

电力视频大数据分布式检索系统设计与实现

冯亚洲, 岳 东

(南京邮电大学 先进技术研究院, 江苏 南京 210023)

摘 要:随着智能电网的迅速发展和逐渐完善,海量的电力视频大数据每时每刻都在产生,电力行业对视频处理也提出了更高的要求。云计算平台 Hadoop 具有海量数据存储和运算、高可靠性、高拓展性等特点,为解决电力视频大数据的检索问题提供了新的研究思路。在介绍平台环境之后,着重阐述了整个系统的设计与实现。以 Hadoop 大数据平台为基础,将视频文件存储在 HDFS 中,利用 FFmpeg 进行解码,辅之以 OpenCV 函数库进行视频帧的特征提取,使用直方图法在 MapReduce 计算框架上实现关键帧的提取。最后基于 Lucene 实现关键帧的索引,系统将检索结果通过 Web 检索界面呈现给用户。设计并实现的基于 Hadoop 的电力视频大数据检索系统,能够容纳海量电力视频大数据的存储和运算,并且能够实现电力视频大数据的快速检索。

关键词:电力视频大数据;视频检索;Hadoop;OpenCV;Lucene

中图分类号:TP302

文献标识码:A

文章编号:1673-629X(2016)12-0186-04

doi:10.3969/j.issn.1673-629X.2016.12.040

Design and Implementation of Distributed Retrieval System for Massive Power Video

FENG Ya-zhou, YUE Dong

(Institute of Advanced Technology, Nanjing University of Posts and Telecommunications,
Nanjing 210023, China)

Abstract: With the rapid development and gradually perfection of smart grids, massive power video is being produced all the time, and the power industries are also requested higher demand for video processing. Hadoop as a cloud computing platform, has great advantage of mass data storage and computing, high reliability and high expansibility, which provides a new research idea to solve the problem of massive power video retrieval. After introducing platform environment, it mainly focuses on the design and implementation of the whole system. Based on the Hadoop big data platform, the video files is stored in HDFS. The system decodes with FFmpeg and extracts key frames with MapReduce and OpenCV function library. At last, the system presents the retrieval result to users through the Web retrieval interface after the index of key frames based on Lucene. The system introduced can store and operate massive power video, realizing quick retrieval of that.

Key words: massive power video; video retrieval; Hadoop; OpenCV; Lucene

0 引 言

随着计算机技术的迅速发展以及互联网科技的普遍应用,每天都会产生大量的以图片和视频等形式表现的多媒体数据。在电力行业应用中,输变电状态监测、智能营业厅、各类机房等都对视频系统提出了更高要求。传统的检索系统由于缓慢的检索速度、有限的可扩展性、无法实现实时性和较差的稳定性等问题,已经无法满足人们越来越复杂多样的要求^[1]。现有的视

频检索主要借用基于文本数据库的检索方法,检索过程中会消耗大量的 CPU 资源。

云计算具有分布式、并行处理能力,可以将任务分配到各个工作节点同时完成任务,为输电线路电力视频检索提供一种全新的研究思路^[2]。Hadoop 分布式文件系统(Hadoop Distributed File System, HDFS)是一个可扩展的分布式文件存储系统,可以在廉价的普通硬件上运行。尽管大多数的技术人员并不了解系统底

收稿日期:2015-11-10

修回日期:2016-03-16

网络出版时间:2016-11-21

基金项目:国家自然科学基金资助项目(51507084)

作者简介:冯亚洲(1992-),男,硕士研究生,研究方向为电力大数据的视频检索;岳 东,教授,博士生导师,长江学者,研究方向为智能电网大数据分析、复杂系统与多智能体理论等。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20161121.1633.014.html>

层的详细信息,但是 Map/Reduce 函数可以很容易地实现大规模数据的并行处理和计算,具有高可靠性、高扩展性、高效性以及高容错性等优点,在大规模数据处理领域得到了广泛应用。

1 平台环境介绍

1.1 硬件

文中的检索原型开发及运行等都是在大数据实验中心的硬件环境下进行,目前大数据研究中心拥有数十台高档机架式服务器、专业 GPN 图像处理服务器等专业大数据存储和处理设备,构建了基础处理平台以及面向互联网、电力、教育科研、交通安防及过程控制等实验环境。表 1 为大数据研究中心设备数量及配置汇总。

表 1 大数据研究中心设备数量及配置汇总

硬件名称	数量/台	设备配置
NAME 节点	1	2 * E5-2620v2, 128 G 内存, 2 * 4 T 7 200 K SATA 硬盘, 双千兆网口, 冗余电源, RAID 0
DATA 节点	10	2 * E5-2620v2, 64 G 内存, 2 * 4 T 7 200 K SATA 硬盘, 双千兆网口, 冗余电源, RAID 0
管理节点	1	2 * E5-2620v2, 32 G 内存, 4 * 600 G 10 K SATA 硬盘, 双千兆网口, 冗余电源, RAID 5
图形处理服务器	2	2 * E5-2620v2, 128 G 内存, 2 * 300 G 10 K SATA 硬盘, 2 块 NVIDIA Tesla GPU K40 双千兆网口, 冗余电源, RAID 0
防火墙	1	标准 2U 机箱, 单电源, 标准配置 6 个 10/100/1 000 M 自适应电口, 支持一个扩展槽
交换机	2	24 个 10/100/1 000 Base-T 以太网端口; 4 个复用的 1000Base-X 千兆 SFP 端口; 2 个扩展插槽

1.2 软件

OpenCV 是一个用于图像处理、分析、机器视觉方面的开源函数库。采用的编写语言是 C 及 C++, 可以在 Windows、Linux、mac OSX 系统上运行^[3]。因为它更专注于设计成为一种用于实时系统的开源库, 所以该库的所有代码都是经过优化的, 并且计算效率很高。OpenCV 拥有丰富的函数, 可调用的 API(应用程序编程接口)有 500 多个, 在物体识别、图像分割、立体视觉、机器人和运动分析等计算机视觉和图像处理领域应用广泛^[4]。由于机器学习和计算机视觉密切相关, OpenCV 也提供了机器学习库 (Machine Learning Library, MLL)。

OpenCV 主体分为四个模块, 分别包含不同的算法、函数或者执行工具。例如, CV 模块包含了图像处理方法和计算机视觉算法; MLL 作为机器学习库, 包含聚类工具; HighGUI 包含图像和视频读写、处理等函数; CXCore 提供了所有 OpenCV 运行时的一些最基

本的数据结构和出错处理的基本函数^[5]。

Hadoop 分布式计算平台以 HDFS 和 MapReduce (Google MapReduce 的开源实现) 为核心。一个 HDFS 集群拥有一个主节点 (NameNode) 和若干个从节点 (DataNode)。NameNode 作为文件系统的管理者, 管理文件系统的元数据, 包括命名空间、集群配置信息和存储块的复制等; DataNode 是文件存储的基本单元, 周期性地将其存储的 Block 信息发送给 NameNode^[6]。MapReduce 框架采用 Master/Slave 结构, 包含一个 JobTracker 和若干个 TaskTracker。主节点作为任务节点管理调度每一个分配到作业中的任务, 并且能够重新安排之前失败的任务; 从节点作为工作节点完成分配到的任务, 并且 TaskTracker 必须与 DataNode 部署在同一台计算机上^[7-8]。

可以看出, HDFS 和 MapReduce 是 Hadoop 分布式系统体系结构的核心。HDFS 作为存储基础提供海量的数据存储, MapReduce 作为处理引擎对海量数据进行分布式计算处理。MapReduce 在 HDFS 提供的文件操作和存储等支持下进行任务处理, HDFS 在 MapReduce 完成任务的调度、监控、执行等工作过程中构建分布式系统的基础, 二者相互作用, 完成了 Hadoop 分布式集群的主要任务, 将系统底层细节透明的分布式基础架构呈现在用户眼前^[9]。

2 原型系统关键技术设计

2.1 关键帧提取

视频由于其本身“非结构化”的特点, 在对其进行分析和检索之前, 首先要对其进行结构层次的描述和组织, 将其划分成相互独立的视频片段并选取合适的帧。视频结构化就是把一个连续视频流按照其内容展开的不同, 分成若干语义段落单元^[10]。视频的层次结构自顶而下分成视频 (Video)、场景 (Scene)、镜头 (Shot) 以及帧 (Frame), 其粒度越来越精细^[11]。

该系统采用 X^2 直方图法提取关键帧, 用式 (1) 计算两视频帧图像间的距离:

$$d(I_i, I_j) = \sum_{k=1}^n \frac{(H_i(K) - H_j(K))^2}{H_j(K)}$$

(1)

$p(k)$ 表示随机变量 k 的概率密度函数, 对于数字图像, k 可以是灰度级数、区域灰度和梯度等特征。当选用颜色直方图特征时, $H(K) = - \sum_{i=0}^{L-1} p(k_i) \log p(k_i)$ 表示量化后图像信号的熵, $d(I_i, I_j)$ 表示两视频帧之间的距离, I_i, I_j 表示两幅不同的视频帧。首先设定一个阈值 T , 当计算出来的视频帧间距离大于 T 时, 认为该帧能够反映该视频镜头的主要内容, 提取出来作为视频关键帧; 如果小于 T 则依次继续执行, 直到最后。

2.2 基于 Hadoop 平台分布式关键帧提取

在 Hadoop 平台中,HDFS 作为分布式文件系统可以存储任何种类的大数据,而 Map/Reduce 计算框架可以满足并行运算的要求。因此文中采用以 HDFS 为基础的 video 大数据存储平台,利用 Map/Reduce 计算框架进行视频关键帧提取。

提取技术方案如下所述:

HDFS 存储视频数据:HDFS 采用冗余备份的策略保证了分布式文件系统的安全性,非常适合大数据文件的存储^[12]。数据在上传到 HDFS 的过程中,HDFS 会根据用户置顶的 Block 的大小(默认 64 M)自动对其进行分割处理。在均衡整个系统负载的情况下,均匀地将数据分布存储到集群中的 DataNode。

分布式关键帧提取:利用 FFmpeg 第三方解码库^[13],实现视频文件的分割,进而将分割后的文件传送到多台计算节点上进行分布式的关键帧提取,最后将提取到的关键帧存储在 HDFS 上。

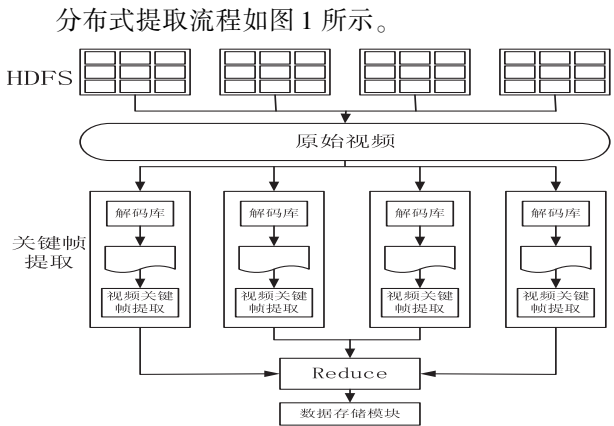


图 1 分布式视频关键帧提取

2.3 基于 Lucene 的视频检索

Lucene 是一个高效、技术成熟的全文检索工具包,提供完整的查询引擎和索引引擎。具有如下几个优点:

- (1) Lucene 定义了一个特殊的索引文件格式,该格式独立于应用平台。
- (2) 设计了独立于语言和文件格式的文本分析接口,用户只需要实现文本分析的接口就可以扩展新的语言和文件格式。
- (3) Lucene 的设计者已经默认实现了一套强大的查询引擎,用户不需要编写大段代码就可以获得 Lucene 的搜索能力^[14]。

优秀的面向对象设计的系统框架,而且独立于应用系统的特点大大降低了 Lucene 的耦合度,可以使用户方便地扩展新功能,很轻松地将 Lucene 嵌入到实际应用中以实现全文检索功能。

Lucene 系统结构组织图及各个模块所属的系

统部分如图 2 所示。Lucene 将所有的源代码分成 7 个主要模块。其中,索引核心主要用于操作索引的创建与维护^[15]。

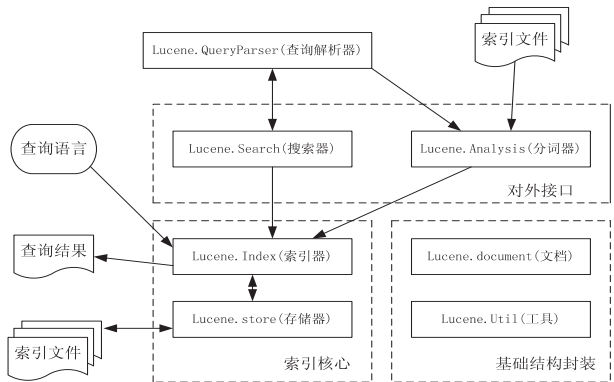


图 2 Lucene 系统的结构组织图

3 原型系统实现

3.1 原型系统框架图



图 3 原型系统框架图

系统是在 Hadoop 平台的支持下完成,HDFS 平台在存储海量视频数据中起到了重要作用。大量的原始视频数据存储在 HDFS 平台上,构成视频检索的视频库,用户搜索的视频都在这个视频库中产生。

有了原始数据,系统将对其进行处理。系统将原始视频分成不同的视频片段,并且利用 OpenCV 视觉库和直方图算法提取出能代表该视频的关键帧,并存储在 HDFS 平台上。关键帧的提取将直接关系到用户检索的质量,因此这一步相当重要,这也是整个系统最关键的一步。视频关键帧即一幅幅图片,而且相对于海量的原始数据,关键帧大大提高了检索效率,且检索精度也令人满意。

存储在 HDFS 平台上的关键帧既与原始视频片段形成一一对应的关系,又在时刻等待着示例图片的输入,等待着被索引。关键帧索引的方法有很多,该系统利用 LIRE 图像检索系统建立视频关键帧索引,用户通过该索引搜索相似的图像,从而定位到相似图片的对应视频。

3.2 Web 检索界面介绍

首先是系统首页,“以图搜视频”是检索系统名

称。从这个首页及名称可知,要完成视频检索,用户需要提供示例图片,这对提高检索效率和检索精度都起到了至关重要的作用。在这个界面,用户上传示例图片,系统将搜索包含与示例图片类似、相像的视频帧的视频片段。

系统的第二个界面是搜索结果返回页,如图 4 所示。该界面除了显示用户上传的示例图片,大部分界面显示与用户上传示例图片相关的视频关键帧。这些关键帧按照与示例图片相关程度由高到低的顺序排列,在本页用户就可以观察出哪些关键帧符合用户的主观意志,即哪段视频是用户想要得到的。

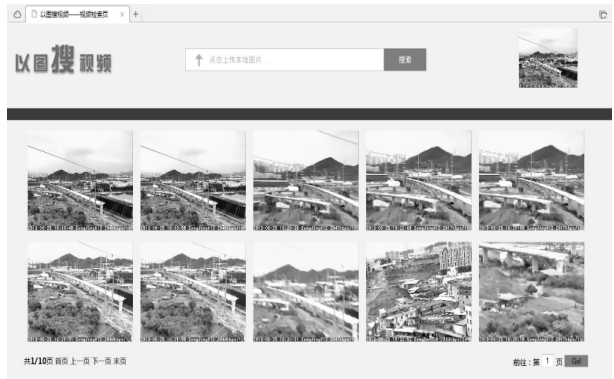


图 4 检索结果页

当用户选择了与输入示例图片最为相似的视频关键帧,系统将显示第三个界面—视频展示页,即包含用户选择的视频关键帧的视频片段。这个界面包含了与用户输入示例图像相关的视频信息,包括该段视频的 HDFS 文件路径,并且用户可以直接从这个界面下载得到与输入图像相关的视频片段,如图 5 所示。

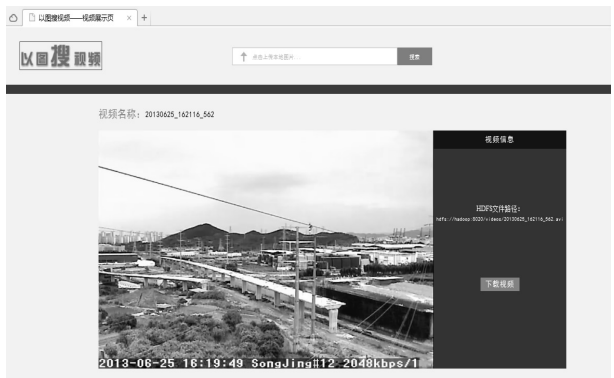


图 5 视频展示页

4 结束语

传统的视频检索方法已经不能满足电力行业对电力视频大数据检索的需求。文中设计并实现的电力数据环境下的视频检索原型系统,以 Hadoop 大数据平台

为基础,充分利用 HDFS 分布式存储以及 MapReduce 并行计算框架的特点完成检索任务。实验结果表明,系统能够实现电力视频大数据的快速检索,并且检索质量较高,能够达到传统检索方法无法达到的效果。为了更好地实现检索系统的高效性、实时性,在以后的工作中,需要从 Hadoop 大数据平台运行机制出发,重点解决如何更加有效地提高检索效率的问题。

参考文献:

[1] 王 梅,朱信忠,赵建民,等. 基于 Hadoop 的海量图像检索系统[J]. 计算机技术与发展,2013,23(1):204-208.

[2] 范 敏,徐胜才. 基于 Hadoop 的海量医学图像检索系统[J]. 计算机应用,2013,33(12):3345-3349.

[3] 刘永勤,袁 卫. 基于 CortexM3 自动追踪系统的实现[J]. 计算机与数字工程,2014,42(6):1068-1070.

[4] 王 燕,曹银杰,郎丰法,等. 基于 Emgu CV 的数字相机图像采集[J]. 电子科技,2012,25(4):31-32.

[5] Bradski G,Kaehler A. 学习 OpenCV[M]. 于仕琪,刘瑞祯,译. 北京:科学出版社,2008.

[6] 万川梅,张 莉. 大数据存储技术标准化的探讨[J]. 数字技术与应用,2014(1):222.

[7] Zhang J,Liu X L,Luo J W,et al. DIRS:distributed image retrieval system based on MapReduce[C]//Proceedings of the 5th international conference on pervasive computing and applications. Piscataway:IEEE,2010:93-98.

[8] Pal A,Agrawal P,Jain K,et al. A performance analysis of MapReduce task with large number of files dataset in big data using Hadoop[C]//2014 fourth international conference on communication systems and network technologies. [s. l.]:IEEE,2014:587-591.

[9] 张学亮,陈金勇,陈 勇. 基于 Hadoop 云计算平台的海量文本处理研究[J]. 无线电通信技术,2014,40(1):54-57.

[10] 雷少帅. 基于内容的视频检索关键技术研究[D]. 太原:太原理工大学,2012.

[11] Sze K W,Lam K M,Qiu G P. A new key frame representation for video segment retrieval[J]. IEEE Transactions on Circuits and Systems for Video Technology,2005,15(9):1148-1155.

[12] 金松昌. 基于 HDFS 的多用户并行文件 IO 的设计与实现[D]. 长沙:国防科学技术大学,2010.

[13] Lei Xiaohua,Jiang Xiuhua,Wang Caihong. Design and implementation of a real-time video stream analysis system based on FFMPEG[C]//2013 fourth world congress on software engineering. [s. l.]:IEEE,2013:212-216.

[14] 张建军,王剑霞. 浅谈 Lucene 在号百搜索引擎系统中的集成[J]. 科技资讯,2012(21):12.

[15] 李永春,丁华福. Lucene 的全文检索的研究与应用[J]. 计算机技术与发展,2010,20(2):12-15.