



基于 Hadoop 云计算平台的图像分类与标注^{*}

陆寄远¹, 黄承慧¹, 侯 昉¹, 李 斌²

(1. 广东金融学院计算机科学与技术系 广州 510521;

2. 甲骨文研究开发中心(深圳)有限公司 深圳 518057)

摘 要: 为有效处理并利用互联网海量的图像和视频数据, 提出了一种基于 Hadoop 云平台的图像分类和标注解决方案。针对如何高效地进行训练集提取这一重要问题, 搭建了基于云计算的图像抓取平台, 利用互联网的图像资源作为原始数据集, 为提取训练集图像提供足够的数据; 实现了基于概率潜在语义分析模型的训练集图像提取功能, 对原始数据集进行基于主题的聚类, 帮助用户快速选取训练集图像; 加入了 SVM 分类模型, 利用提取出来的训练集对未标注图像进行分类标注, 实现了完整的系统。实验结果表明, 该方案能够满足海量图像数据分类和标注的功能和性能需求。

关键词: 云计算; 训练集提取; 支持向量机; 视觉特征提取

doi: 10.3969/j.issn.1000-0801.2014.02.008

Image Classification and Annotation Based on Hadoop Cloud Computing Platform

Lu Jiyan¹, Huang Chenghui¹, Hou Fang¹, Li Bin²

(1. Department of Computer Science and Technology, Guangdong University of Finance, Guangzhou 510521, China;

2. Oracle Research and Development Center (Shenzhen) Co., Ltd., Shenzhen 518057, China)

Abstract: In order to effectively deal with the massive image and video data in internet, a solution for image classification and annotation based on Hadoop cloud platform was proposed. Firstly, a system based on cloud computing was given to crawl the raw data image from WWW. Secondly, a training image extractor based on pLSA (probabilistic latent semantic analysis) was presented to help users to get training set images effectively. Thirdly, the SVM model for the system was integrated, which could be used to classify or annotate novel images. According to the experiment, the system can both meet the functional requirement and achieve optimized performance for image data classification and annotation.

Key words: cloud computing, training set extraction, support vector machine, visual feature extraction

1 引言

随着网络的普及以及多媒体数据获取设备的发展, 图像和视频的数量都飞速增长, 多媒体资料的存储和检索成为热门的研究领域。基于内容的图像检索、对象识别、标注

等都是现在的研究重点。各种分类算法、模型和系统不断涌现, 如基于 SVM (support vector machine, 支持向量机)、pLSA (probabilistic latent semantic analysis, 概率潜在语义分析) 或决策树的图像分类方法^[1]。其中, 基于内容的图像检索是形成基于内容的标注和图像之间的映射。有了这些

^{*} 国家自然科学基金资助项目 (No. 61173081), 广东省自然科学基金资助项目 (No. S2011020001215, No. S2012040007847)



标注,就可以将用户的查询分解到标注的概念,以检索出结果。

图像分类和标注的问题可以理解为模式识别的问题,计算机无法像人类一样具有抽象概括的能力,只能利用图像的底层特征进行识别分类。现有的图像分类系统大部分是按照如图 1 所示的工作流程:从图像中提取视觉描述子向量;利用已经学习到的码本(codebook)对这些描述子向量进行编码,使得相似的描述子向量得到相近的标签;根据每个标签的出现频率统计出图像内容的全局直方图,得到图像的视觉特征表示;将直方图导入分类模型中,估计该图像类别标签。

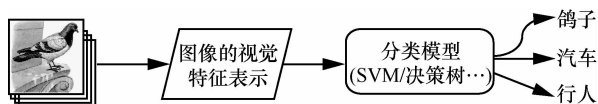


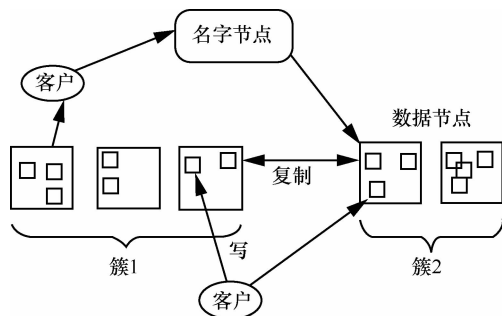
图 1 图像分类系统的一般工作流程

目前已经有两类图像检索系统:基于文本的图像检索(text based image retrieval, TBIR)系统和基于内容的图像检索(content based image retrieval, CBIR)系统。在基于文本的系统中,图像要进行人工标注,然后通过这些标注信息进行检索。然而,在 CBIR 中是由图像的视觉特征(如颜色、纹理、形状等)建立索引。很多人在该领域中取得了出色的研究成果,现在已经可以使用的 CBIR 系统有 QBIC、Informedia-II、ALIPR、GazoPa 等^[23]。随着研究的深入,人们发现 CBIR 系统存在两个明显的不足:低层可视特征和高层语义概念之间存在巨大鸿沟;与人类视觉机制具有明显的不一致性。为了缩小图像底层特征和用户检索语义概念之间的“语义鸿沟”,部分研究者开始进行语义图像检索的研究。微软亚洲研究院开发了一个 Web 图像检索系统^[4],目的是将传统 Web 图像检索返回的结果重新进行聚类。搜索的结果被聚类成不同的语义类别,对每个类别,都会选出几张代表性的图片,使用户能够马上了解到这一类的主题。每一类里面的图片则根据它们的视觉特征进行组织,使其显示结果更符合用户的需求。其中,训练集的大小和质量是影响分类效果的重要因素,现在基本使用人工收集训练集,这是件复杂的工作,当需要分类的对象越多,要收集的训练集越大,消耗的人力也越多时。本文的重点是要解决该问题,高效地获取训练分类模型所需的训练集图像。本文所提出的解决方案为:一是依靠云计算技术解决训练集大小的问题,二是通过 pLSA 主题聚类的方式实现人机交互的训练集选取,从而提高效率。

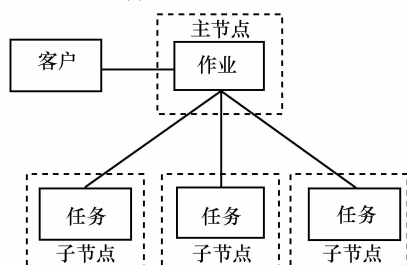
2 基于 Hadoop 的图像分类与标注系统

云计算是一种新的 IT 资源提供模式,依靠强大的分布式计算能力,使成千上万的终端用户能够依靠网络连接的硬件平台的计算能力实施多种应用。Hadoop^[5]是一个分布式系统基础架构,由 Apache 基金会开发。使用者可以在不了解分布式底层细节的情况下,搭建分布式计算平台。Hadoop 的核心组件有两个:Hadoop 分布式文件系统(HDFS)和 MapReduce,如图 2 所示。HDFS 是一个隐藏下层负载均衡、冗余复制等细节的分布式文件系统,适合部署在廉价的机器上。它能提供高吞吐量的数据访问,非常适合大规模数据集上的应用,并对上层程序提供一个统一的文件系统 API(应用程序接口)。从图 2(a)可以知道,HDFS 只有一个名字节点,负责管理元数据操作和控制数据块的放置,由数据节点实际保存数据块。另外,MapReduce 代表了 map 和 reduce 两种操作,如图 2(b)。大多数分布式运算可以抽象为 MapReduce 操作。map 是把输入分解成中间的 key/value 对,reduce 把 key/value 合成最终输出。这两个函数由程序员提供给系统,下层设施把 map 和 reduce 操作分布在集群上运行,并把结果存储在分布式文件系统上。从图 2(b)可以知道,用户提交 MapReduce 任务给主节点,JobTracker 负责将任务分配到各个子节点上,实现并行处理。

考虑到 Hadoop 的开源以及容易进行开发的特点,同



(a) HDFS 的架构



(b) MapReduce 框架

图 2 Hadoop 的核心架构

时为了保证平台的稳定性,本系统以 Hadoop + Ubuntu 的方式进行构建。Ubuntu 是一个完全以 Linux 为基础的操作系统,可自由地获得,并提供社区和专业的支持。本系统的框架设计如图 3(a)所示。首先,利用云技术发挥互联网中多台硬件的计算能力,加快图像抓取的速度,并获取所需要的原始图像集。其次,当抓取到足够的原始数据集以后,使用训练集提取模块通过交互的方式帮助用户选取恰当的训练集。第三,通过分类器学习模块训练分类器。最后,分类标注模块利用这些分类器对新图像进行分类标注。系统的硬件拓扑,如图 3(b)所示。从图中可以看出,图像抓取平台中有一个主节点,抓取任务通过安全外壳(SSH)协议提交到主节点,由该节点负责将任务分配到所有的子节点。

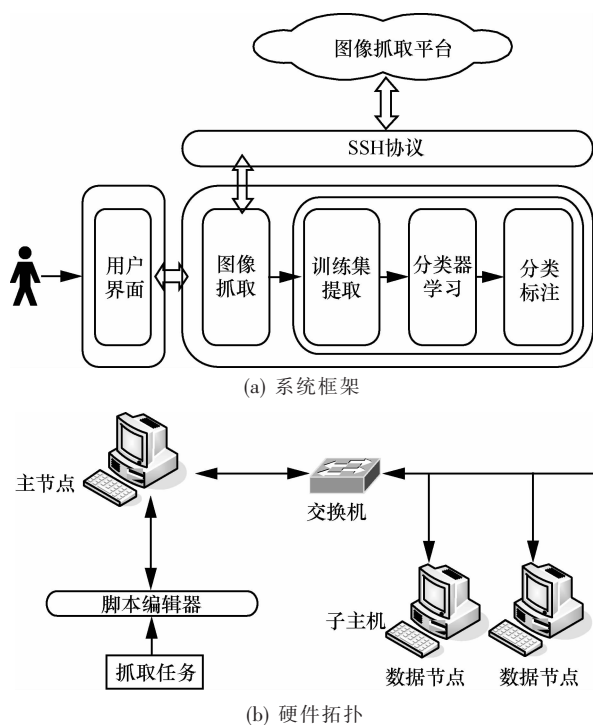


图 3 基于 Hadoop+Ubuntu 的系统架构

从图 3(a)可以知道,分类标注系统要满足 4 个功能需求:原始图像抓取、训练集提取、分类器模型学习和分类标注。用户通过训练集提取可以从原始数据集中生成训练集,然后学习分类器模型,利用分类器模型进行图像的分类和标注,并对分类标注结果进行存储,以供用户查询或者检索系统。本节将介绍分类标注系统功能模块设计。结合系统的架构,把系统分为图像抓取模块、训练集提取模块、分类器学习模块、分类标注模块,如图 4 所示。

其中,图像抓取模块负责将用户的抓取任务提交到图

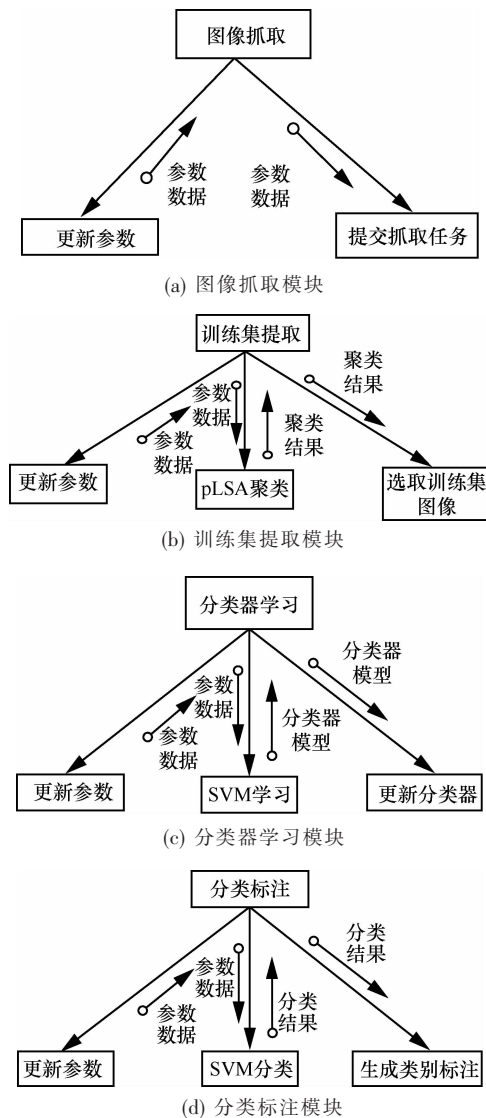


图 4 分类标注系统的模块结构

像抓取平台,通过 SSH 协议连接到云计算平台的主节点,从互联网中抓取所需的原始图像集;训练集提取模块负责对原始图像集进行基于 pLSA 模型的主题聚类分析,通过用户交互的形式选取出训练集图片;分类器学习模块的任务是根据用户提供的训练集图像学习分类器模型,并保存为分类器模型文件;分类标注模块完成对图像或者图像序列进行分类标注的任务,并生成分类标注文件。

训练集提取模块分为 3 个子模块:更新参数模块、pLSA 聚类模块和选取训练集模块。用户通过更新参数模块设置训练集提取任务的参数,pLSA 聚类模块根据用户的参数设置进行原始图像的主题聚类分析,完成后用户可以通过训练集选取模块选取所需的训练集图片。

分类器学习模块进一步分为更新参数、SVM 学习和更



新分类器模型 3 个子模块。用户通过更新参数模块设置分类器学习任务的参数,SVM 学习模块根据用户的参数设置从训练集中学习分类器模型,运行成功后更新分类器模块负责存储更新分类器模型。

分类标注模块分为更新参数、SVM 分类和生成类别标注 3 个子模块。用户通过更新参数设置模块设置任务的参数,SVM 分类模块根据用户的参数设置对图像 (或图像序列)进行分类标注,并通过生成类别标注模块生成类别标注文件。

3 实验结果

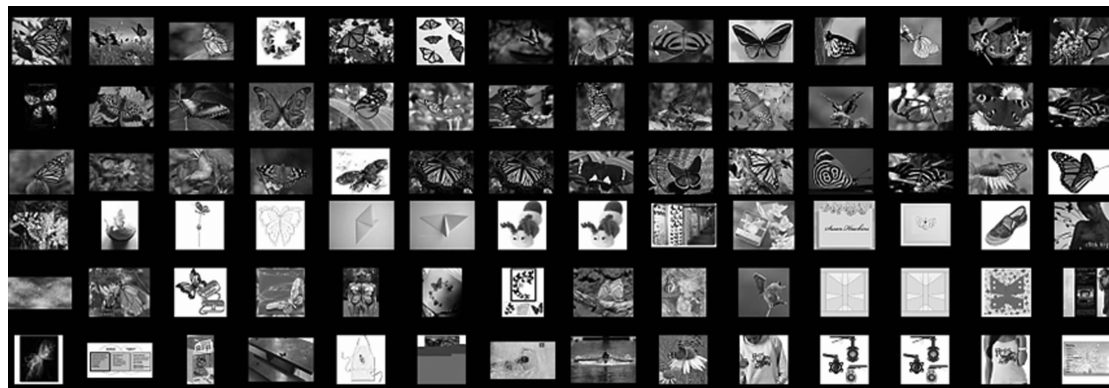
系统使用操作系统平台为 Ubuntu Desktop 9.10,分布

式系统平台为 Hadoop 0.19.2。图像抓取模块主要用 Java 开发,开发工具为 Eclipse,运行环境为 sun-6-jdk、sun-6jre。选取 2 组测试集,第一组是“Caltech-256 object category dataset”^[6],第二组是利用网络抓取平台抓取的原始数据集,如图 5 所示。通过实验发现,对原始图像集进行 10 个主题的聚类,取得了较好的效果。下面将展示摩托车、蝴蝶测试的结果。

利用训练集提取模块对原始图像集进行 10 个主题的聚类,结果见表 1。从表 1 中可以看出,大部分聚类还是比较理想,可以直接去掉如聚类 1、聚类 3、聚类 4、聚类 7、聚类 8、聚类 10,从而实现了“按类选取”,加快了训练集的筛选速度。



(a) 摩托车图像集



(b) 蝴蝶图像集

图 5 原始图像集

表 1 摩托车和蝴蝶的 10 个主题聚类测试结果

图像集	测试结果	聚类 1	聚类 2	聚类 3	聚类 4	聚类 5	聚类 6	聚类 7	聚类 8	聚类 9	聚类 10
摩托车	符合	44	62	11	6	38	14	1	0	68	5
	不符合	115	43	131	32	27	57	74	26	20	75
蝴蝶	符合	11	1	3	2	10	16	7	45	24	45
	不符合	8	50	40	54	21	3	39	29	13	39

4 结束语

随着网络的普及和多媒体数据获取设备的发展,图像和视频的数据量都在飞速增长,多媒体资料的存储索引成为热门的研究领域。本文针对当前图像分类技术中都要面对的提取训练集问题,提出了一个基于 Hadoop 云平台的解决方案。该方案基于现有文本图像搜索引擎的图像抓取器,实现基于云计算的图像抓取平台,并在此基础上利用 pLSA 模型,采用 MSER 和 STAR 区域特征,实现了主题聚类的训练集提取,同时通过对原始图像集进行基于主题的聚类,使得用户可以“按类”筛选训练集,比“逐张”筛选要有效率得多。

参考文献

1 Moosmann F, Nowak E, Jurie F. Randomized clustering forests for image classification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 30(9): 1632~1646

2 ALIPR. Automatic photo tagging and visual image search. <http://www.alipr.com/>, 2012

3 GazoPa. Similar image search. <http://www.gazopa.com/>, 2010

4 Cai D, He X F, Ma W Y, et al. Organizing www images based on the analysis of page layout and web link structure. Proceedings of 2004 IEEE International Conference on

Multimedia and Expo, ICME'04, Sorrento, Italy, 2004

5 Grangier D, Bengio S. A discriminative kernel-based approach to rank images from text queries. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 30 (8):1371~1384

6 Griffin G, Holub A, Perona P. Caltech-256 Object Category Dataset. Technical Report, California Institute of Technology, 2007

[作者简介]



陆寄远,男,博士,广东金融学院计算机科学与技术系系统分析师,主要研究方向为视频编码、图像处理。

黄承慧,男,博士,广东金融学院计算机科学与技术系讲师,主要研究方向为视频信息检索、多媒体数据挖掘。

侯昉,男,博士,广东金融学院计算机科学与技术系讲师,主要研究方向为视频网络、图像通信。

李斌,男,甲骨文研究开发中心(深圳)有限公司高级软件工程师,主要研究方向为图像处理。

(收稿日期:2013-08-23)

· 简讯 ·

2013 年 12 月通信业主要指标完成情况(一)

指标名称	单位	本年本月止累计到达	比上年同期累计	本月
电信营业收入	亿元	13 494.0	12.0%	1 177.4
其中:电信主营业务收入	亿元	11 689.1	8.7%	1 014.8
电信固定资产投资完成额	亿元	3 754.7	3.9%	1 111.7
固定本地电话通话时长合计	万分钟	30 230 919.0	-15.2%	2 482 940.0
区间电话通话时长	万分钟	3 174 731.7	-20.0%	259 301.2
区内电话及拨号上网通话时长	万分钟	27 056 187.2	-14.6%	2 223 638.8
固定长途电话通话时长合计	万分钟	5 905 388.3	-15.7%	496 858.8
国内长途电话通话时长	万分钟	5 771 161.5	-15.7%	487 057.2
国际电话通话时长	万分钟	81 947.3	-18.3%	5 980.9
港澳台电话通话时长	万分钟	52 279.5	-18.3%	3 820.8
移动电话通话时长合计(含本地)	万分钟	289 877 348.2	5.0%	24 877 969.6
移动电话国内长途通话时长	万分钟	67 290 406.3	5.8%	5 861 657.8
移动电话国际电话通话时长	万分钟	130 949.3	7.9%	10 881.3
移动电话港澳台电话通话时长	万分钟	97 145.9	-3.5%	8 374.3
移动短信业务量	万条	89 166 641.3	-0.6%	6 021 925.8
移动互联网接入流量	万吉字节	132 138.1	71.3%	13 494.4

注:(1)收入增长率按可比口径计算。(2)固定长途电话通话时长和移动电话通话时长均包含相应的 IP 电话通话时长。(3)通话时长各项指标均为去话通话时长。

(来源:工业和信息化部运行监测协调局)