash University, 2002.

[49] HAWKINS D M. Fitting multiple change-point models to data[J]. Computational Statistic & Data Analysis, 2001, 37(3): 323-341.

[50] 刘涵, 刘丁, 李琦. 基于支持向量机的混沌时间序列非线性预测[J]. 系统工程理论与实践, 2005, 25(9): 94-99.

[51] 张志华, 郑南宁, 郑海兵. 径向基函数神经网络(RBF) 的一种极大熵学习算法[J]. 计算机学报, 2001, 24(5): 474-479.

[52] 陈哲, 冯天瑾, 张海燕. 基于小波神经网络的混沌时间序列分析与相空间重构[J]. 计算机研究与发展, 2001, 38(5): 591-596.

[53] TAKENS F. Detecting strange attractors in turbulence[C] // RAND D, YOUNG L S. Dynamical systems and turbulence, lecture notes in mathematics. New York : Springer-Verlag, 1981: 336-381.

[54] 蒋嵘, 李德毅. 基于形态表示的时间序列相似性搜索[J]. 计算机研究与发展, 2000, 37(5): 601-608.