

基于 Hadoop 的海量图像检索系统

王 梅,朱信忠,赵建民,黄彩锋

(浙江师范大学 数理与信息工程学院,浙江 金华 321004)

摘 要:在传统图像检索系统中,由于采用单节点架构,面对海量图像数据检索时存在检索速度慢、并发性差等问题。文中提出了一种基于 Hadoop 的图像检索方法,将图像检索技术与 MapReduce 框架相结合,图像特征库存储于分布式文件系统 HDFS 中,计算节点采用基于 Hadoop 的分布式存储调度算法,增强对多数据的并发处理能力,同时对计算后的数据进行压缩处理。实验表明,该方法在处理大数据图像检索时,与单节点检索系统相比,能够有效降低检索时间,提高检索速度。

关键词:Hadoop; MapReduce; 分布式计算; 图像检索

中图分类号: TP391

文献标识码: A

文章编号: 1673-629X(2013)01-0204-04

doi:10.3969/j.issn.1673-629X.2013.01.050

Massive Images Retrieval System Based on Hadoop

WANG Mei, ZHU Xin-zhong, ZHAO Jian-min, HUANG Cai-feng

(College of Mathematics and Physics and Information Engineering,

Zhejiang Normal University, Jinhua 321004, China)

Abstract: In a traditional single-node architecture image retrieval system, facing the problems of slow retrieval speed, poor concurrency etc when retrieved massive image data, proposed an image retrieval method based on the Hadoop, combining image retrieval technology with MapReduce frame, image feature database stored in the distributed file system HDFS, computing nodes using the scheduling algorithm based on Hadoop distributed storage, enhanced concurrent processing capability for multiple data, at the same time compressed the calculated data. Test and experiment results show that, the method in dealing with large data retrieval, compared with the single-node retrieval system, can effectively reduce the search time, improve the retrieval speed.

Key words: Hadoop; MapReduce; distributed computing; image retrieval

0 引 言

图像检索是直接根据初始查询图像的视觉特征,在图像库中找出与之相似的图像。利用图像自身去检索图像,快速有效地提高了图像检索的性能,但在图像检索的过程中,同时将消耗大量的 CPU 资源。随着计算机科学技术和数字图像采集技术的迅速发展以及互联网的普及应用,每天从各行各业都产生出大量的多媒体数据,这些数据大部分是以图片和视频等形式表现的,传统基于单节点架构的图像检索系统存在检索速度慢、并发性差,实时性和稳定性无法保障等问题,已经不能满足人们对于检索性能的要求^[1]。因此一种基于内容的实现图像快速检索、并行处理、及时响应方

法成为了研究热点。

Hadoop 是 Apache 软件基金会 (Apache Software Foundation) 组织下的一个开源项目,提供分布式计算环境下的可靠、可扩展软件,是一个能够让用户轻松架构和使用的分布式计算平台,能够支持上千个节点以及 PB 级数据量的运算^[2,3]。Hadoop 分布式计算平台适合将各种资源、数据等部署在廉价的机器上,进行分布式存储和分布式管理,具有高可靠性、高扩展性、高效性以及高容错性等优点,有效提高图像检索的速度。文中在研究开源框架 Hadoop 的基础上,分析传统图像检索系统的基础上,结合基于内容的图像检索技术和 MapReduce 计算框架^[4],将图像特征库存储于 HDFS 中,开发实现了基于 Hadoop 的海量图像检索系统。

1 系统框架与模块设计

1.1 分布式系统整体框架

基于 Hadoop 的图像检索系统的设计目标是实现海量、异构、分布的图像资源的快速检索和及时响应。系统采用分布式构架,由上而下分别由表现层、业务逻

收稿日期:2012-04-21;修回日期:2012-07-25

基金项目:浙江省自然科学基金(Y1101269);省科技计划项目(2008C14063);浙江省重中之重学科资助项目

作者简介:王 梅(1988-),女,硕士研究生,研究方向为云计算资源调度策略;朱信忠,副教授,硕士生导师,研究方向为模式识别与数字工程、多媒体图像检索等。

辑层以及数据及数据处理层组成,整体框架如图 1 所示。

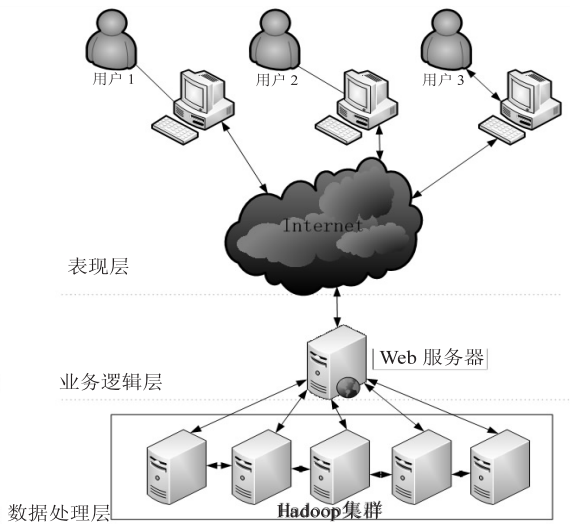


图 1 基于 Hadoop 的图像检索系统整体构架

前端用户通过 Internet 获取服务,用来上传示例图像和接收 Web 服务器的处理结果。在服务器端,业务逻辑层主要根据用户检索请求执行相应业务处理。数据及数据处理层包括 HDFS 存储和管理,海量图像数据导入和请求模块以及 HDFS 管理模块。数据处理层也是系统最核心的部分,负责图像数据的分块、图像特征的提取、匹配以及结果的返回。

1.2 HDFS 模块设计

HDFS 集群有两类节点,采用 Master/Slave 架构,以管理者-工作者模式运行,即一个 NameNode(管理者)和多个 DataNode(工作者)组成。当用户通过客户端发出请求对文件进行读写操作时,集群通过 NameNode 和 DataNode 的交互实现文件读写操作。HDFS 框架结构如图 2 所示。

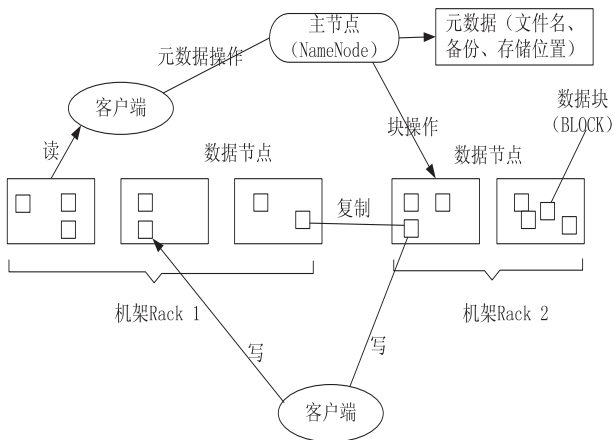


图 2 HDFS 框架结构

NameNode 是整个 HDFS 的核心,用于管理数据节点和客户端对文件的访问,管理文件系统的命名空间,维护整个文件系统的数据结构,记录和保存系统中所有的文件和元数据。这些信息以备份文件的形式保存

在设置备份的 NameNode 节点上,Hadoop 对多份复制的数据块自动进行维护,当集群中的某一节点数据丢失造成任务失败后,NameNode 会自动地重新部署计算任务。在 Hadoop 启动时,通过各 DataNode 收集各数据块的信息。

DataNode 是文件系统的工作节点,根据需要负责存储或检索数据块(受客户端或 NameNode 调度),各数据块的存储位置随 Hadoop 系统的调整而改变,DataNode 周期性向 NameNode 上报心跳。当客户端发送请求,NameNode 的监听程序被开启,当监测到客户端的请求时,NameNode 就将 HDFS 分布式文件系统的目录信息、磁盘空间信息、备份因子、空闲的节点数目等信息返回给客户端,客户端根据返回信息,使用 Hadoop 程序指令进行本地数据的处理、HDFS 数据导出导入等相应的操作。

1.3 MapReduce 模块设计

MapReduce 模块主要用于大规模数据集的并行计算,在本系统中主要负责在图像检索过程中对图像匹配及相似性度量的计算,并将匹配处理结果按照相似度从大到小排序后返回给用户^[5]。MapReduce 模块实现并行计算的工作框图如图 3 所示。

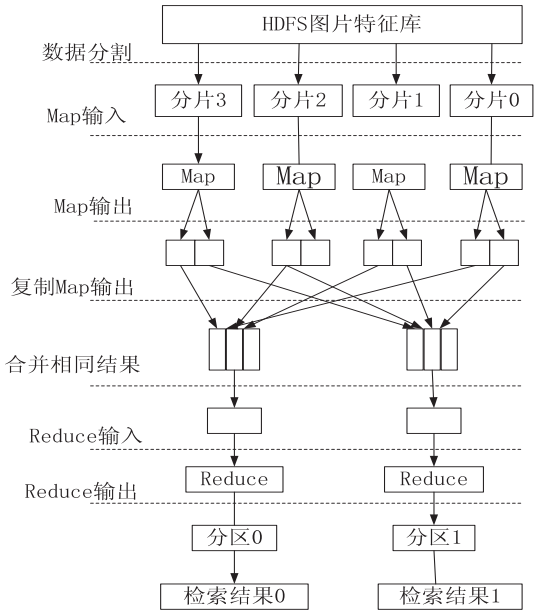


图 3 实现并行计算的 MapReduce 工作框图

首先,MapReduce 程序运行时,对 HDFS 中的图像特征库进行数据分割,得到图像特征数据的分片,然后由 DataNode 节点将每个数据分片传送至各 TaskTracker 节点进行运行,由 Map 读取数据分片,将数据分片分解为以 Key/Value 形式存在的数据特征,其中键值 Key 表示图像特征在数据分片中的偏移距离,键值 value 为图像的特征值。

然后 map() 函数调用这些 Key/Value 值,通过 map 函数判断图像特征是否满足检索条件,满足即计

算形状、纹理、颜色的相似度,并将匹配的中间结果以 Key/Value 键值对的形式保存在本地系统中,其中 Key 值表示相似度,Value 值表示图像特征库中的图像名。

然后对于 Map 任务输出的中间结果进行相同合并(相同结果只取一条),将处理好的中间结果传递给 Reduce 任务。

最后 Reduce 任务收到 Map 输出的中间结果后对其进行排序,排序规则按照相似度大小处理,并将排序好的检索结果存储于 HDFS 中。

2 系统实现关键技术

2.1 Hadoop 的大文件分布式存储

海量图像检索系统处理的数据可以达到 PB 级以上,传统的单节点存储无法达到要求。本系统存储平台选用 Hadoop 框架,对原始的大文件进行分块处理,采用基于 Hadoop 的分布式存储调度算法,该算法能够提高系统对多数据的并发处理能力^[6],同时采用压缩存储对多数据进行处理。

本系统在处理大文件分块存储时采用的存储方法^[7]是将一大文件分块处理成若干个数据分块,并将这些属于同一大文件的数据分块以一个文件的形式存储,利用分布式存储调度算法,将分块后的所有数据存储在不同的存储节点上,并实施相应的备份机制。每一个数据分块以<Blk_ID, MetaData>标识的一条<key, value>键值对进行存储,其中, Blk_ID 代表数据分块的流水号标识; MetaData 代表数据分块的二进制数据。因此,数据分块的存储类型采用<int, byte[]>型,由于每个数据分块的存储标识以键值对存储且一一对应,所以只要通过数据分块的流水号标识,就可得到该数据分块的字节数据,大文件分布式存储流程如图 4 所示。

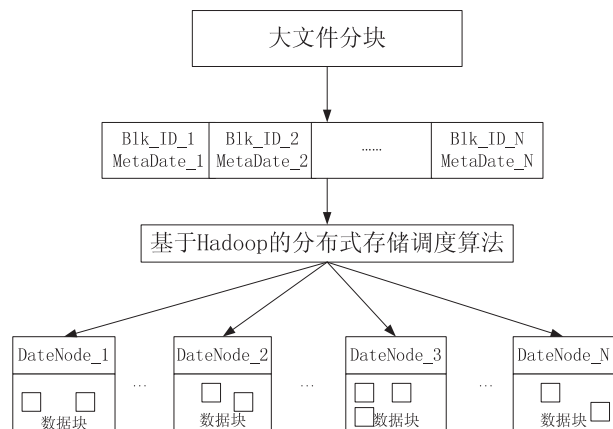


图 4 大文件分布式存储方法

2.2 图像特征提取与匹配

在传统图像检索过程中,对图像特征提取通常采用提取图像全局特征的方法来进行索引,如颜色、纹理

等^[8]。但是这种方法没有考虑到图像的局部区域特征,而只是简单地提取图像视觉特征,从而得到的检索效果往往不尽人意。因此在对图像进行查询时,必须考虑其区域特征。图像的感兴趣区域可以灵活地描述图像的细节内容。在描述图像的感兴趣区域时,通过提取图像的兴趣点集合的凸包,确定的凸包的区域即图像的感兴趣区域^[9]。

文中在提取图像感兴趣区域的颜色特征采用的方法,首先对图像进行空间转换,将 RGB 颜色空间转换成 HSV 颜色空间。然后对 HSV 颜色空间进行量化为 72 种主要颜色,并对感兴趣区域的颜色值按

$$H_k = num_k / num, k = 0, 1, 2, \dots, L - 1 \quad (1)$$

作直方图统计。其中: num_k 表示感兴趣区域颜色 K 的像素数量; num 是区域所有像素的数量; L 为量化后的颜色柄数。

利用灰度共生矩阵来表示图像的纹理特征。假设灰度图像为 $f(x, y)$, 其灰度级数为 L , 则有 $f(x, y) \in [0, L - 1]$ 对图像中的任一区域 R , 定义 S 为区域中具有特定空间联系的像素对的集合, 其归一化共生矩阵可用如下公式表示:

$$CM_{(\delta, \theta)}(i, j) = \frac{\text{card}\{[(x_1, y_1), (x_2, y_2)] \in S | f(x_1, y_1) = i \& f(x_2, y_2) = j\}}{\text{card}(S)} \quad (2)$$

式(2)中 $i \in [0, L - 1], j \in [0, L - 1], x_2 = x_1 + d \cos \theta, y_2 = y_1 + d \sin \theta$, $\text{card}(S)$ 为集合 S 中对 $CM_{(\delta, \theta)}(i, j)$ 有贡献的元素个数。

按公式(2)计算图像感兴趣区域的灰度共生矩阵,提取以下 4 个统计特征量表示为^[10]:

①能量。

$$E = \sum_{i=1}^D \sum_{j=1}^D [m(i, j)]^2 \quad (3)$$

②惯性。

$$I = \sum_{i=1}^D \sum_{j=1}^D (i - j)^2 \cdot m(i, j) \quad (4)$$

③熵。

$$S = - \sum_{i=1}^D \sum_{j=1}^D m(i - j) \cdot \log[m(i, j)] \quad (5)$$

其中,当 $m(i, j) = 0$ 时,有 $\log[m(i, j)] = 0$ 。

④匀度。

$$H = \sum_{i=1}^D \sum_{j=1}^D \frac{m(i, j)}{1 + (i - j)^2} \quad (6)$$

其中量化级数 D 为 8, 由上述特征量组成感兴趣区域的 4 维纹理特征向量, $F = [F_1, F_2, F_3, F_4]$ 。

文中给出图像相似性度量函数 $S(Q, I)$, 其中 Q 表示待检索图像, I 表示图像特征库中的图像, 相似性度量函数定义为:

$$S(Q, I) = w_1 \cdot S_{\text{color}}(Q, I) + w_2 \cdot S_{\text{cm}}(Q, I) \quad (7)$$

式(7)中 $S_{\text{color}}(Q, I)$ 表示两幅图像感兴趣区域的颜色特征的相似度,公式如下:

$$S_{\text{color}}(Q, I) = \sum_{k=0}^{L-1} \min(H_k(Q), H_k(I)) \quad (8)$$

而纹理特征相似度 $S_{\text{cm}}(Q, I)$ 用欧几里得距离计算,公式如下:

$$S_{\text{cm}}(Q, I) = \sqrt{\sum_{i=1}^4 (QF_i - IF_i)^2} \quad (9)$$

w_1 和 w_2 表示可调的权值,且满足 $w_1 + w_2 = 1$,文中这两个权值取 0.5。根据公式(7)计算待检索图像 Q 和图像特征库中的图像 I 相似度是 $S(Q, I)$,然后将图像检索结果按照相似度的大小排序。

2.3 MapReduce 并行计算实现

在基于 MapReduce 框架下实现大数据集的图像检索,最主要的是并行计算 Map 函数和 Reduce 函数的实现,Map 函数的功能是负责将数据分散处理,Reduce 函数的功能则是负责将处理后的中间结果进行聚集,编程过程中只需实现 Map 和 Reduce 两个接口,即可完成 TB 级数据的计算。

根据上述实现并行计算 Map/Reduce 模块设计工作框图(图3),可以将图像匹配计算的 Map 函数和 Reduce 函数实现过程用如图5所示的流程图表示。

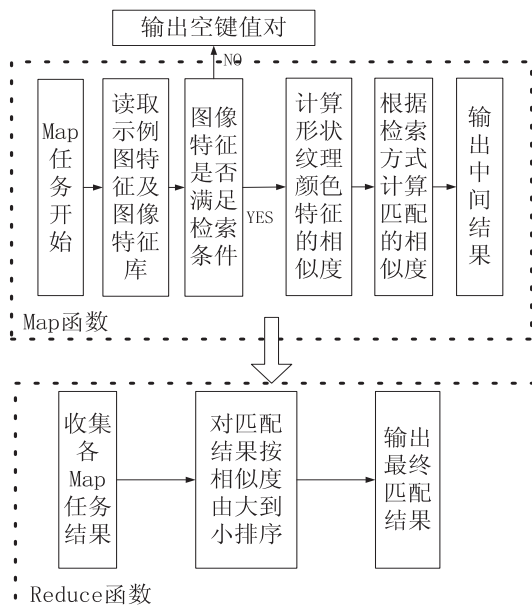


图5 MapReduce 算法流程

文中利用 Hadoop 提供的分布式计算框架,定义了 map 函数、reduce 函数以及实现 Map/Reduce 作业的 run 函数^[11,12]。在整个 Map/Reduce 过程中,map 的输入是关键。map 函数由 Mapper 接口实现,通过调用一个 map() 方法对每一个<Key, Value>键值对进行处理,并将处理后的中间结果写入到 context 中,并将结果存储于 TaskTracker 节点的本地文件系统之中。re-

duce 函数由 Reducer 接口来实现,调用 reduce() 方法重载^[13],把 map 输出的结果进行组合。实现 map 函数与 reduce 函数后,将 Mapper 和 Reducer 交给 Run() 函数实现 MapReduce 作业^[14]。

3 实验结果及分析

实验集群环境由四台普通 PC 机搭建(1 个 Master 节点,3 个 Slave 节点)。节点机器配置如下:CPU:2.1 GHz Pentium;内存:2 GB;硬盘:320 GB;以太网卡:100 Mb/s 全双工;操作系统:RedHat Linux。

为测试集群系统的性能,使用不同数量级的数据对系统进行测试,通过测试不同数量的图片特征库在不同节点数以及 B/S 单节点模式下图像检索的耗时,并进行对比。实验结果如图6所示。

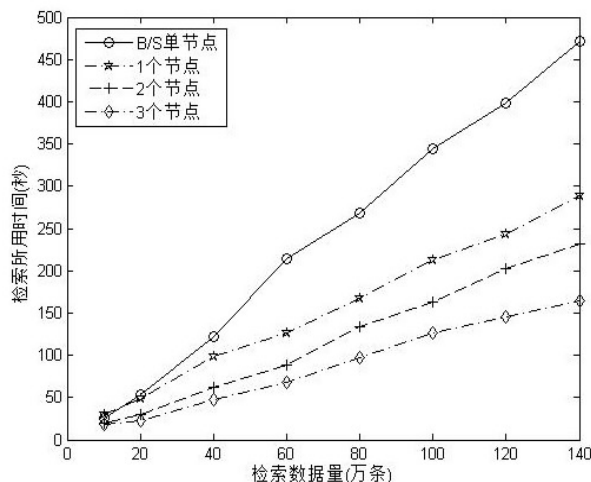


图6 不同节点下的检索时间对比

由图6可以看出,横坐标表示检索的数据量,纵坐标表示检索所用的时间:

(1)当数据量分别为10和20万条,且Hadoop集群中仅配置有1个节点时,检索速度与传统B/S单节点系统相比消耗时间稍长。主要是由于Hadoop集群的分布式特性,在执行MapReduce程序先要消耗一定的时间对任务进行初始化和分配,Hadoop集群的海量数据处理能力并没有发挥出来。

(2)随着数据量增大,Hadoop的性能优势开始发挥出来,从图中可以看出,当数据量从40万条到120万条时,传统B/S单节点模式检索时间极其漫长,甚至在一些条件下造成系统死机,不能完成检索任务,而在Hadoop集群中随着节点的增多图像检索的速度呈线性增长,远远低于单节点的检索时间,数据量越大,检索的效率越明显。

4 结束语

文中设计并实现了基于Hadoop的海量图像检索

系统,将大数据集图像检索任务进行分解,结合 Hadoop 分布式存储和 MapReduce 并行计算框架,通过各节点协同完成图像检索任务。通过选取不同数量级的图像数据进行测试,并与传统 B/S 单节点的图像检索系统进行对比验证,实验表明本系统与传统 B/S 单节点检索系统相比,能够有效改善检索的速度慢、并发性差等问题,有效提高了图像检索的速度、并发性以及处理海量数据的能力。

未来的工作重点在于研究在并行处理框架下如何提高现有图像检索算法的执行效率以及解决 Map 任务与 Reduce 任务之间数据传输过慢的问题,减少更多由于传输信息产生的时间消耗。

参考文献:

- [1] Wang Fei, Vuk E, David B, et al. Large-scale multi-model mining for healthcare with MapReduce [C]//Proceedings of the 1st ACM International Health Informatics Symposium. [s. l.]:[s. n.],2010:479-483.
- [2] HDFS[EB/OL]. 2011-12-08. <http://hadoop.apache.org/>.
- [3] White T. Hadoop:the Definitive Guide[M]. [s. l.]:O'Reilly Media, Inc.,2009.
- [4] MapReduce[EB/OL]. 2011-12-08. <http://hadoop.apache.org/>.

(上接第 203 页)

Linux 操作系统,该平台可以通过嵌入式 Web 服务器来远程浏览监控网页。实现了现代家居的智能化监控。基本能够满足现代家用的需求,具有广泛的推广和应用价值。

参考文献:

- [1] 刘於勋,李智.基于嵌入式 WebServer 的粮仓温湿度监测系统[J].计算机技术与发展,2009,19(7):213-215.
- [2] 纪金水.基于 ZigBee 无线传感器网络技术的系统设计[J].计算机工程与设计,2007,28(2):404-408.
- [3] 曾桂根,吴霜.基于嵌入式 Linux 的 3G 接入方案的设计与实现[J].计算机技术与发展,2010,20(9):193-196.
- [4] 纪晴,段培永,李连防,等.基于 ZigBee 无线传感器网络的智能家居系统[J].计算机工程与设计,2008,29(12):3064-3067.
- [5] 陈敏.基于嵌入式 Linux 和 GPRS 的数字家庭远程监控系统研究[D].南京:南京理工大学,2011.

hadoop.apache.org/mapreduce/.

- [5] Chu C T, Kim S K, Lin Y A, et al. Map-reduce for machine learning on multicore[M]. [s. l.]:the MIT Press,2007.
- [6] 陈全,邓倩妮.云计算及其关键技术[J].计算机应用,2009,29(9):2562-2566.
- [7] 王德政,申山宏,周宁宁.云计算环境下的数据存储[J].计算机技术与发展,2011,21(4):99-102.
- [8] 陈宇萍.外观设计专利图像检索系统研究[J].科技管理研究,2005(4):162-164.
- [9] 方骥,戴青云.基于图像内容的外观专利自动检索系统[J].计算机工程与应用,2004(34):209-211.
- [10] 邹武,李龙澍,周闪闪.一种基于颜色直方图的图像检索方法[J].计算机技术与发展,2009,19(4):38-40.
- [11] 王贤伟.基于 Hadoop 的外观专利图像检索系统的研究与实现[D].广州:广东工业大学,2011.
- [12] 杨锋,吴华瑞,朱华吉,等.基于 Hadoop 的海量农业数据资源管理平台[J].计算机工程,2011(12):242-245.
- [13] 石柯,徐胜超,唐晓辉,等.一种分布式环境下的新型高性能计算平台[J].小型微型计算机系统,2006,27(9):1782-1787.
- [14] 黄智维,倪子伟.网格计算环境下资源管理的研究[J].计算机技术与发展,2009,19(3):200-203.

- [6] 王小红,周渊,方晓翠.嵌入式视频监控系统的设计和实现[J].通信技术,2011(6):105-109.
- [7] 彭宇,罗清华,潘大为. ZigBee 网络低功耗节点设计[J].仪器仪表学报,2009(6):588-591.
- [8] 南忠良,孙国新.基于 ZigBee 技术的智能家居系统设计[J].电子设计工程,2010,18(7):117-119.
- [9] 王海涛,朱兆优.基于 ZigBee 的 LED 节能街灯控制系统[J].东华理工大学学报,2009,32(2):394-396.
- [10] 孟雷,忽海娜. ARM-Linux 嵌入式系统 BootLoader 的配置与移植[J].计算机技术与发展,2008,18(10):205-206.
- [11] Jahnke J H, d'Entremont M, Stier J. Facilitating the programming of the smart home[J]. IEEE Wireless Communications, 2002,9(6):70-76.
- [12] Furber S. ARM System-on-Chip Architecture[D]. USA: Addison-Wesley Press,2000.
- [13] Henkel J. Selective revealing in open innovation processes: The case of embedded Linux[J]. Research Policy, 2006,35(7):953-969.

基于 Hadoop 的海量图像检索系统

作者: [王梅](#), [朱信忠](#), [赵建民](#), [黄彩锋](#)
作者单位: [浙江师范大学 数理与信息工程学院, 浙江 金华 321004](#)
刊名: [计算机技术与发展](#)
英文刊名: [Computer Technology and Development](#)
年, 卷(期): 2013(1)

本文链接: http://d.g.wanfangdata.com.cn/Periodical_wjfz201301052.aspx