e-Science应用中野外数据采集传输系统设计与实现

秦 刚 王金一 杨智超 代梁峰 樊道毅 阎保平
(中国科学院计算机网络信息中心 北京 100190)

摘 要 e-Science是近年新兴的研究热点,主要研究IT技术与各学科领域的结合与应用。野外数据的采集与传输是在涉及野外观测的学科领域中开展e-Science应用所面临的亟需解决的问题。提出了一种融合多种通讯方式的支持野外数据采集与传输的系统架构,以实现自动化、高效的数据采集与传输。提出的系统已经使用JAVA语言开发出了原型系统,经过测试有较好的实用效果。

关键词 e-Science应用,野外数据采集,无线网络,数据通信,数据处理,JAVA语言
中图法分类号 TP393 **文献标识码** A

Design and Implementation of Field Data Collection and Transmission System for e-Science Applications

QIN Gang WANG Jin-yi YANG Zhi-chao DAI Liang-feng FAN Dao-yi YAN Bao-ping
(Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China)

Abstract e-Science is one of the research hotspots in recent years, which focuses on how to apply IT on other disciplines. How to collect and transmit field data is an immediate problem to be solved in e-Science for those disciplines involving field observation. This paper brought out a system for field data collection and transmission in an automatic and high efficient way combined with several kinds of communication methods. A demonstration system was developed in JAVA language to implement this system and was verified to be good efficient in test.

Keywords e-Science application, Field data collection, Wireless network, Data communication, Data processing, JAVA language

1 引言

随着人类社会的不断发展,人类对客观世界的探索和认识也不断地向新的深度和广度拓展,科学研究的问题也日益趋于复杂化。传统的科学研究方法和手段已经暴露出其明显的不足:一方面,小范围、封闭式的科学研究活动使得信息不能快速流动和充分地共享,造成重复劳动、科研效率低下的问题,延长了重大科研成果产出的时间。另一方面,传统的观测试验、理论分析的科研手段在面对很多复杂问题时已经显得无能为力,迫切需要对大规模数据的处理分析、计算模拟和仿真等新的科研方法的支持[1]。随着IT技术(包含计算机、通信、信息技术等)的发展,用先进的信息化基础设施构建新型的信息化科学研究环境已经成为可能。

e-Science就是在这种背景下产生的。简而言之,e-Science就是科学研究信息化。英国科学家约翰.泰勒指出:"e-Science是在重要的科学领域中的全球性合作以及使这种合作成为可能的下一代基础设施。e-Science意味着科学研究越来越多地通过因特网进行分布在全球的合作,并充分利用极大规模的数据、万亿次规模的计算资源和高性能的可视化设施[2]"。美国科学基金会的报告中也明确指出"e-Science是一种新的科学研究环境,在这种新的研究环境中,研究人员能通过高性能的网络进行先进计算、协同工作、实现数据获取和管理的服务"[3]。

在地球物理、生态环境、气象水文等涉及到野外观测数据采集的e-Science应用领域中,如何利用IT技术手段为这些学科提供新的数据获取方法,对于提高这些学科的成果产出效率、拓展研究范围有重要的意义。

2 需求分析

在以往传统的科研方式中,科研人员每年定期去野外采集诸如温度、湿度、土壤温度、光照强度等观测数据。随着自动化电子仪器的出现,科研人员也在野外部署一些实验仪器,这些实验仪器将观测到的数据保存到自带的存储卡中,由科研人员定期将数据拷贝回来。有的设备带有GPRS功能,能够自动将采集到的数据通过移动通信网络发送到指定的计算机上。这些自动化仪器在一定程度上减轻了科研人员的工作负担,提高了工作效率。

但随着科研活动的深入,研究范围的扩展,需要在一些基础设施条件薄弱的地方部署观测设备,获取观测数据。例如,中国科学院在西藏、青海等偏远的地方建立了野外台站,野外

本文受中国科学院知识创新工程青年人才领域前沿项目基金(CNIC_QN_10003)资助。
秦 刚(1978-),男,博士,副研究员,主要研究方向为计算机网络、e-Science应用,E-mail:gqin@cnic.cn;王金一(1977-),男,博士生,副研究员,主要研究方向为计算机网络;杨智超(1987-),男,硕士生;代梁峰(1987-),男,硕士生;樊道毅(1987-),男,硕士生;阎保平(1950-),女,博士,研究员,主要研究方向为数据库技术、e-Science应用。

台站有供工作人员居住和设备存放的管理中心，这些管理中心一般具有电话线路、移动通信信号等基础设施。而科研仪器设备则部署在野外台站周围几公里甚至几十公里的山区或者林区，那些地方基本没有通讯信号，目前还是依靠人工定期拷贝数据的方式来获取数据。其缺点非常明显：

1）工作效率低下，人员成本很高；

2）缺乏对科研仪器设备运行状态的监控，如果设备在运行期间出现故障或者被盗，等到下一次科研人员去拷贝数据时才能发现，这将造成一些关键时间节点的数据丢失；

3）科研仪器设备数据采集参数一旦设定则难以更改，如需要修改只能由人工去现场进行修改；

对于此类有野外科研数据采集需求的 e-Science 应用，如何提供一种有效的、自动化程度高的野外数据采集方式，是本文要研究的主要问题。

3 系统设计

3.1 系统概述

本文设计的 e-Science 野外数据采集处理系统整体结构如图 1 所示。

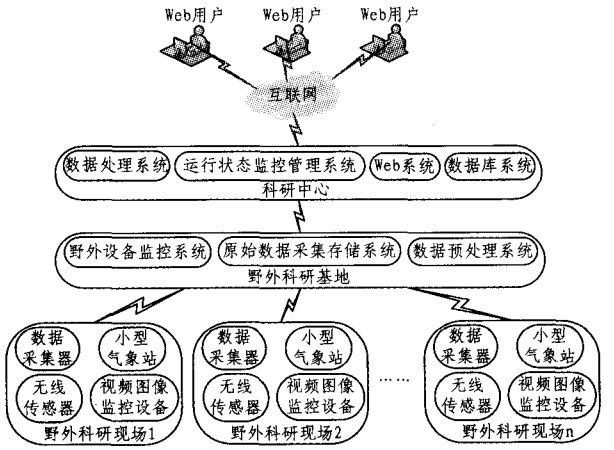

图 1 系统结构图

根据地理位置和功能的不同，可以将系统从总体上分为 3 个部分，下面分别进行介绍。

3.1.1 野外科研现场

野外科研现场，包括部署在野外的科研仪器设备，如各种传感器、数据采集器、监控设备等。它们分布在野外，负责原始观测数据的采集。

在一个野外科研应用中，可能包括多个野外科研现场。

3.1.2 野外科研基地

野外科研基地，一般位于具备一定基础设施条件的地方，比如靠近村镇或者道路，有电力供应、通信线路等。在本文设计的系统中主要完成如下几种功能。

1）野外设备监控：负责对野外科研现场终端设备运行状态的监控，有异常情况会有系统告警，由值班人员及时进行设备维护检修；

2）原始数据采集存储：负责接收从各个野外科研现场采集到的数据。数据采集可以采取两种方式：主动方式，即采取轮询的方式，由系统主动向野外终端设备发送数据请求；被动方法，即系统等待野外终端设备发送数据。可以根据终端设备和实际科研需求来选择不同的数据采集方式。同时，为了保证数据的可靠性，在野外科研基地保留有原始数据的存储备份；

3）数据预处理：负责对接收到的原始数据进行预处理，包括无效数据的剔除、根据实际科研需求对数据进行整合、压缩等。同时，将经过预处理的数据通过网络进行传输。

3.1.3 科研中心

科研中心，是科研人员进行长期科研活动的固定场所，如研究所或者实验室等，在本文设计的系统中主要包括以下几个子系统。

1）运行状态监控管理系统：负责整个系统运行状态的监控管理，根据不同的应用需求，可以设定不同的管理粒度。比如可以只对科研中心和野外科研基地的设备、系统运行情况进行管理，也可以具体到每个野外科研现场的设备运行状态的管理。

2）数据处理系统：负责科研数据的加工、处理，包括使用各种算法和模型对原始数据进行建模、分析等。

3）数据库系统：负责原始科研数据及各种系统相关信息的存储和管理。

4）Web 系统：提供科研数据共享和成果展示的平台，使科研人员和其他用户可以通过 Web 的方式进行查看。

下面讨论系统 3 个部分之间的通信问题。

3.2 野外科研现场与野外科研基地之间的数据通信

野外科研现场分布在野外，一般无法通过有线网络的方式来实现通信。在一些野外应用中，使用 GPRS[3,4] 来实现数据的传输，但是存在如下几点不足：

1）GPRS 的传输速率太低，一般只有十几 kbits/s，对于数据量稍大的观测设备，例如视频、图像等，GPRS 传输就无能为力了；

2）GPRS 为单向通信，终端设备没有公网 IP 地址，只能由终端设备主动发起连接才能够建立与远程的通信链路，如果要保持联系，必须每隔一定的时间发送“心跳”包，否则终端设备的 GPRS 连接会中断，频繁地发送“心跳”包会给终端设备带来额外的开销；

3）GPRS 信号不稳定，科研现场地处野外，移动通信基站的运行状况和信号覆盖程度等不如城市等人口密集的地方，经常会出现信号差、无法完成数据传输，甚至没有信号的情况；

4）从费用角度考虑，如果为每个野外观测设备都配备 GPRS 通信设备，那么整个系统的开销必然增加；

5）在许多科研现场由于地处偏僻，根本就没有移动通信网络，这种情况下自然也无法使用 GPRS 来完成数据传输。

考虑到上述问题，在本系统中，设计使用无线网桥和无线基站的通信方式来实现从野外科研现场到野外科研基地的数据传输，这种方式的优势如下：

1）无线网桥传输距离可达十几公里，而且可以通过中继的方式实现更远距离的传输，扩大了数据传输范围；

2）相邻的观测设备可以通过有线的方式互联，将数据汇聚到一点之后，再通过无线网桥传送出去；

3）无线网桥和无线基站之间构建的无线通信链路，能够满足野外科研现场与野外科研基地之间双向通信的需求；

4）根据周边环境的不同，无线通信链路的带宽可以达到几 Mbits/s 到几十 Mbits/s，远远高于 GPRS 的带宽，能够支

持大数据量的传输；

5)从费用角度考虑，无线网桥和无线基站只需要一次性投资，即可免费长久使用，而且在网络的管理和维护方面都能够实现自主控制。

3.3 野外科研基地与科研中心的数据通信

野外科研基地一般具有较好的基础设施条件，例如一般都有持续的电力供应，有电话线接入等。

对于野外科研基地与科研中心的数据通信，有两种方式。

1)有线网络方式，对于条件较好的野外科研基地，可以通过接入 ISP 的有线网络，例如专线、ADSL 等，实现与科研中心的数据通信；

2)移动通信网络方式，对于不具备有线网络接入的野外科研基地，可以通过 GPRS 或者 3G 方式实现与科研中心的数据通信。

对比 3.2 节讨论的在野外科研现场使用 GPRS 实现数据传输的方式，在野外科研基地采用移动通信方式有以下几点不同。

1)传输数据的有效性：在野外科研基地可以对从野外科研现场采集到的原始科研数据进行本地备份，然后可以对原始数据进行预处理，例如排除无效数据、合并数据等，将处理过的数据再进行传输，提高了数据传输的有效性；

2)移动通信质量：因为野外科研基地一般接近城镇，信号质量会比在野外科研现场好很多，保证了网络连接和数据传输的稳定性；

3)费用：在野外科研基地将数据汇聚之后，只需要一个 GPRS 传输设备即可实现到科研中心的数据传输，费用会大幅度减少。

4 原型系统实现

4.1 原型系统构建

以本文提出的野外数据采集传输系统为基础，作者构建了一个原型系统，如图 2 所示。原型系统具体说明如下：

1)野外科研现场使用支持 Zigbee[5] 协议的无线传感器网络[6] 设备来模拟；

2)野外科研基地使用一台数据采集与预处理服务器来模拟；

3)科研中心使用一台数据处理服务器、一台数据库服务器和一台 Web 服务器来模拟。

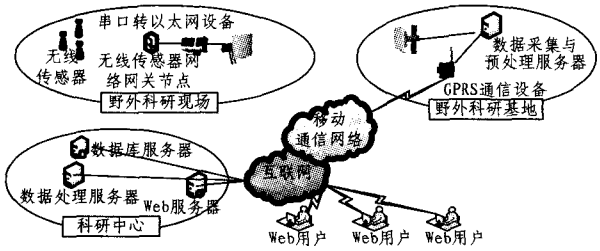

图 2 原型系统结构图

原型系统中所使用的无线传感器网络包括一个支持 RS232 串口输出的网关节点、一个温度传感器节点和一个湿度传感器节点。两个传感器节点与网关节点之间通过 Zigbee 协议组网并实现数据传输。网关节点通过一个 RS232 串口实现数据的输出。为了将该网关节点连接到网络中，在原型系统中使用了一个 RS232 转以太网的串口服务器设备，该设备可以实现从 RS232 接口读写串口数据，并封装成 IP 包，再通过网络进行转发。

在原型系统中使用一对无线网桥来模拟从野外科研现场到野外科研基地的无线连接。野外科研基地与科研中心的数据传输使用 GPRS 通信设备来实现。

4.2 原型系统软件设计与实现

原型系统的软件包括两部分：一部分是运行于野外科研基地的数据采集与预处理系统，另一部分是运行于科研中心的数据处理系统、数据库系统和 Web 应用。应用系统使用 JAVA 语言开发，数据库系统使用 MySQL。

数据采集与预处理系统的处理流程和功能如下：该系统与串口转以太网设备建立 Socket 连接，监听本地端口；本地端口在收到数据包后进行解析，在保存原始数据的同时，根据设定的规则进行数据的预处理，将处理后的数据生成固定大小的数据包，通过服务器的串口输出给 GPRS 通信设备，由该设备经过移动通信网络和互联网传送到科研中心数据处理服务器的指定端口；该系统也可以将科研中心或者野外科研基地的操作指令，如修改采样间隔、传感器能量阈值等，通过 Socket 连接发送到串口转以太网设备，由该设备发送给无线传感器网络网关节点，再由其通过 Zigbee 协议下发到无线传感器；该系统还负责对野外科研现场设备的运行状态进行监控，如果在一定的时间间隔内未收到任何数据，则会发出告警，提示需要对野外科研现场的设备进行故障检测。

数据处理系统的处理流程和功能如下：该系统监听本地端口，对接收到的数据进行解析，插入数据库中相应的数据表；操作指令的下发；野外科研现场和野外科研基地运行状态信息的维护和管理。

Web 应用提供如下功能：用户的认证管理；用户对设备的授权访问；野外科研基地与野外科研现场设备运行状态的监控；野外科研现场设备实时数据的展示；选定时间段的历史数据分析与展示等。

目前，原型系统软件已完成了部署和运行测试，表现稳定。图 3、图 4 所示为原型系统的 Web 应用中传感器的实时数据和历史数据展示。

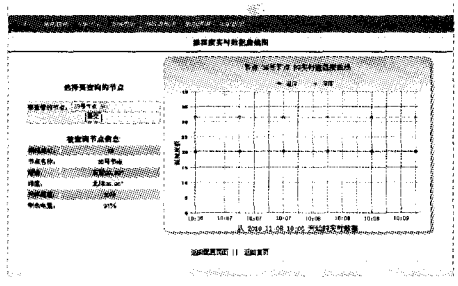

图 3 实时数据展示

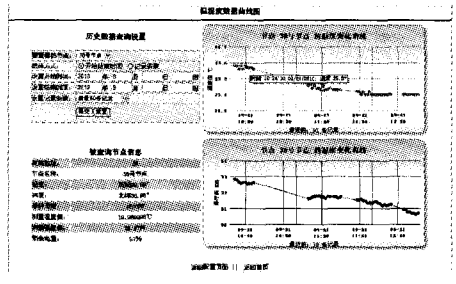

图 4 历史数据展示

(下转第 149 页)

② Master 接收到挖掘请求后，向 NameNode 申请所需的 XML 数据文件，同时访问空闲节点列表，将 ServiceNode 的元数据（机器名、IP 地址是否空闲）返回到 Master。Master 将元数据发送给算法存储节点，算法存储节点将 Apriori 算法发送到原始数据所在节点。

③各 ServiceNode 首先扫描本地数据库，统计库中事务的个数、每个项的出现次数，然后根据挖掘流程和 Apriori 算法，得到局部的候选 1-项集，再把统计结果和局部候选 1-项集发送到 Master 计算得出全局 1-项集，然后再把全局频繁 1-项集发送到各个 ServiceNode 生成更精确的局部频繁 1-项集，再由局部 1-项集得出局部候选 2-项集，扫描本地数据库中的事务，统计每个项的出现次数，把新的局部候选 2-项集和统计结果发往 Master……如此重复，直到生成符合用户定义的满足最小支持度的频繁项集，最后根据置信度阈值生成规则。

④Master 将得到的关联规则返回给用户。

3.6 算法结果

该系统由 7 台服务器（均安装 Linux 以及 Hadoop 云计算系统）组成，其中 1 台作为客户端和主控节点，1 台作为算法存储节点，5 台作为服务节点 ServiceNode。在并行执行过程中，时间消耗主要在各节点之间建立连接以及数据的传输。首先，将所有数据放在主节点上直接调用 Aprior 算法，计算出执行时间；然后将数据集分割成 5 个子文件分别保存在 5 个 ServiceNode 上，将 Aprior 算法从算法存储节点上并行传到 1、3、5 个 ServiceNode 上执行，计算出时间；最后将 Aprior 算法分别拷贝到 5 个 ServiceNode 上，将数据文件传输到 1、3、5 个 ServiceNode 上执行，计算出时间。通过 3 个实验对比，可以发现执行效率随着数据量的增明显得到提高。同时，随着数据量的增加，向存储节点传输算法的时间也明显少于向算法节点传输数据。

本文基于云计算平台改进的 Aproior 算法，由于其对各个节点频繁项集的筛选都是在全局端进行的，因此既不会流失有效的关联规则，也不会产生无效的关联规则[7]。

结束语 传统数据挖掘系统运行于 UNIX 小型机的集中平台上，这在海量数据以及应用愈加复杂的 Web 挖掘中受到很多限制。与传统 Web 数据挖掘相比，基于云计算的 Web 数据挖掘系统通过“云”中多个资源完成原先由一个节点承担的挖掘工作，使资源得到了充分利用，提高了数据挖掘过程的效率。基于云计算的数据挖掘工作意义重大，它不仅能够提高挖掘效率，还克服了网格环境的弊端，能够面向商业应用，更具有价值。

参 考 文 献

[1] 李健，徐超，谭守标. 一种 Web 数据挖掘系统的设计和研究[J]. 计算机技术与发展，2009，19(2)

[2] 张涛. Web 数据挖掘现状分析[J]. 科学之友，2009，6(17)

[3] 潘正高. Web 数据挖掘技术综述[J]. 电脑知识与技术，2009，5(15)

[4] 席景科，闫大顺. Web 数据挖掘中数据集成问题的研究[J]. 计算机工程与设计，2006，8(27)

[5] 纪俊. 一种基于云计算的数据挖掘平台架构设计与实现[D]. 青岛：青岛大学，2009

[6] 郑晶. 基于网格的并行数据挖掘算法的实现[J]. 福建工程学院学报，2010，2(8)

[7] 齐玉成，郑丽英，高三营. 基于网格的数据挖掘算法[J]. 电脑知识与技术

[8] Cannataro M, Talia D, Trunfio P. KNOWLEDGE GRID: High Performance Knowledge Discovery on the Grid[C]// Lecture Notes In Computer Science, Vol. 2242, Proceedings of the Second International Workshop on Grid Computing. 2001: 38-50

[9] Ye Yan-bin, Chiang C-C. A Parallel Apriori Algorithm for Frequent Item sets Mining[C]//Proceedings of the Fourth International Conference on Software Engineering Research Management and Applications(SERA'06). 2006: 87-94

[10] Armbrust M, Fox A, Griffith R, et al. Above the Clouds: A Berkeley View of Cloud Computing

[11] 万至臻. 基于 MapReduce 模型的并行计算平台的设计与实现[D]. 杭州：浙江大学，2008

[12] 王鹏. 云计算的关键技术与应用实例

[13] 郑庆华，刘均，田锋，等. Web 知识挖掘：理论、方法与应用[M]. 北京：科学出版社，2010

（上接第 135 页）

结束语 本文针对有野外数据采集需求的 e-Science 应用，提出了一种数据采集传输系统的设计思想。在此基础上构建了原型系统，并完成了相关软件的开发，实现了系统设计的基本功能。本文提出的系统，对于提高野外 e-Science 应用的工作效率、系统管理能力有重要的意义。下一步，将选择在青海湖国家级自然保护区、黑河流域等实际的野外科研环境中部署该系统，并根据野外应用的特点进行系统的优化和改进。

参 考 文 献

[1] 宋琳琳. E-Science 发展情况简介[J]. 图书馆学研究，2005(07)：21-23

[2] Taylor J. The Definition of e-Science [OL]. http://www.lesc.ic.ac.uk/admin/escience.html, 2005-10-13

[3] Hey T. Trefethen A E. Cyberinfrastructure for e-Science[J]. Science, 2005, 308(5723): 817-821

[4] ETSI. GSM 02.60: Digital cellular telecommunications system (Phase 2+): General Packet Radio Service (GPRS)[S]. Service Description Stage 1. 1998

[5] ETSI. GSM 03.60: Digital cellular telecommunications system (Phase 2+): General Packet Radio Service (GPRS)[S]. Service Description Stage 2. 1998

[6] IEEE 802.15.4. Standard-2003, Standard for Part 15.4. Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for Low-Rate Wireless Personal Area Networks (LR-WPANS)[S]. 2003

[7] 孙利民，李建中，陈渝，等. 无线传感器网络[M]. 北京：清华大学出版社，2005