文章编号:1671-4598(2017)10-0278-03 DOI:10.16526/j.cnki.11—4762/tp.2017.10.071 中图分类号:TP391 文献标识码:A

基于云计算的大数据自动分类处理系统设计

罗 弦,查志勇,徐 焕,刘 芬,詹 伟
(国网湖北省电力公司 信息通信公司,武汉 430000)

摘要:随着现代网络技术不断进步,系统数据量也在逐渐增多;传统的大数据自动分类处理系统已经无法满足现阶段用户需求,其软件与硬件的设计都比较单一,存在能源消耗大、分类速度慢、处理时间长、内存占用率高等问题,为此,提出基于云计算的大数据自动分类处理系统的设计;首先设计系统硬件结构,主要包括数据采集器、数据处理器以及数据自动存储模块,并详细的介绍了各硬件结构;然后利用时域特征提取数据的算法对频域特征数据进行提取,从而实现数据自动分类处理系统的软件设计;最后对两种系统性能进行对比实验;实验结果证明,基于云计算的大数据自动分类处理系统的资源不仅占用率低,内存消耗小,而且数据库内存较大;该系统不但可以提高数据自动分类精准度,还能加快数据分类速度,从而使系统拥有更好的分类性能。

关键词:云计算;大数据;自动分类;数据处理;系统设计

Design of Large Data Automatic Classification and Processing System Based on Cloud Computing

Luo Xian, Zha Zhiyong, Xu Huan, Liu Fen, Zhan Wei
(Information & Communication Branch, Hubei EPC, Wuhan 430000, China)

Abstract: With the continuous improvement of modern network technology, the amount of data in the system is increasing gradually. Traditional big data automatic classification processing system has been unable to meet the needs of users, the software and hardware design are single, there exists large energy consumption, slow speed of classification, long processing time and memory usage rate is high, therefore, automatic classification is proposed based on cloud computing of large data processing system design. Firstly, the hardware structure of the system is designed, which mainly includes data collector, data processor and data automatic storage module, and introduces the structure of each hardware in detail. Then, the data is extracted using the time-domain feature extraction algorithm to realize the software design of data automatic classification and processing. Finally, two kinds of system performance design are compared. The results show that the resources of large data automatic classification and processing system based on cloud computing have low occupancy rate, small memory consumption and large memory of database. The design of this system can not only improve the accuracy of automatic classification of data, but also speed up the classification of data, so that the system has better classification performance.

Keywords: cloud computing; big data; automatic classification; data processing; system design

0 引言

近几年随着网络技术的不断进步,各种系统中的数据量也在逐渐地增多,但是面对丰富的数据资源却让使用者很困惑,大量的数据呈现无序、分散的状态,从而增加了使用者对数据信息利用的难度[1-3]。传统的大数据自动分类处理系统的结构具有单一性,其能源的消耗、分类的速度、处理的时间、内存的占用率都不能满足当下大量数据分类的需求[4-5]。随着时间的流逝,大量数据逐渐形成了特殊的特征趋势,传统大数据自动分类处理系统的不稳定性很难对数据进行自动的分类,因此,能否设计出优于传统大数据自动分类处理的系统,是数据自动分类领域应该重点关注的内容[6-7]。

文献[8]中提出了一种基于关联规则的大数据自动分类处理系统的设计,该系统具体的数据挖掘过程是:利用迭代来获取数据的全部项集,其支持的力度高于既定阈值的力度即可,通过对项集的频繁搜索即可获得符合使用者的最优规则,并依据数据挖掘的关联规则对大数据进行自动的分类处理。但是该设计方法受到系统硬件条件的制约,运行的效果较差,能源消耗较多。文献[9]中提出了一种基于向量的数据自动分类处理系统的设计,该系统设计的风险较小,不会受到数据维数的影响。其设计的过程中,分类的数据将置于两种数据样本之间距离较远的位置,并经过高维空间的变化,低维线性存在的不可分的问题就迎刃而解了,从而实现大数据的自动化分类,但是该系统的设计会严重造成数据分类的单一性,性能效果不佳。文献[10]中提出了一种基于信息互动的大数据特征提取系统的设计,该系统以信息互动为准则,对数据特征进行分类与对比,并利用迭代算法对系统的软件进行设计,进而对数据进行准确的分类。虽然该系统的准确率较高,但是资源的占用率较少以及稳定的性能较低。

针对上述存在的问题,我提出了基于云计算的大数据自动分类处理系统的设计。首先设计了系统的硬件结构,主要有数据采集器、数据处理器以及数据自动存储模块,并详细的介绍了各硬件的结构;然后利用时域特征提取数据的算法对频域特征数据进行提取,从而实现数据自动分类处理的软件设计;最后对两种系统性能设计了对比实验。实验结果证明,基于云计

收稿日期:2017-07-01; **修回日期**:2017-07-29。

作者简介:罗 弦(1982-),男,湖北武汉人,硕士研究生,工程师,主要从事大数据与网络安全方向的研究。

算的大数据自动分类处理系统的设计不但提高了分类的精准度，而且降低了能源的消耗，其系统应用将会有更广阔的前景。

1　系统硬件设计

大数据自动分类处理系统的硬件是基于云计算设计的，云计算是在网络相关服务的程序下，对资源提供动态的易扩展的方式，并根据使用者的需求，将大数据进行分布式的配置，并以 SOA 组件模型的体系为基础，增加云计算的兼容性，从而提高大数据自动分类处理的稳定性。系统硬件的设计框图如图 1 所示。

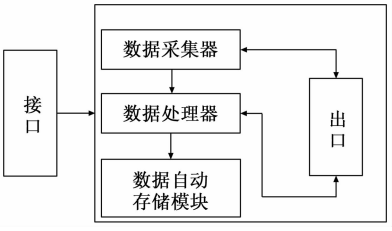

图 1　系统硬件的设计框图

1.1　数据采集器的设计

数据采集器的设计主要包括钛网的管制的芯片以及单片机，通过云计算的接口向大数据自动处理器传送采集到的数据。数据采集器的电源产生的是 5 V 的电压，并经过单片机的引脚传送到单片机上方的电压调节器中，为单片机上方的工作提供 3 V 的电压。再将单片机上方的 3 V 电压通过引脚传送到其它剩余所需 3 V 电源的器件中供其使用。单片机经过传送的信息与引脚传送电压结束后与其它的单片机进行信息之间的交换。基于云计算的网络信号经过电路调整后，使用 p25 的引脚传送到单片机上方的 A/D 转换器当中，并通过 A/D 转换器将网络信号转换成数据，从而实现了基于云计算环境下的大数据的采集。

1.2　数据处理器设计

基于云环境下的数据处理器主要用于对采集到的大数据进行处理。处理器主要选用的是某网络公司生产的 IXP2400 的处理器，采用共享效率高的数据线程以及微引擎的数据信号对收集到的数据进行处理。通过控制处理器对采集到的数据进行处理，这个过程是可完全编程的，处理器工作的模式也可以利用编程来实现。

1.3　数据自动存储模块的设计

数据的存储模块采用的是 C8051F 系列的单片机来完成数据的存储。C8051F 系列的单片机集合成了完全混合的 soc 芯片，其内置的 FLASH 存储的程序具备较大的存储空间。C8051F 系列的单片机与 AT 45DB80 的硬件工作原理如图 2 所示。

由图 2 可知，将 C8051F 系列的单片机 $P_{0.2}$、$P_{0.3}$、$P_{0.4}$ 引脚采用设置开关为 MOSI 的信号主线，每条主线都与 AT45DB80 的硬件的始终进行串联并将信号输出。将 $P_{3.0}$、$P_{3.1}$、$P_{3.2}$ 和 AT45DB80 的硬件中的芯片连接，并进行复位。C8051F 系列的单片机采用的是存储器瞬间开启的一次性数据储存，其过程是：先将串行外设接口的表示进行清除，然后再向数据的自动储存器中输入字节，如果检测出的串行外设接口由 AT45DB80 硬件组成，那么一次的数据自动储存结束。

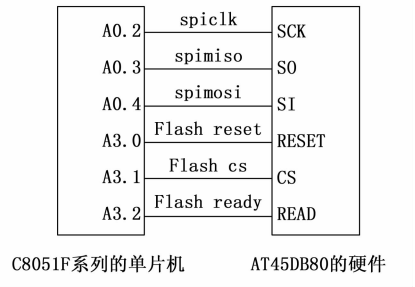

图 2　单片机与硬件工作原理图

2　系统软件设计

基于云计算的大数据自动分类处理系统的软件设计需要对大数据进行特征提取，然后进行分类处理。虽然大数据的特性在数据处理的时候较为复杂，但是对于自动分类处理的软件设计过程来说是必不可少的。其过程为：首先将数据进行人工的分类，来获取数据的样本，然后为了消除多余数据之间存在的可能性的矛盾对样本进行聚类，并对系统选取特征性的数据，最后对性能改进型评估，以便性能的改善。

2.1　基于时域特征提取数据算法的设计

时域的特征主要包括瞬时能量、平均方值的大小以及过零率以及高过零的帧数比。瞬时能量的单位为帧，对于大多的数据提取方法来说，一般参照瞬时能量每帧的点幅数值 z 的平方以及同一时间段的大数据增减的能量值 E，公式为：

$$E = \sum_{i=0}^{I-1} z^2(i) \tag{1}$$

由公式（1）可以看出瞬时的能量均方根值（RMS）的表示式为：

$$RMS = \sqrt{\frac{E}{I}} \tag{2}$$

其中：i 为帧数；I 为参照的点数。

大数据的特征可分为语音数据与文字数据，都可以通过帧数的大小呈现出来，因此瞬时能量的表达式能够准确的将语音数据的特征通过上述的公式准确的提取出来。

过零率 Q 的含义是能够在特定的时间内将大数据的正负幅度值的变化次数迅速的计算出来，其表示式为：

$$Q = \frac{1}{2Z-2}\left\{\sum_{i=0}^{I-1} \left|\operatorname{sgn}[Z(i)] - \operatorname{sgn}[Z(i-1)]\right|\right\} \tag{3}$$

公式（3）中，sgn 表示的是特定的参数；若数据的变量大于等于 0 时，特定的参数 sgn 的数值为 1；若数据的变量小于 0 时，特定的参数 sgn 的数值为 −1。

高过零是在一定的时间内，其过零率的瞬时能量的数值超过其他平均数值的帧数的比（$HZCRR$），其表达式为：

$$HZCRR = \frac{1}{2I}\sum_{I=0}^{I-1}[1 + \operatorname{sgn}(Q(i)) - 1.5avQ] \tag{4}$$

公式（4）中：avQ 代表的是过零率的平均数值；$Q(i)$ 表示的是帧数为 i 时的过零率。

大数据中语音信号的数据往往呈现的是交替形式的出现，这就导致了过零率的波动呈上升趋势，高过零的帧数过大；而文字信号的数据波动的情况并不明显，高过零的帧数较小。根据上述的内容可以对大数据的特征进行分类。

2.2 基于频域特征数据提取的设计

频域特征主要运用的是线性预测和梅尔频率倒谱系数计算方法的数据提取，该方法能够对频率产生的瞬时能量进行数据的辅助提取。

梅尔频率倒谱系数是针对等距划分频带数据提取特征应用的一种计算方法，该方法拥有较高的抵抗干扰的能力，因此，常将该计算方法作为数据特征提取的主要手段之一。如果想要获取梅尔频率倒谱系数，需要对大数据的软件进行加重、帧数分类、添窗等设计，这时获取到的帧数时域信号用 $W(i)$ 表示。帧数的时域信号进行经过傅里叶的转变之后即可获取到离散的频谱，并用 $W(k)$ 来表示，那么表达离散频谱的公式为：

$$W(k)=\sum_{i=0}^{I-1}W(i)e^{\frac{-2\pi ik}{I}},0\leqslant k\leqslant I \tag{5}$$

式中，k 为傅里叶变换点数；e 为频率。

利用 $W(k)$ 能够计算出离散频率的数值，即为 $W^2(k)$，此时的输出数据的能量为：

$$E'=ln\Big[\sum_{i=0}^{I-1}\mid w(i)\mid^2H(k)\Big],0\leqslant k\leqslant M \tag{6}$$

其中：H 为处理后的输出数据的能量值；M 为处理的次数。根据数据分类处理的顺序，可得到梅尔频率倒谱系数计算的表达式为：

$$T(i)=\sum_{i=0}^{I}E'\cos\frac{\pi i(m-4)}{M},0\leqslant k\leqslant M \tag{7}$$

公式（7）中，m 为数据处理的顺序。由此可得出线性预测系数的表达式为：

$$T'(k)=\sum_{i=1}^{m}\alpha_iT'(k-1),k=0,1,\cdots,m \tag{8}$$

公式（8）中，m 为线性预测数据的阶段；$T'(k)$ 为第 k 个序列实数的组合；i 为自然数。通过上述的内容，可完成系统软件的设计。

3 实验结果与分析

为了验证大数据自动分类处理系统设计的有效性进行了实验，其中数据来自于网络知识库，系统是由 3 台计算机组成，其中系统的硬件配置有：Intel Dual－core 2.6 GHz 型号的处理器、16 GB 的内存大小。

3.1 参数的设置

将实验的数据进行编号，分别为：T_0、T_1、T_2、T_3、T_4、T_5、T_6；数据的种类分别为：经济学数据、农业经济数据、贸易经济数据、世界经济数据、工业经济数据、交通运输经济数据；数据的大小分别为：1686、1789、1893、1595、1537、1678。

3.2 数据的分析

根据上述的参数，分别对传统的大数据自动分类处理系统与基于云计算的大数据自动分类处理系统的稳定性进行了分析。

由表 1 可知：传统的大数据自动分类处理系统在六次的实验中，其数据分类的准确率随着实验次数的增多，数据分类的准确率和数据分类的预测值变高，而系统数据的召回率始终维持在 91％左右。

由表 2 可知：基于云计算的大数据自动分类处理系统在六次的实验中，其数据分类的准确率随着实验次数的增多，数据分类的准确率和数据分类的预测值变高，而系统数据的召回率则高达 99％。

表 1 传统的大数据自动分类处理系统

实验序号	数据分类的准确率/％	系统数据的召回率/％	数据分类的预测值
T_0	90.34	90.28	90.28
T_1	91.12	90.96	90.57
T_2	91.34	91.35	91.69
T_3	92.28	91.47	92.78
T_4	92.56	91.34	93.29
T_5	93.13	92.56	93.83
T_6	94.52	91.15	94.12

表 2 基于云计算的大数据自动分类处理系统

实验序号	数据分类的准确率/％	系统数据的召回率/％	数据分类的预测值
T_0	91.17	92.34	92.15
T_1	92.54	94.13	92.54
T_2	93.52	96.35	93.69
T_3	94.74	97.20	94.72
T_4	95.56	98.34	95.89
T_5	9613	98.52	96.38
T_6	97.52	99.05	97.89

3.3 实验结果

由上述的实验过程可以分析出大数据自动分类实质上就是一个映射的过程，根据数据特征的提取可以充分的体现出基于云计算的大数据自动分类处理的准确程度。一般情况下采用数据分类的准确率与系统数据的召回率这两个指标作为对系统评估的判断。由上述实验内容中的表 1 与表 2 可以看出，采用基于云计算的大数据自动分类处理系统对各种数据进行了分类，并得到数据分类的准确率与召回率的优势都明显高于传统的大数据自动分类处理系统。

为了进一步验证基于云计算的大数据自动分类处理系统设计的有效性，对 CPU 的占用率与内存占用率的情况进行对比。

表 3 两种系统的资源占用率的对比结果

系统设计的流程	传统的大数据自动分类处理系统		基于云计算的大数据自动分类处理系统	
	CPU 的占用率	内存占用率	CPU 的占用率	内存占用率
大数据的收集	80	48	69	35
数据特征的提取	62	35	58	20
数据的自动分类	63	40	58	30

由表 3 可知：基于云计算的大数据自动分类处理系统的 CPU 的占用率结果的范围为：58％～69％，内存占用率的范围为：20％～35％；而传统的大数据自动分类处理系统的 CPU 的占用率结果的范围为：62％～80％，内存占用率的范围为：35％～48％。

传统的大数据自动分类处理系统与基于云计算的大数据自动分类处理系统在内存的损耗与分类的速度上也大不相同，如图 3 所示。

（下转第 288 页）

作为试验依据，以确保定型试验质量。三是建立健全规章制度，将有关机关部门和使用、保障、论证、试验、研制承担单位的职责、分工以试验管理文件明确，并下发执行，同时研究确定工作程序[9]。

5.2 确定一体化试验质量管理

参照设计定型试验的相关文件、标准，对纳入一体化试验总体规划的试验项目进行质量管理。一是在试验进场前，被试品通过必要的试验，能够证明一体化试验的被试品关键技术问题已经解决，主要战术技术指标能够达到规定要求。二是被试品技术状态已确定，经订购方认可，被试品技术状态更改及其评审、验证和确认的完整记录。三是试验实施过程严格执行试验大纲、技术准备操作程序、试后处理方案等，特殊过程和关键过程控制符合要求，运行的质量管理文件应符合 GJB1452A－2004《大型试验质量管理要求》等设计定型试验的相关文件和标准规定。

6 结束语

针对现代水中兵器试验鉴定发展需求，通过分析水中兵器试验现状，提出一体化试验需求和构想，探讨一体化试验多源信息融合的基本方法，提出一体化试验管理要求，为解决一体化试验总体设计问题提供一种解决方案。一体化试验探索是当前武器装备试验转型发展的重要课题，一体化试验模式需要从装备牵引、理论研究和工程实践 3 个层面同时推进，所涵盖试验组织、管理、技术、保障等各方面内容，有待进一步深入研究。

参考文献：

[1] 李洪涛，高顺林，谢君红. 水中兵器一体化试验研究 [J]. 国防科技，2013，1：1－4.

[2] 刘保根，张召奎. 论水雷一体化试验 [J]. 水雷战与舰船防护，2013，5：82－84.

[3] 雷 帅，丁士民. 基于一体化试验鉴定发展的试验鉴定的发展规律 [J]. 国防科技，2011，32 (4)：36－39.

[4] 晓 丽，李晓斌. 美军武器装备一体化试验与评价综述 [J]. 海上靶场学术，2009，6：12－15.

[5] 崇 虎，张佐成. 飞航导弹研制和定型一体化试验研究与探讨 [J]. 飞航导弹，2005，3：8－12.

[6] 武小悦，刘 琦. 装备试验与评价 [M]. 北京：国防工业出版社，2008.

[7] 蔡 洪，张士峰，张金槐. Bayes 试验分析与评估 [M]. 长沙：国防科技大学出版社，2004.

[8] 张湘平. 小子样统计推断与融合理论在武器系统评估中的应用研究 [D]. 长沙：国防科技大学，2003.

[9] 王国盛，洛 刚. 美军一体化试验鉴定分析及启示 [J]. 装备指挥技术学院学报，2010，21 (2)：95－98.

(上接第 280 页)

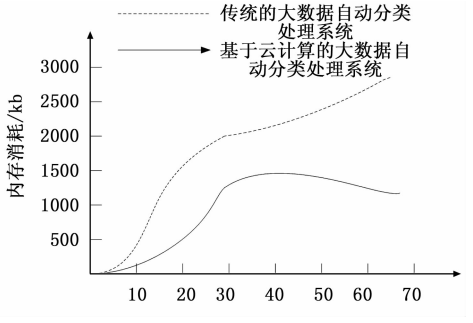

图 3 两种系统能耗与速度的对比结果

由图 3 可以看出，基于云计算的大数据自动分类处理系统的内存消耗明显高于传统的方法，其分类的时间比传统的方法节省很多。

由此可得出实验结论为：基于云计算的大数据自动分类处理系统的资源占用率低，内存消耗较小，且数据库的内存较大，该设计方法不仅提高了数据自动分类的准确度，还加快了数据分类的速度，具有较好的分类性能。

4 结束语

基于云计算的大数据自动分类处理系统的设计与传统的大数据自动分类处理系统相比具有良好的稳定性，其资源的占用率也比较低，分类的速度较快。数据自动处理后的显示端是用户直接应用的端口，该端口的任务就是对大数据进行收集与获取结果进行显示并标注分类。

对系统进行硬件设计就是为大数据提供自动分类处理数据的平台，并将数据的特征进行分类处理，将处理的结果传送给逻辑层的处理端。而系统的软件设计就是为了实现数据自动分类处理更加的准确，为此使用了时域特征提取数据的算法，利用该算法对频域特征数据进行提取。基于云计算的大数据自动分类处理系统的设计不但提高了分类的精准度，而且降低了能源的消耗，为我国未来的数据处理方式提供了强有力的依据。

参考文献：

[1] 肖乃慎，李 博，孔德诗，等. 大数据背景下的电网客户用电行为分析系统设计 [J]. 电子设计工程，2016，24 (17)：61－63.

[2] 刘 莉，杨傲雷，屠晓伟，等. 面向 INS 数据分类的鲁棒性无监督聚类方法 [J]. 仪器仪表学报，2016，37 (1)：152－160.

[3] 余 翔，白友良，李 成，等. 多维有序聚类法在地质数据分类中的应用 [J]. 计算机应用，2015 (s1)：152－155.

[4] 陈学斌，王 师，董岩岩，等. 面向大数据的并行分类混合算法研究 [J]. 微电子学与计算机，2016，33 (4)：138－140.

[5] 孟丽丽，宋 锋. Web 网络大数据分类系统的设计与改进 [J]. 现代电子技术，2016，39 (22)：36－40.

[6] 张 青，吕 钊，ZHANGQing，等. 基于主题扩展的领域问题分类方法 [J]. 计算机工程，2016，42 (9)：202－207.

[7] 张明卫，朱志良，刘 莹，等. 一种大数据环境中分布式辅助关联分类算法 [J]. 软件学报，2015，26 (11)：2795－2810.

[8] 李 悦，孙 健，邱志祺. 基于关联规则的数据挖掘技术的研究与应用 [J]. 现代电子技术，2016，39 (23)：121－123.

[9] 蒋 亮，蒙祖强，胡玉兰，等. 一种基于向量夹角的快速计算等价类算法 [J]. 小型微型计算机系统，2015，36 (10)：2360－2364.

[10] 张科星. 网络大数据平台中的特征数据分类系统设计与实现 [J]. 现代电子技术，2017，40 (8)：25－28.