(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 102651093 A
(43)申请公布日 2012.08.29

(21)申请号 201210093084.9

(22)申请日 2012.03.31

(71)申请人 上海海洋大学
地址 201306 上海市浦东新区临港新城沪城环路 999 号

(72)发明人 黄冬梅 田瑜基 王建

(51)Int.Cl.
G06Q 10/00(2012.01)
G06F 17/30(2006.01)

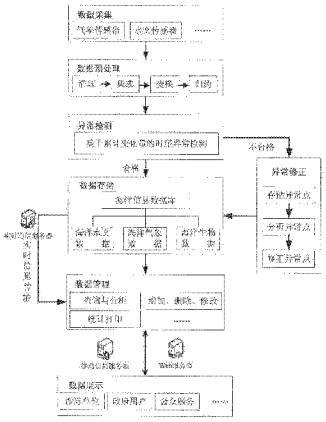
权利要求书 1 页 说明书 4 页 附图 2 页

(54)发明名称

一种基于时间序列异常检测技术的海洋信息管理系统

(57)摘要

本发明提供一种基于时间序列异常检测技术的海洋信息管理系统。该系统主要包括数据采集、数据预处理、异常检测、数据存储、数据管理和数据展示六个功能模块。海洋数据质量直接影响海洋信息管理系统的科学性。由于海洋信息数据库数据量太大，往往易受噪声、丢失数据和不一致数据的侵扰，因此，提出了一种基于累积变化量的时间序列异常检测方法，能够有效的检测出海洋数据中的异常点，再选择适当的修正方法对异常点进行修正，为海洋信息数据库提供干净、准确、简洁的数据，从而建立一套完善的海洋信息管理系统，有效的对海洋数据进行管理，为我国的数字海洋建设提供有力支持。



CN 102651093 A

1. 一种基于时间序列异常检测技术的海洋信息理系统，其特征在于，应该包括：

一数据采集模块，负责采集海洋数据；

一数据预处理模块，负责对海洋数据进行预处理；

一异常检测模块，负责检测海洋数据中的异常点；

一数据存储模块，负责存储海洋数据；

一数据管理模块，负责对海洋数据进行查询、分析、增、删、改及统计打印等操作；

一数据展示模块，负责对海洋数据的分析与查询结果进行展示。

2. 根据权利要求 1 所述的数据采集装置，其特征在于包括台站，浮标，船舶，卫星等海洋数据采集装置，具体包括各种传感器：气象传感器，水文传感器、生物传感器等海洋数据采集中用到的各类传感器。

3. 根据权利要求 1 所述的数据预处理方法，其特征在于包括数据清理、数据集成、数据交换和数据归约四个步骤。

4. 根据权利要求 1 所述的异常检测模块，其特征在于使用基于累计变化量的时间序列异常点检测方法，对海洋数据进行异常检测，对于合格的数据进行数据存储，对于不合格的数据进行异常点存储、分析及修正。

5. 根据权利要求 1 所述的数据存储装置，其特征在于使用传统的数据存储。针对经过异常检测后的合格数据及异常修正后的合法数据，统一采用传统的数据库（例如 oracle）行存储，再对存储的数据进行数据管理。

6. 根据权利要求 1 所述的数据管理模块，其特征在于包括对传统数据的统计分析和查询。根据涉海部门、政府部门、以及公众用户提出的要求进行分析。具体分析包括某一海洋要素的某一历史时间变化趋势预测和分析，某一海洋要素正常值分析。并对海洋数据进行增加、删除、修改、数据导入和统计打印等管理。

7. 根据权利要求 1 所述的数据展示模块，其特征在于根据用户需求，对权利要求 6 的数据分析及查询等结果进行展示。通过 WEB 服务器，移动信息服务器，无线网络将结果在终端进行展示。终端展示模块包括计算机、智能手机、PDA 等移动终端智能设备。

一种基于时间序列异常检测技术的海洋信息管理系统

技术领域

[0001] 本发明涉及海洋信息数据的预处理方法，时间序列异常检测技术和海洋数据管理方法。

背景技术

[0002] 目前数据挖掘的研究主要集中在数据挖掘算法的探讨，而忽视了对数据预处理的研究。而实际系统中的数据一般很少能直接满足数据挖掘算法的要求，严重影响了数据挖掘算法的执行效率，甚至会造成挖掘结果的偏差。据统计，数据预处理所花费的时间和成本占数据挖掘全过程的60％左右。因此，对数据源进行有效的归纳和预处理，已经成为数据挖掘系统实现过程中的关键问题。

[0003] 随着国家用海需求的日益增长及海洋经济的快速发展，对海洋局的管理和服务能力提出了更高的要求。为了满足海洋局对海洋数据管理和海洋数据分析统计的需求，建设海洋信息管理系统势在必行。然而，高质量的决策必然依赖于高质量的数据，如何提高海洋数据的质量控制效率和水平，更高效的利用海洋数据，使之符合挖掘算法的规范和要求，是国家973项目的一个重要研究内容。海洋数据质量的好坏直接影响海洋信息管理系统决策的科学性，目前国内还没有系统的海洋数据质量控制方法，一般采用手工校正处理进行控制，针对大量的海洋数据，使用基于累计变化量的时间序列异常检测技术，对采集的海洋数据进行异常检测，将合格的数据及异常修正后的合法数据，存储到海洋信息数据库中，对海洋数据进行管理及应用展示。

发明内容

[0004] 本发明提供一种基于时间序列异常检测技术的海洋信息管理系统。该系统主要包括数据采集、数据预处理、异常检测、数据存储、数据管理和数据展示六个功能模块。其中，数据预处理是为海洋信息管理系统提供高质量数据的关键。海洋数据质量直接影响海洋信息管理系统的科学性。由于海洋信息数据库数据量太大，往往易受噪声、丢失数据和不一致数据的侵扰，因此，提出了一种基于累积变化量的时间序列异常检测方法，能够有效的检测出海洋数据中的异常点，再选择适当的修正方法对异常点进行修正，为海洋信息数据库提供干净、准确、简洁的数据，从而建立一套完善的海洋信息管理系统，有效的对海洋数据进行管理，为我国的数字海洋建设提供有力支持。

附图说明

[0005] 图1为本发明的海洋信息管理系统架构图。

[0006] 图2为本发明的基于累计变化量的时间序列异常点检测方法的流程图。

具体实施方式

[0007] 本发明公开了一种基于时间序列异常检测技术的海洋信息管理系统，下面结合附

图对实施方式进行说明。

[0008] 请参考图 1。图 1 为本发明的海洋信息管理系统架构图。包含数据采集，数据预处理，异常检测，数据存储，数据管理，数据展示六个功能模块。

[0009] 数据采集模块包括气象传感器、水文传感器和生物传感器。气象传感器采集气象类数据，包括风速风向，气温，降水量和雾等数据；水文传感器采集数据包括水温、盐度、海流、波浪、潮位、含沙量和悬沙等；生物传感器采集浮游动物、浮游植物和底栖生物等数据。

[0010] 通过分析，发现海洋信息管理系统中的元数据存在以下问题：

[0011] 1. 海洋信息管理系统的数据库在某些字段上存在空值。所以需要对这些数据进行一些转换和集成工作，对空值字段需进行数据的智能填充。

[0012] 2. 各个站点关于台站信息的数据在结构上基本相同，但在数据的完整性和一致性上很差。

[0013] 3. 来自不同数据表的同类数据，具有不同的数据类型。如同样是表示日期数据，有的用日期型，有的用字符型。

[0014] 4. 各台站的海洋数据中或多或少的含有噪声数据，在装入数据仓库前必须进行清洗。

[0015] 综上所述，海洋信息管理系统中的原始数据存在数据不一致性、数据空缺、数据冗余等情况。可见，海洋数据并不能直接用于后继的数据开采，对海洋数据的预处理是进行数据挖掘的前提。

[0016] 数据预处理模块主要是通过对数据进行清理、集成、变换和归约等四个方面的工作来实现。数据清理例程通过填写缺失的值、光滑噪声数据、识别或删除离群点并解决不一致性来“清理”数据。主要是达到如下目标：格式标准化，异常数据清除，错误纠正，重复数据的清除。数据集成例程将多个数据源中的数据结合起来并统一存储，建立数据仓库的过程实际上就是数据集成。通过平滑聚集，数据概化，规范化等方式将数据转换成适用于数据挖掘的形式。数据挖掘时往往数据量非常大，在少量数据上进行挖掘分析需要很长的时间，数据归约技术可以用来得到数据集的归约表示，它小得多，但仍然接近于保持原数据的完整性，并结果与归约前结果相同或几乎相同。

[0017] 异常检测模块主要是使用基于累计变化量的时间序列异常点检测方法，对采集的海洋数据进行异常检测，对于合格的数据进行数据存储，对于不合格的数据，对其进行异常点存储，并进行异常分析，再选择适当的修正方法对异常点进行修正。

[0018] 数据存储模块主要是将经过时间序列异常检测后的合格数据及异常修正后的合法数据，存储到海洋信息数据库中。

[0019] 数据管理模块包括海洋气象、海洋水文和海洋生物等数据进行查询，数据分析。对于查询功能，通过精确查询和模糊查询两种查询方式，实现对海洋数据进行全方位多条件的查询。数据分析功能是通过台站比较和多年比较，对某一海洋要素的某一历史时间变化趋势和某一海洋要素正常值进行分析，将海洋数据的规律总结出来，并给予用户提示信息，为决策者提供帮助。另外，还可以对数据进行增加，删除，修改，数据导入，统计打印等功能。其中，数据导入功能可以对数据进行批量增加，可以将整个 Excel 表中的数据导入到数据库中，使得批量数据的导入工作更加快捷方便，提高工作效率。

[0020] 数据展示模块将数据分析模块的结果通过图表多种形式进行展示，展示的客户端

包括涉海单位、政府用户、公众等。

[0021] 数据展示模块和数据管理模块之间采用 GIS 服务器，Web 服务器，移动信息服务器等实现实时通信和展示。

[0022] 请参考图 2。图 2 为本发明的基于累计变化量的时间序列异常点检测方法的流程图。

[0023] 在数据挖掘过程中，常常存在与数据模型或数据一般规律不符合的数据对象，这类与其它数据不一致的数据对象就称为异常数据，它们往往容易被人们所忽略。然而，这些数据对象可能是具有特殊意义的，而且相对于那些普通的数据而言，这类异常的数据往往提供了更多的有用信息，它们往往更具有研究价值。

[0024] 按照异常的表现形式不同，时间序列的异常可以分为序列异常，点异常和模式异常。本发明主要是针对海洋时间序列数据的特点，设计了基于累计变化量的时间序列异常点检测方法，用于检测海洋时间序列中的异常点。

[0025] 定义 1 :海洋时间序列异常点定义

[0026] 给定一段海洋时间序列 X = <x1 = (v1,t1),x2 = (v2,t2), …,xn = (vn,tn)>，点 xi = <vi, ti> 表示时间序列在 ti 时刻的观测值为 vi。用 <N1, N2, …, Nk) 表示点 xi 的 k 个邻居点集合，其观测值集合记为 <vN1,vN2, …,vNk>,给定阈值 T,若点 xi 与其 k 个邻居点的累积变化量 (Accurnulative Change) 大于 T,则判定点 xi 为这段时间序列中的一个异常点，这一判定条件用公式表示为 :

[0027] $$\text{Accurnulative Change} = \frac{w_1 \cdot |v_i - v_{N_1}| + w_2 \cdot |v_i - v_{N_2}| + \ldots + w_k \cdot |v_i - v_{N_k}|}{w_1 + w_2 + \ldots + w_k} > T$$

[0028] 式中的 <w1,w2, …,wk> 为权值向量，赋予每个变化量不同的权重。一般来说，在时间轴上，越接近点 xi 的邻居点赋予的权值越大 ;阈值 T 是用户给定的一个常量，点 xi 的累积变化量和阈值的大小关系，是判定 xi 是否为一个异常点的依据。

[0029] 本发明涉及一个平均变化量的统计量，该变量是各个相邻观测值之间的差值和的平均值。在定义 1 的基础上，本发明提出了一种基于累积变化量的海洋时间序列异常点检测方法。主要步骤如图 2 所示。基于累计变化量的时间序列异常点检测方法的步骤 :首先读取数据，并计算数据的平均变化量。然后遍历每一个数据点，查找到其邻居点，计算累积变化量的值，根据平均变化量计算出阈值 T,比较累积变化量和 T 的大小关系，判定异常点并存储。

[0030] 异常数据产生的原因很多，可能是由于在数据阅读、记录、计算、误操作时产生的错误等人为因素，还可能是由于数据内在特性而造成。根据定义 1,一个海洋时间序列中的点 xi 被判定为一个异常点，则点 xi 与其邻居点的累积变化量的值一般较大，导致这一结果的原因也有很多种，结合海洋时间序列数据的特点可能的原因归为以下三类 :

[0031] 1. 数据录入时的错误导致。

[0032] 2. 自然因素导致。

[0033] 3. 其它人为因素导致。

[0034] 经过分析，异常点的修正方法主要有以下四种 :

[0035] 1. 根据其它数据来源，手工修正，或由领域专家估计修正，但过程复杂、耗时长、代价高。

[0036] 2. 用该序列其它时间数据平均值补修正。但是，对于连续的异常点，有时该方法也不能达到满意的效果。

[0037] 3. 用其它相关序列的数据平均值补缺失。

[0038] 4. 可以通过回归分析、贝叶斯形式化方法工具或判定树推导出可能数据值以修正异常值。

[0039] 综上所述，本发明通过对海洋信息管理系统中的元数据进行详细的分析，发现海洋信息数据库中大量的海洋数据存在数据不一致、数据空缺和数据冗余等问题。为了更好地对海洋数据进行有效的归纳和预处理，提出了一种基于累积变化量的时间序列异常点检测方法。这种方法能够有效的检测出海洋数据中的异常点，然后对异常点进行分析，再选择适当的修正方法对异常点进行修正，保证了海洋数据的质量，再进一步将异常检测后的合格数据及异常修正后的合法数据存储到海洋信息数据库中，对数据进行管理及展示，建立了一套完善的海洋信息管理系统。该系统可以指导海洋相关部门业务流程的科学化和规范化，为海洋相关部门管理决策提供科学的支持。

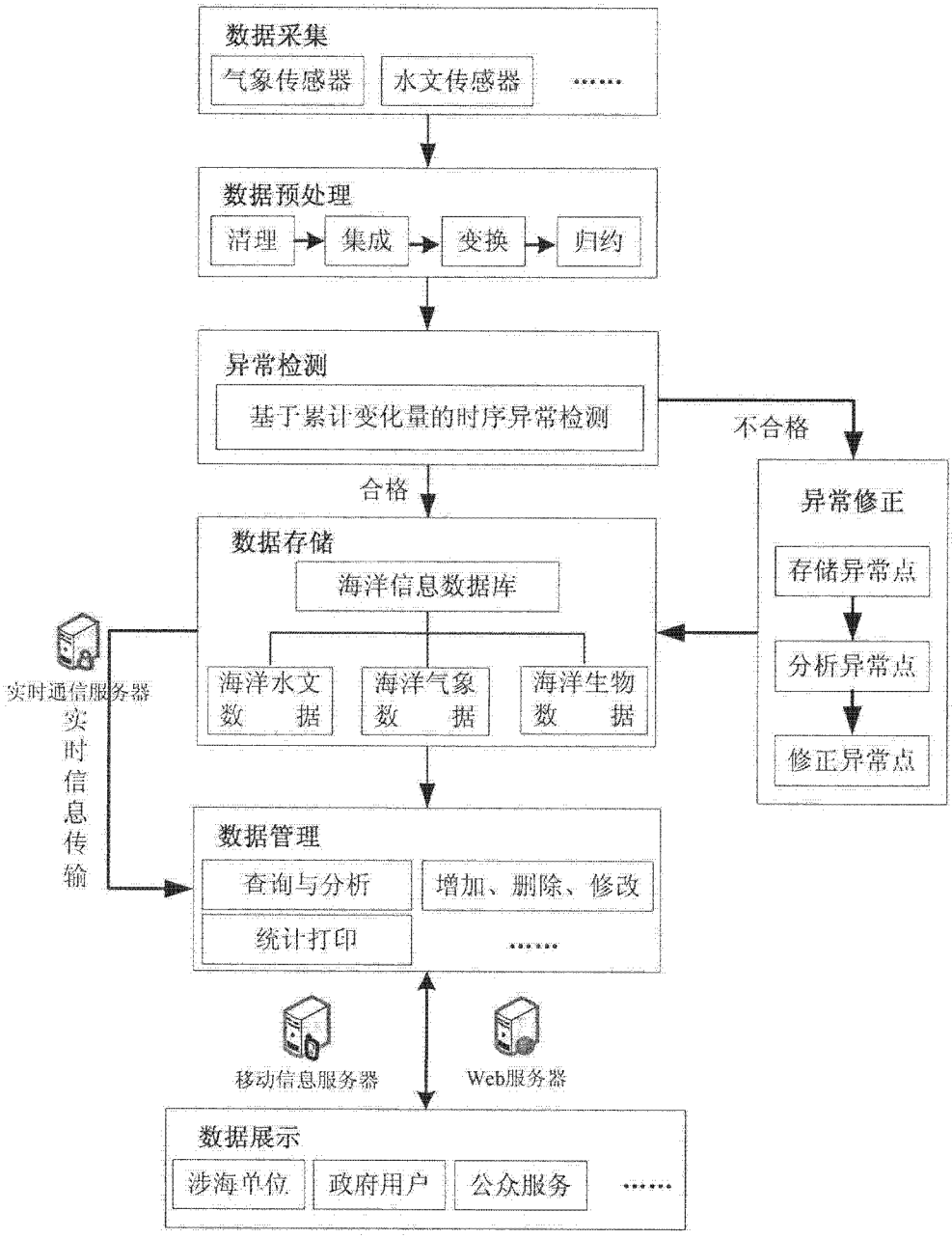

数据采集
气象传感器
水文传感器
……
数据预处理
清理
集成
变换
归约
异常检测
基于累计变化量的时序异常检测
不合格
合格
异常修正
存储异常点
分析异常点
修正异常点
数据存储
海洋信息数据库
海洋水文数据
海洋气象数据
海洋生物数据
实时通信服务器
实时信息传输
数据管理
查询与分析
增加、删除、修改
统计打印
……
移动信息服务器
Web服务器
数据展示
涉海单位
政府用户
公众服务
……

图 1

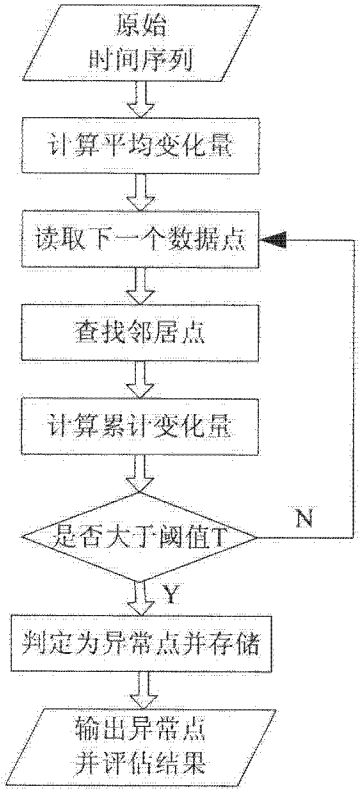

原始
时间序列
计算平均变化量
读取下一个数据点
查找邻居点
计算累计变化量
是否大于阈值T
N
Y
判定为异常点并存储
输出异常点
并评估结果

图 2