无线监测系统的数据处理方法研究

连 乐 付 杰
(兰州交通大学电子与信息工程学院 兰州 730070)

摘 要 传统数据库、联机事务处理(On-Line Transaction Processing,OLTP)方式已经无法满足用户对数据查询分析的需求,因此提出一种将 ROLAP(Relational Online Analytical Processing)技术与改进的数据挖掘算法相融合的新型数据处理方式。即使用 ROLAP 引擎将星型关系数据库结合成多维数据结构,利用改进后的 K-means 算法进一步分类和聚集数据库中未缓存的数据,并结合线性回归算法统计数据变化率,实现监测系统的预警功能。仿真结果表明,新型数据处理系统不仅能够挖掘更多类的数据信息,而且其预警时间相比传统预警方式有了显著提高。

关键词 ROLAP,数据挖掘算法,预警功能
中图法分类号 TP274 **文献标识码** A

Research on Data Processing Method of Wireless Monitoring System

LIAN Le FU Jie
(School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China)

Abstract The traditional database and online transaction processing (OLTP) have been unable to meet users' demands for data query and analysis. This paper put forward a new kind of data processing mode combining ROLAP technology and an improved algorithm of data mining approach. It uses the ROLAP engine to combine star database into a multidimensional data structure, and uses the improved K-means algorithm to classify and aggregate uncache data in the database. Combining linear regression algorithm, the data rate of change is counted to achieve earlier warning function of monitoring system. Simulation results show that the new data processing system can mine more data and information, and the early warning time has been significantly improved compared with the traditional alarm mode.

Keywords ROLAP, Data mining algorithm, Warning function

1 引言

无线传感网络(Wireless Sensor Network,WSN)通常由硬件和软件两部分组成,硬件模块中的传感器节点通过无线传输将数据传送到汇聚节点,然后由通信模块将数据传输到客户端,并通过数据库软件接收。目前,处理基本事务的传统数据库存在的不足包括:关系型数据库侧重于捕获数据与记录[1],只能进行简单的显示或计算;查询并处理大量数据时,数据库要提供时间段内所有的有效数据,计算过程缓慢;数据库存储数据时没有特定的要求,如类型或者时间。

有限的数据信息对于大量数据的研究来讲是不足的,同时也会影响系统的决策功能。关系数据库可直接运用于普通 OLTP(On-Line Transaction Processing)数据库,并利用关系数据库模拟 OLAP 中多维数据集的管理[2],使用过程简便;但数据挖掘则是对大量数据进行有目的的提取、分拣、归类等,以挖掘数据中隐含的有用信息,借助多种数据分析工具来发现元数据信息和模型之间的关系。二者均是数据处理中不可或缺的工具。

近年内的相关研究包括:陈阵[3]提出并探讨了 OLAP 技术的特点,其满足了用户报表多维分析计算的需求;邱怀姗[4]讨论了 OLAP 在大量数据下的设计实现和数据挖掘算法对数据处理的效果,并对两种方法做了横向比较;刘建宇详细介绍了近两年 OLAP 技术在具体数据处理案例中的应用成果[5];张阳结合 K-means 算法的不足[6],提出一种减少聚类迭代数与聚类数据量的新型计算方法。以上研究表明,OLAP 技术与数据挖掘一直是数据处理中必不可少的代表性工具,近年来得到了不断的更新和改进。综合以上研究的特点,本文提出了将 OLAP 技术与数据挖掘算法相融合的新型数据处理方式,并针对传统 K-mean 算法进行改进。

1 系统模块

1.1 硬件模块框架

在水质监测平台,硬件部分用于采集实际环境中与水质相关的几项参数,下面用温度、湿度这两项参数进行分析研究[7]。

硬件结构的设计包括:传感器模块,用于采集源数据;ZigBee 模块,用于传感器节点、路由节点以及汇聚节点间的数据传输(无线收发模块(GPRS 模块),用于硬件模块与客户端之间的数据传输;GPRS 模块与 ZigBee 模块之间采用 RS232 串口通信)[8];电源模块,用于提供不同电压值,为硬件系统各模块供电,如图 1 所示。

本文受国家自然科学基金(61366006),甘肃省自然科学基金(1610RJZA046),甘肃省建设科技攻关项目(JK2016-7),甘肃省高等学校科研项目(216132)资助。

连 乐 硕士生,主要研究方向为电路与系统,E-mail:1229739617@qq.com。

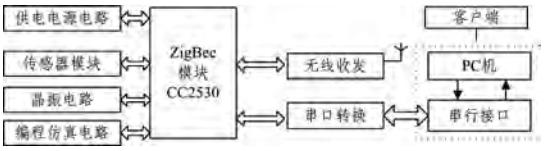

图 1　硬件结构图

1.2　软件算法模块

1.2.1　改进的 K-means 算法

K-means 算法的主要特点是通过两组数据间的相异度对其分类[9]，即给定一个元素集合 D，其中每个元素均具有可观察属性，也即各维度属性。运用简单算法将 D 划分为 K 个子集(簇)，要求子集中的元素间的相异度尽可能低，而不同子集的元素的相异度则较高。

假设获得两组元素项：$X=\{x_1,x_2,\cdots,x_n\}$，$Y=\{y_1,y_2,\cdots,y_n\}$，它们各自具有 n 个特征属性，则 X 和 Y 的相异度定义为：

$$d(X,Y)=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\cdots+(x_n-y_n)^2}$$

这种相异度判定仍然存在缺陷，因此增加了规格化计算，通过公式将各属性映射到[0,1]区间，其映射公式为：

$$a_i'=\frac{a_i-\min(a_i)}{\max(a_i)-\min(a_i)}$$

其中，$\max(a_i)$和 $\min(a_i)$是所有元素项中第 i 属性的最大值和最小值，这样便将元素规格化到区间[0,1]上。

利用簇中心对最终聚类结果的影响，得到改进后的 K-means 算法的步骤如下：

步骤 1　确定 K 值，一般在$[2,\sqrt{n}]$区间内逐个选取(n 为数据个数)；

步骤 2　选出 K 个初始聚类中心；

步骤 3　完成初步聚类；

步骤 4　采用 Sil(Silhouette)指标进行评选；

步骤 5　取 $K=K+1$，继续进行 K-means 算法聚类；

步骤 6　回到步骤 4，再进行评选，最终得到最优 K 值以及其对应的聚类中心。

算法流程如图 2 所示。

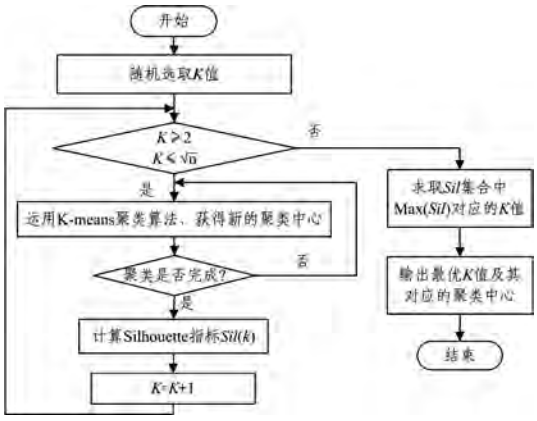

图 2　改进的 K-means 算法的流程图

1.2.2　回归算法

回归算法用一个函数拟合给出的点集 D，使得点集与其误差最小，如果得到的函数曲线是一条直线，则称为线性回归。

回归算法的实现步骤如下：

步骤 1　回归算法的关键就是找出使误差最小的 W，此处的误差也就是现有值 Y 和新读取值 y 之间的差值。其平方误差为：$\sum_{i=1}^{m}(y_1-x_j^{\mathrm{T}}w)^2$。

步骤 2　直接运用公式

$\theta=(x^{\mathrm{T}}x)^{-1}x^{\mathrm{T}}\vec{y}$求解平方误差公式，当然也可以使用最小二乘法或者梯度下降法等。

步骤 3　得到拟合曲线。

2　ROLAP 与算法融合

针对传统数据库的不足，我们建立了框架将 ROLAP 技术与数据挖掘算法相融合：通过 ROLAP 技术建立星型多维数据模型；用改进的聚类算法对部分未缓存的数据分类，提高数据的精确度；使用回归算法将重新采集到的数据与缓存的数据进行拟合，通过结果曲线来判断系统的安全性。

2.1　ROLAP 的实现

2.1.1　物理数据的存储

ROLAP 系统基于 Mondrian 工具实现。ROLAP 结构中最关键的部分是“OLAP 引擎”[10]。客户端是适应多维显示的工具，它可直接从“OLAP 引擎”读取数据，再以多维的方式将其显示。图 3 为水质监测平台数据的星型模型图。

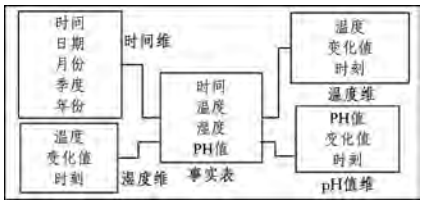

图 3　星型模型图

图 3 将水质研究对象作为一个事实表，存储了温度、湿度、PH 值和时间 4 个维表的主码，由 4 个维表与事实表构成整体星型模式。在每个维表中又设置相应的存储项，如温度维包括温度值、变化值和具体的时刻值，其中变化值表示 T 时刻读数与 $T+1$ 时刻读数的差值，用于后期优化系统的安全性。

2.1.2　多维处理

图 4 为 ROLAP 系统的操作流程。在建立了多维模型后，还需得到 result 单元格，以建立整体框架的元数据体系。

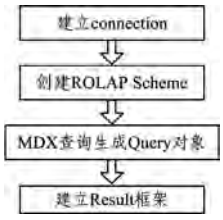

图 4　ROLAP 系统的操作流程图

得到 Result 单元格后，继续将其切分为片段，即部分维度的查询结果单元格。

2.1.3　寻找聚集计算目标

在 ROLAP 系统中查找 Result 框架并将其分解成片段，聚集算法则针对片段所形成的簇，处理未缓存的数据，这一过程包括将未缓存“条纹”分解为片段。Mondrian 工具中将“条纹”分解为“片段”的调用栈代码为：

load()-mondrian. rolap. agg. Segment

load()-mondrian. rolap. agg. Aggregation

loadAggregation ( )-mondrian. rolap. agg. Aggregation-Mana-ger

loadAggregation ( )-mondrian. rolap. FastBatchingCell-Rea-der. Batch

loadAggregation( )-mondrian. rolap. FastBatchingCellrea-der

executeBody(Query)-mondrian. rolap. Result

2.2 K-means 算法的实现

以监测系统中的温度数据为例，获得簇数据集以及聚集计算目标数据后，采用 Weka 系统设置模型参数，实现算法建模、仿真[11]。数据聚类分析结果如图 5 所示。

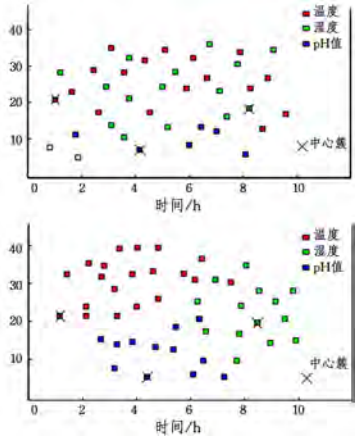

图 5 聚类分析结果

2.3 线性回归算法的实现

实现了聚类算法后，未缓存的数据已经被归类到相异度最低的缓存单元格，通过数据处理后其效果得到进一步改善。之所以加入回归算法是为了应用回归算法本身的可预测性。

传统的报警系统为数据设置阈值，当数据超出阈值范围时启动报警装置。但这种装置反应迟钝，没有足够的缓冲时间。此前，我们设定在各维度中加入了变化量这一属性，取各温、湿度值 t_n 与 t_{n+1} 间的差值。在回归算法中，对于变化量属性的数值，可进一步求取变化量的变化值，也就是读数的变化率。监测回归曲线时，当变化率超出设定范围时，说明数据异常，则激活报警系统，使问题于萌芽状态解决，提高了系统的安全性。

同样地，以采集的温度数据为例，用 Weka 系统来实现回归算法，其回归分析结果如图 6 所示。

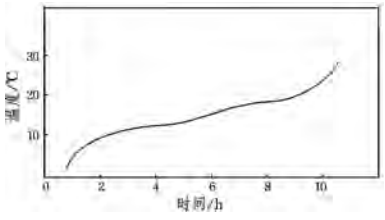

图 6 线性回归曲线

最后，记录并统计报警反应的时间，结果如表 1 所列。

表 1 二者的反应时间对比

报警方式	参数反应	时间/ms
传统报警装置	温度	36
	湿度	42
	pH 值	56
回归算法	温度	17
	湿度	16
	pH 值	16

从表中可知，传统的预警系统与基于算法分析统计的预警系统的反应时间相差较大，基于算法分析的预警方式可以更好地保护系统实施的安全性。

结束语 文中提出了一种将 ROLAP 技术与改进型数据挖掘算法相结合的新型数据处理方法，它不仅能在多维的基础上展示查询，且改进型 K-means 算法能更高效地将未缓存数据聚集分类。在融入了回归算法后，得到更敏感的预警系统。根据所有仿真结果以及预警时间的对比可知，新型数据处理方法成功融合了两种数据仓库工具，提高了数据存储、分类的维度和精准度，信息查询的准确度以及系统的反应速度为研究更多数据仓库工具的融合提供了技术依据，以挖掘更多、更深层次的数据信息，并适应数据化时代的发展。

参 考 文 献

[1] 沈敬伟，周廷刚，温永宁，等. 基于面向对象数据库的空间数据管理[J]. 西南大学学报(自然科学版)，2013，35(4)：132-137.

[2] 游进国，董朋志，胡宝丽，等. 语义 OLAP 缓存技术研究[J]. 小型微型计算机系统，2015，36(7)：1470-1475.

[3] 陈阵. 基于关系型数据库 OLAP 策略的研究与实现[D]. 大连：大连理工大学，2002.

[4] 邱怀姗. OLAP 和数据挖掘技术在高校科技管理决策中的应用[D]. 北京：北京化工大学，2003.

[5] 刘建宇. 基于关系数据库的 OLAP 的研究[J]. 煤炭技术，2013，32(4)：152-154.

[6] 张阳，何丽，朱颢东. 一种改进的 K-means 动态聚类算法[J]. 重庆师范大学学报(自然科学版)，2016(1)：97-101.

[7] 花俊，胡庆松，李俊，等. 海洋牧场远程水质监测系统设计和实验[J]. 上海海洋大学学报，2014，23(4)：588-593.

[8] 王建平，徐其林，张茂林. 基于 EPA 标准的 ZigBee 网络构建方法的研究[J]. 计算机测量与控制，2008，16(1)：121-123.

[9] 纪雯，王建辉，顾树生，等. 基于系统聚类的自动知识获取方法简[J]. 控制工程，2016，23(10)：1527-1532.

[10] 殷婷，肖敏，陈岭，等. 基于 CQPM 的 OLAP 查询日志挖掘及推荐[J]. 浙江大学学报工学版，2012，46(11)：2052-2060.

[11] 王会青，陈俊杰，侯晓晶，等. 决策树分类的属性选择方法的研究[J]. 太原理工大学学报，2011，42(4)：346-348.

(上接第 561 页)

[3] SHEU J B. Dynamic Relief-Demand Management for Emergency Logistics Operations under Large-Scale Disasters[J]. Transaction Research Part E，2010，46(1)：1-17.

[4] 戴更新，达庆利. 多资源组合应急调度问题的研究[J]. 系统工程理论与实践，2000，20(9)：52-55.

[5] 刘春林，何建敏，施建军. 一类应急物资调度的优化模型研究[J]. 中国管理科学，2001，9(3)：29-36.

[6] BARBAROSOLU G，ARDA Y. A two-stage stochastic programming framework for transportation planning in disaster response[J]. Journal of the Operational Research Society，2004，55(1)：43-53.

[7] 何建敏，刘春林，曹杰，等. 应急管理与应急系统——选址，调度与算法[M]. 北京：科学出版社，2005：101-109.

[8] 罗朝晖，董鹏，黎放. 有运力限制条件下的军械紧急调运优化模型研究[J]. 海军工程大学学报，2006，18(3)：1-6.