科学数据网格中面向天文应用的问题求解环境设计与实现*

秦 欣, 罗 泽, 南 凯, 阎保平
(中国科学院 计算机网络信息中心, 北京 100080)

摘 要: 通过对虚拟天文台中科学课题的需求分析, 提出一种基于数据网格针对天文应用的问题求解环境的框架; 以"2MASS 巡天星表搜寻 OB 星协候选体研究银河系的旋臂结构"科学课题为例, 在科学数据网格上设计和实现了面向天文应用的问题求解环境, 并对实现中的关键技术进行研究。

关键词: 问题求解环境; 数据网格; 工作流
中图分类号: TP393 **文献标志码**: A **文章编号**: 1001-3695(2007)04-0242-04

Design and Implementation of Problem Solving Environment for Astronomy Application Based on Science Data Grid

QIN Xin, LUO Ze, NAN Kai, YAN Bao-ping
(Network Information Center, Chinese Academy of Sciences, Beijing 100080, China)

Abstract: This paper proposed a framework of Problem Solving Environment (PSE) for astronomy application based on science data grid, after analyzing the requirement of science research in virtual observatory. It designed and implemented a material PSE on science data grid for the astronomic application of "Detect arm structure of the Milky Way through OB associations mining from 2MASS", researched key problems in the implementation.

Key words: problem solving environment; data grid; workflow

在过去十多年中, 随着天文探测器制造技术的高速发展, 数据收集能力也遵循着类似摩尔定律而增长, 天文学开始进入一个信息富庶的时代。天文数据具有海量、开放、高复杂性、多波段高相关性、较好归档等特点, 决定天文学是数据网格完美的测试平台, 同时网格技术为大规模天文应用的实现提供了关键的技术支撑平台, 使得天文研究和网格技术[1]互相结合, 诞生了虚拟天文台[2]。

科学数据网格是以中国科学院"十五"信息化建设重大项目"科学数据库及其应用系统[3]"为背景而建设的一个数据网格; 同时, 科学数据网格项目得到国家"863"计划的支持, 是"十五""863"重大专项"高性能计算机及其核心软件"中的一个应用网格。在科学数据网格的建设中, 中国科学院计算机网络信息中心与中国科学院国家天文台合作, 开发虚拟天文台作为科学数据网格的示范应用系统。虚拟天文台中开展许多试验和应用, 如图 1 所示, 需要创建一个协同的环境来反复实现复杂问题求解处理。在这个环境中, 天文工作者依照事先拟订好的试验步骤, 协同、反复多次使用大量的分布、异构数据检索、数据计算、可视化、数据挖掘等资源, 甚至可接受本地开发的程序, 组合成有效的处理, 进而开展天文研究工作。

2MASS 巡天星表搜寻 OB 星协候选体研究银河系的旋臂结构(简称 2MASS-OB 天文课题)是研究银河系整体的结构特征, 旋臂的形成机制、银河系的演化等问题的一个重要项目。本文通过对天文课题的需求分析, 提出一种基于数据网格针对天文应用的问题求解环境的框架。

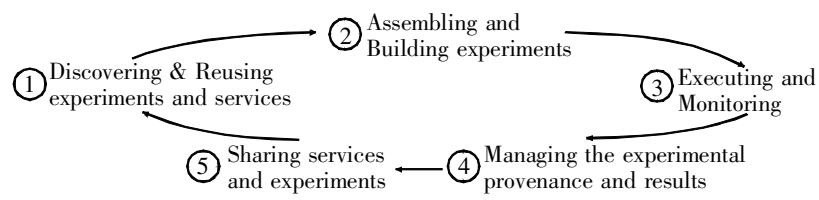

图 1 天文研究的生命周期

1 相关工作

1.1 传统的天文研究计算环境

传统的天文研究多是在分散的、孤立的计算环境中进行的, 往往是个人或者小团体针对特定的天文课题自编程序实现。例如天文课题运用神经网络算法研究星系形态[4], 还有通过查询 2MASS 数据库研究恒星形成[5]。在这种计算环境下, 数据格式不统一、程序和接口不规范、开发效率低、数据和过程可共享性低, 很难被其他人再利用。

收稿日期: 2005-09-29; 修返日期: 2006-01-18 基金项目: 国家"十五""863"计划资助项目(2002AA104240); 中国科学院"十五"信息化建设重大资助项目(INF105-SDB)

作者简介: 秦欣(1981-), 女, 硕士研究生, 主要研究方向为网格技术与应用(qinxin@cnic.cn); 罗泽(1976-), 男(傣族), 博士, 主要研究方向为网格中间件、大规模数据库应用及计算机网络; 南凯(1974-), 男(满族), 副研究员, 主要研究方向为高性能计算应用、大规模数据库应用和计算机网络; 阎保平(1950-), 女, 研究员, 博导, 主要研究方向为下一代互联网络、大型网络及系统集成工程化规范管理、大规模数据库应用技术、多媒体数据库、智能搜索引擎.

1.2　虚拟天文台现状

1999 年各国天文学界纷纷提出了自己的虚拟天文台计划, 开始天文学领域的全球协作。2002 年 6 月, 国际虚拟天文台联盟(IVOA) 成立; 2002 年 10 月, 中国虚拟天文台成为国际虚拟天文台联盟成员。为了逐步实现不同组织之间的互操作, 国际虚拟天文台(IVO) 目前就许多问题在国际范围内达成一致, 如制定数据和接口的统一标准, 协调天文的软件包、源代码库、开发工具等。除此之外, 结合 Grid/Web 技术, 以服务的形式封装分布的异构的数据资源, 进一步加强共享和协作。

然而, 在目前虚拟天文台的应用项目中多是采用定制任务引擎的方法完成服务的组合。这种固定的序列组合方式有两个缺点: ①易重构性不好。当科学工作者需要对应用的步骤进行调整和改动时, 必然也将对这个定制的任务引擎进行代码级的改动。②用户界面不友好。通过命令和程序控制后台的任务提交和执行; 不支持可视化; 在设计和部署天文试验上给不熟悉命令和程序的天文科学工作者带来极大的不便。

1.3　Web Service 工作流

工作流有数十年的历史, 1993 年工作流管理联盟的成立标志着工作流技术开始进入相对成熟的阶段。工作流是一类能够完全或者部分自动执行的业务过程, 根据一系列的过程规则, 文档、信息或数据在不同的执行者之间传递和执行。传统工作流的主要应用集中在管理和生产过程的自动化, 特别是在企业和大组织内部。1995 年后, 企业应用集成成为工作流的一个新的应用领域, Web 技术因其开放、跨平台的特性使基于 Web 的工作管理系统已经成为一种必然的发展趋势。

具有代表性的 Web Services 工作流描述语言有 WSFL 和业务流程建模语言 BPML。WSFL(Web Services Flow Language) 由 IBM 提出, 符合 Web Services 层次模型, 基于 XML 描述 WS 交互的一种网络服务工作流描述语言。WSFL 使用一种流模型(Flow Models) 和总体模型(Global Models) 来描述网络服务的组成。流模型说明了如何使用服务所提供的功能来描述商业交易流程; 总体模型则详细说明所有交易伙伴的交易情形, 即服务如何与其他服务交互。流模型定义了一系列活动表现复合网络服务的运作, 并确定活动执行的顺序。它分别使用 controlLink 和 dataLink 定义各种活动之间的控制和数据流。

业务流程建模语言(Business Process Management Language, BPML) 是一种基于 XML 为业务流程建模的元语言。BPML 为协同事务性业务流程提供了一个虚拟执行模型, 它基于一个叫做事务有限状态机的概念。

因此, Web Services 工作流是实现协同工作环境的一个良好方法, 它能方便地与原有的应用、信息集成。然而它是基于 Web Services 层次模型, 适应商业流程, 开销往往比较大, 而且不是免费的, 不太满足科学研究的需求。

2　设计与实现

2.1　天文课题的需求

2MASS-OB 课题是由中国国家天文台提出的重要天文应用。其基本思想是利用已观测到的 OB 星在 2MASS 天文数据库中查找 OB 星的光谱范围, 再将属于这个有效光谱范围的银河系天体可视化, 通过相应的数据挖掘, 帮助研究银河系的旋臂结构[6]。

由此得出天文研究的迫切需要: ①提供分布、异构数据资源的统一访问和集成。因为原始数据信息一般都分布在不同的地理位置, 数据的存储格式及存储管理系统各不相同。②易用、可视化的问题求解环境, 帮助科学家快速构造和修改试验流程。在这个平台上定义通用的过程模型, 支持用户在图形编辑环境下从库中选择各种有效的组件构造具体科学问题的求解模型, 并自动收集执行的元数据。

2.2　问题求解环境层次模型

通过对天文课题的需求分析, 设计出在科学数据网格环境下面向天文应用的问题求解环境, 以改善传统的天文研究环境, 并借鉴 Web Services 工作流的优势, 丰富以服务为中心的网格体系结构。

问题求解环境本质上是透明、可扩展、易重构的网格环境。图 2 所示的四层模型描述了在问题求解环境中不同抽象级别上资源和它们之间的互操作。

(1) 应用表示层

应用表示层面向最终用户, 提供友好的、可视化的界面, 便于快速构造、修改和管理科学问题求解的流程。它从面向问题的层面上展现工作流, 隐藏服务间交互的复杂细节。输入起始试验数据后, 用户只用对逻辑服务产生或者接收的数据进行适当连接就能完成工作流中服务的组合, 而不必关心服务的实现, 也不用了解面向服务的体系结构(Service-Oriented Architecture) 的概念和技术细节。

(2) 工作流执行层

工作流执行层扩展和检查上层中由用户设定的组合逻辑, 完成工作流程执行细节。网格工作流描述语言记录应用表示层中用户在可视化界面提出的请求, 并转换成包含控制流和数据信息的一组过程规则; 进而工作流引擎把各种互相作用的网格服务或者本地的开发应用(如图 2 中的 “Math Computing”) , 按照规则有机地结合在一起, 其间的文档、信息或数据在不同的网格服务之间传递和执行。该层主要完成语义的实现, 并不关心具体服务的执行、维护。

(3) 基于 OGSA 或 WSRF 的网格中间件层

网格中间件层封装数据资源和数学运算为网格服务(如图 2 中的网格资源 A、B、C、D、E) , 以资源的形式和统一的接口提供给上层用户。它由 Globus Toolkit 实现, Globus 在网格计算逻辑结构中的组成部分包括网格安全基础设施、Globus 资源分配管理 GRAM、元计算目录服务、GridFTP 服务等。

(4) 数据资源层

数据资源层包含是位于网络中分布的、自治的、异构的数据资源, 如 2Mass-OB 试验需要访问的 “J/AJ/117/354”数据表、Hipparcos 和 2MASS 数据库。

2.3　网格数据服务

2MASS-OB 课题需要访问的天文数据库有 Hipparcos 和 2MASS, 如表 1 所示。网格数据服务以可控的方式将这些物理数据资源暴露于网格环境中, 通过公共通用的接口支持对异构物理数据资源的统一访问; 提供基本服务支持构建高层数据集成服务, 如分布查询处理和数据联盟服务。

数据服务的实现使得数据资源的提供者能够按照标准的方式加工、处理和提供数据,数据资源的使用者可以通过统一的方式高效、便捷地访问到不同来源的各种数据。与此同时,天文分析、处理、可视化工具,如 2MASS-OB 科学课题中的分析处理:坐标变换(从银经银纬到空间 *XYZ* 坐标)、空间裁减、坐标统计等,也以网格服务的方式集成到科学数据网格天文的问题求解环境中,构成开展协同研究和试验的重要基础。

表 1 2MASS-OB 中涉及到的数据库和表信息

观测波段	数据量	数据库名	涉及表名	主要字段	字段数	记录数	数据访问服务
可见光	50MB	Hipparcos	J/AJ/117/354	HIP		1 300	Select 查询
			I/239/hip_main	HIP, RA, Decl	78	118 218	Select 查询
红外线	127GB	2MASS	2mass_psc	RA, Decl, Glon, Glat	60	约 4.7 亿	Select 查询

2.4 Scufl 网格工作流描述语言

网格工作流是按照预先的定义将网格服务组合起来完成一个特定的任务的操作流,它是开展新形式下的科学研究活动最有力的协作工具。问题求解环境中,网格工作流描述语言每次会记录科学工作者高层建模的结果,也就是以服务形式封装的天文数据和应用工具的组合方式。网格工作流描述语言需要支持网格服务的添加和拆卸,一旦科学工作者对试验方案进行改造或者某些步骤需要调整,都能快速重构。

Scufl(Simple conceptual unified flow language)是高抽象级别的基于 XML 的数据流描述语言,定义了不同服务间数据交互的流程[7]。Scufl 中主要有五种实体:

(1) 输入(A set of inputs)

(2) 输出(A set of outputs)

(3) 过程(A set of processors)

每个过程代表一个逻辑服务,是工作流中的一个原子步骤。从用户的角度看,一个过程完成的行为就是接收输入数据,经过内部处理后产生数据到输出端。

(4) 数据连接(A set of data links)

它连接数据源到数据目的地。数据源可以是输入数据或者是一个过程的输出端;数据目的地可以是输出数据或者一个过程的输入端。

(5) 协同约束(A set of coordination constraints)

协同约束提供给数据流程一定的依赖规则,而不需要数据连接(Data Links)的参与。例如在图 3 中,过程 addTerm 和过程 getDot 之间、过程 getDot 和过程 destroySession 之间有两个协同约束(带圆点的灰线表示),使得过程 addTerm 完全结束后,过程 getDot 才会开始处理数据,依此类推。

Scufl 吸取了 WSFL 的优点,提供了更简化、抽象级别更高的描述,满足天文工作者期望的抽象表达,重点针对数据流,适合天文应用的语义。

2.5 实例实现

以 2MASS-OB 天文课题为例,说明此问题求解环境中科学工作者如何按照事先拟订好的试验步骤,协同、反复使用大量的分布、异构数据、计算等资源,组合成有效的处理,进而开展天文研究工作。

图 4 是 2MASS-OB 课题的试验步骤。天文工作者依照此步骤,在可视化的界面通过拖、拽的方式选择并组合相关天文服务。他们像在一个虚拟的实验室开展试验,不必关心数据和工具的物理分布和网格结构,就像本机上使用所有可用的数据和应用工具一样。表 2 描述了服务的信息;图 5 是自动生成 Scufl 描述的数据连接。

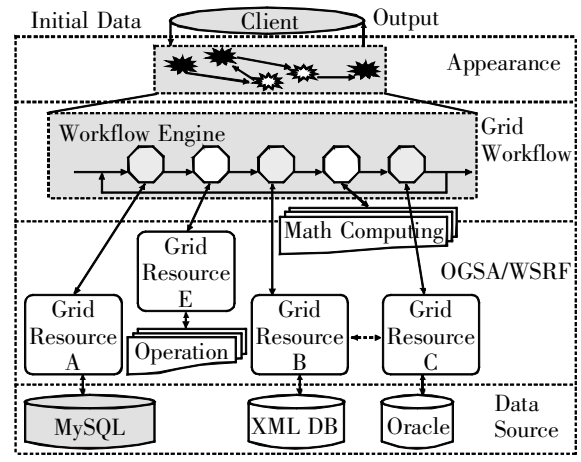

图 2 问题求解环境层次模型图

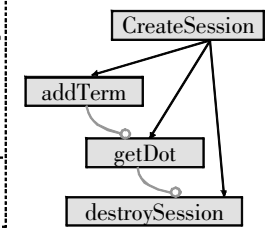

图 3 协同约束示例

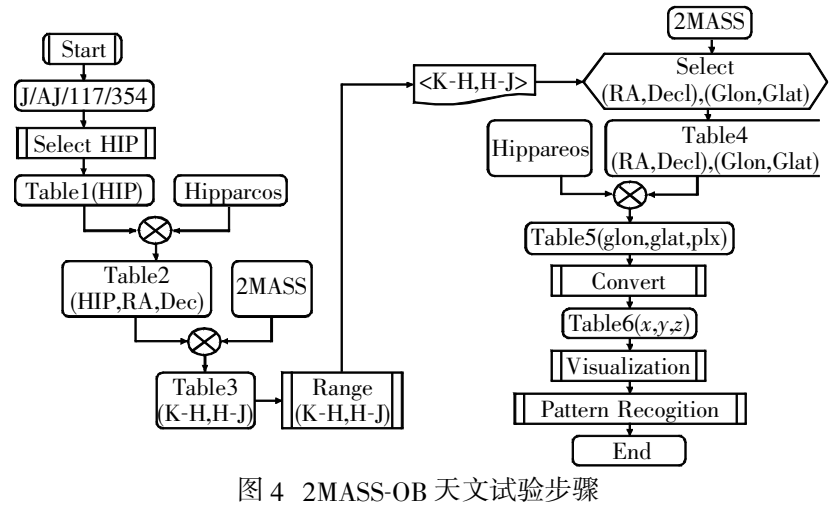

图 4 2MASS-OB 天文试验步骤

表 2 2MASS-OB 天文课题中的网格服务信息

Processor Name	Wsdl URL	Operation
Select_glat/glon/plx_2MASS	http://159.226.2.184:8080/ogsa/services/SelectDataGGP2MASSService?wsdl	Select_glat/glon/plx_2MASS
CChange	http://159.226.10.99:8080/ogsa/services/cc/CoordinateChangeService?wsdl	Change
Strip_KH/HJ	http://159.226.2.185:8088/ogsa/services/Strip_KHHJService?wsdl	Strip_KH/HJ
Computing_KH/HJ	http://159.226.10.99:8083/ogsa/services/Computing_KHHJService?wsdl	Computing_KH/HJ
Select_KH/HJ/HIP_2MASS	http://159.226.2.184:8080/ogsa/services/Select_KHHJHIP_2MASSService?wsdl	Select_KH/HJ/HIP_2MASS
Convert_KH/HJ_SelectStatment	http://159.226.2.184:8080/ogsa/services/Convert_KHHJ_SelectStatmentService?wsdl	Convert_KH/HJ_SelectStatment

2MASS-OB 天文试验步骤的 Scufl 描述如下:

```
<s: linksource = " FirstCondition" sink = " Select_KH/HI/HIP_2MASS.Condition" />
<s: linksource = " Select_KH/HI/HIP_2MASSResultSetURL" sink = " Computing_KH/HI: URL" />
<s: linksource = " Computing_KH/HJ: outputlist" sink = " Convert_KH/HJ_SelectStatment: stringlist" />
<s: linksource = " Computing_KH/HJ: outptlist" sink = " Strip_KH/HJ: stringlist" />
<s: linksource = " Strip_KH/HJ: strippedlist" sink = " Convert_KH/HJ_selectStatment: seperator" />
<s: linksource = " Convert_KH/HJ_SelectStatment: concatenated" sink = " Select_glot/glcn/plx_2MASS: Conditon" />
<s: linksource = " Select_glat/glon/plx_2MASS: ResultSetURL" sink = " CChange: URL" />
<s: linksource = " P1" sink = " t: in1" />
<s: linksource = " P2" sink = " t: in2" />
```

```
<s: linksource = " P3" sink = " t: in3" />
<s: linksource = " P4" sink = " t: in4" />
<s: linksource = " t: cchangeRetum" sink = OutDataURL" />
<s: sourcename = " FirstCondition" />
<s: sourcename = " P1" />
<s: sourcename = " P2" />
<s: sourcename = " P3" />
<s: sourcename = " P4" />
<s: sinkname = " OutDataURL" />
```

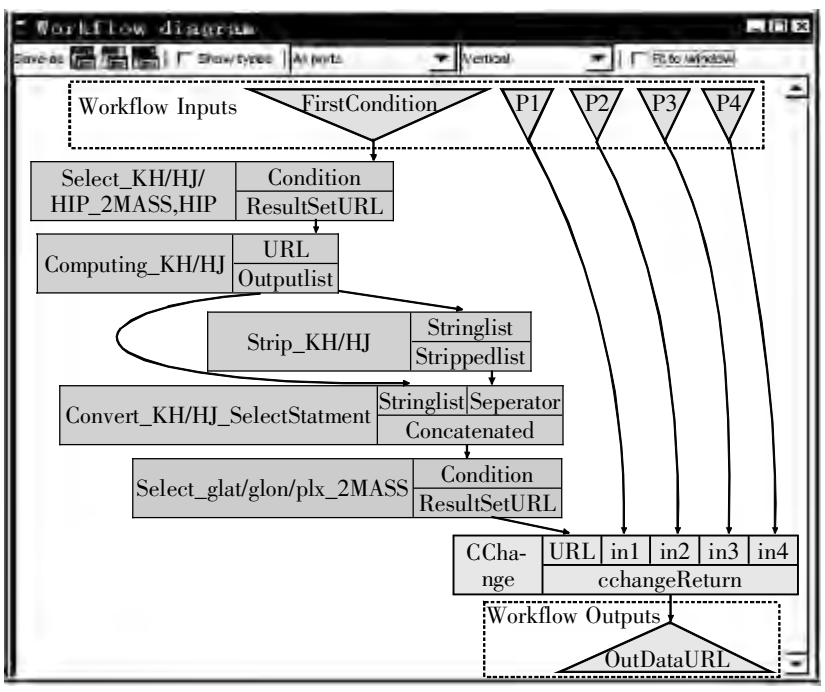

图 5　2MASS-OB 天文课题网格工作流示意图

3　结束语

本文分析了传统天文研究计算环境和当前应用项目中定制任务引擎的方法的劣势,借鉴和发展 Web Services 工作流的优势,通过对天文课题的需求分析提出了一种基于数据网格面向天文应用的问题求解环境的框架。在这个框架中,分布、异构的相关天文数据和分析工具能被发现、定位和在工作流中集成给对网格一无所知的用户,以完成复杂的科学分析和试验。

以 2MASS-OB 天文课题为例,在科学数据网格上设计实现面向天文应用的问题求解环境,用可视化流程建模方法开展面向天文的科学活动,用网格工作流平台解决服务组合问题,并研究数据的统一访问和网格工作流描述语言等关键技术。结果表明在这种环境开展工作,改善了传统计算环境协作能力低的状况,将科研活动的要素连接起来,可有效降低从事科研工作的门槛,减少科研的开发周期,使得科技工作者能够更容易地将自己的创新思想付诸实现。

参考文献:

[1]　OSTER I, KESSELMAN C. The anatomy of the grid: enabling scalable virtual organizations[J]. **International Journal of Supercomputer Applications,** 2001, **15**(3): 1-10.

[2]　迈向国家虚拟天文台: 科学目标、技术挑战和实施计划(草案)[R]. 北京: 国家天文台, 2000.

[3]　南凯, 阎保平. 科学数据库系统平台建设设想: 科学数据库与信息技术论文集(第六集)[C]. 北京: 科学出版社, 2002.

[4]　NAIM A, RATNATUNGA K, GRIFFTHS R. Galaxy morphology without classification: self-organizing maps[J]. **Astrophysical Journal Supplement Serices,** 1997, **111**: 357-367.

[5]　LEE Hsu-tai, CHEN W P, ZHANG Zhiwei, *et al*. Triggered star formation in the orion bright-rimmed clouds[J]. **The Astrophysical Journal,** 2005, **624**(2): 808-820.

[6]　李宗伟, 肖兴华. 天体物理学[M]. 北京: 高等教育出版社, 2000.

[7]　OINN T, ADDIS M, FERRIS J, *et al*. Taverna, lessons in creating a workflow environment for the life sciences[J]. **Concurrency and Computation: Practice & Experience,** 2006, **18**(10): 1067-1100.

(上接第 241 页) 波动曲线拟合方面也效果理想。

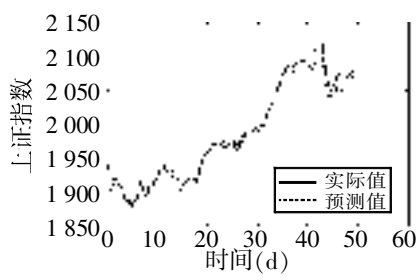

图 2　本文递归模型预测与实际曲线比较

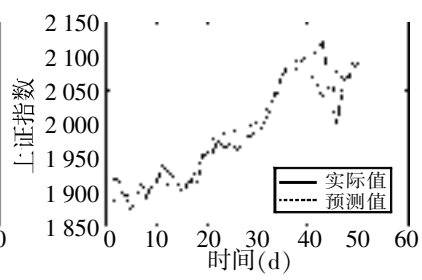

图 3　BP 网络模型预测与实际曲线比较

5　结束语

本文提出了将相空间重构与人工神经网络相结合的股票时间序列短期预测新方法,通过应用于沪市股指实例证明可有效提高预测精度。取得结论如下:

(1) 股票指数时间序列具有混沌特性,用传统的统计方法或者单纯用一种方法(如 BP 网络) 很难得到令人满意的预测结果,需要将不同的方法融合以提高预测精度。

(2) 在股票指数时间序列相空间建立混沌分析基础上的递归神经网络,利用其混沌特征依次步进来构成样本,用递归神经网络来逼近相点的非线性关系,预测结果精度与 BP 网络模型相比有了较大提高。

(3) 该方法是研究非线性时间序列预测的一种尝试,通过对沪市股票指数实例的短期预测,获得了满意的预测结果,从而证实该方法无论在理论分析还是实际应用中都具有实用性和有效性。

参考文献:

[1]　周洪涛, 王宗军. 基于混沌理论的上海股市非线性动力学研究[J]. 系统工程理论方法应用, 2005, **14**(5): 390-394.

[2]　JOHNSON M A. The random walk and beyond: an inside guide to the stock market[M]. New York: John Wiley & Sons, 1988.

[3]　BOX G E P, JENKINS G M. Time series analysis: forecasting and control[M]. San Francisco: Holden-Day, 1990.

[4]　陈可, 张琴舜. BP 神经网络在证券分析预测中应用[J]. 计算机工程, 2001, **27**(11): 95-97.

[5]　TAKENS F. Detecting strange attractors in turbulence[J]. **Lecture Notes in Math,** 1981, **898**(1): 366-381.

[6]　ROSENSTEIN M T, COLLINS J J, DE LUCA C J. Reconstruction expansion as a geometry-based framework for choosing proper delay times[J]. **Journal of Physica D,** 1994, **73**(1-2): 82-98.

[7]　FRASER A M, SWINNEY H L. Independent coordinates for strange attractors from mutual information[J]. **Physical Review A,** 1986, **33**(2): 1134-1140.

[8]　KENNEL MB, BROWN R, ABARBANEL HD I. Determing embedding dimension for phase-space reconstruction using a geometrical construction[J]. **Phyical Review A,** 1992, **45**(6): 3403-3411.

[9]　韩文秀, 郁俊莉, 王其文. 我国资本市场混沌特性研究[J]. 系统工程理论与实践, 2002, **22**(10): 43-48.

[10]　HAGAN M T, DEMUTH D B, BEALE M H. Neural Network Design[M]. Boston: PWS Publishing, 1996.