基于移动平台的图像检索系统

刘强强,余黎青,赵 鹏,刘慧婷
(安徽大学 计算机科学与技术学院,安徽 合肥 230601)

摘 要:近年来移动终端的普及促进了移动平台上图像检索技术的发展。当用户看到感兴趣的商品的时候,他们希望能够使用终端拍下来,然后进行商品的检索并返回一些推荐的商家。为了解决这个问题,面向移动平台,构建了一个图像检索系统,通过手机等移动终端,拍摄或传输图片来检索互联网上相关的图片和信息。该系统构建了一个爬虫系统用来采集图片信息,在安卓平台上直接进行图像特征提取,通过移动终端拍摄的商品图像搜索互联网图像,返回相关网店链接并进行相关商品推荐。该系统对 120 万幅图片采用位置敏感哈希索引、存储和检索,既保证了结果在较小的误差范围内,也极大地降低了时间复杂度。最后用户可以根据推荐的链接进行选购。实验结果表明,该系统能够满足用户的需求,并且具有很强的实用性。

关键词:特征提取;图像检索;图像搜索引擎;爬虫系统
中图分类号:TP391.41 **文献标识码**:A **文章编号**:1673-629X(2016)11-0010-04
doi:10.3969/j.issn.1673-629X.2016.11.003

A Novel Image Retrieval System Based on Mobile Platform

LIU Qiang-qiang, YU Li-qing, ZHAO Peng, LIU Hui-ting
(School of Computer Science and Technology, Anhui University, Hefei 230601, China)

Abstract: The popularity of mobile device in recent years promotes the development of image retrieval technology in mobile platform. When users see some interesting commodities, they hope to take a photo about them and retrieve them to get some recommended shops. In order to solve this problem, a novel image retrieval system is built facing mobile platform. By taking or transferring picture with mobile platform like mobile phones, the relevant web or information including the similar picture with the proposed system is retrieved. In the proposed system, the image features are extracted on the Android platform. The web links including the similar image are returned, and recommends are given. In order to collect images, a crawling system is established. For 1.2 million images, the proposed system adopts Location Sensitive Hash to index and store. The proposed system not only promotes the retrieval performance, but also reduces the time complexity greatly. Users can buy commodities according to the recommended links. The experimental results show the system can meet users' needs and has very strong practicalities.

Key words: feature extraction; image retrieval; image search engine; spider system

0 引 言

现今,图像视频等多媒体信息呈现出爆发式增长态势。快速和精确地找到用户所需查询的图像成为了各个搜索引擎的重要研究领域之一[1]。

智能移动终端的普及,用户检索需求日趋多样化,面向移动平台的图像检索使用户能随时随地检索信息[2]。文中在移动平台 Android 系统上,融合几种全局特征作为实验特征,并将特征提取移植到移动平台上,开发了一个性能良好的爬虫系统来采集数据,采用位置敏感哈希(Location Sensitive Hash, LSH)建立索引,构建一个基于移动平台的图像检索系统[3]。

1 相关技术

1.1 提取全局特征和评估其相似度

图像的特征提取和特征匹配是图像检索的关键步骤[4]。文中提取图片的纹理特征和颜色边缘方向特

收稿日期:2016-01-20 **修回日期**:2016-05-11 **网络出版时间**:2016-10-24
基金项目:国家自然科学基金资助项目(61202227);安徽省自然科学基金(1408085MF122,1508085MF127);安徽大学大学生科研训练计划项目(J10118520149);安徽大学信息保障技术协同创新中心公开招标课题(ADXXBZ2014-5,ADXXBZ2014-6)
作者简介:刘强强(1993-),男,研究方向为智能信息处理、多媒体信息检索;赵 鹏,副教授,研究方向为智能信息处理、机器学习;刘慧婷,副教授,研究方向为数据挖掘。

征。目前采用的颜色模型有很多,包括 HSV、RGB、HIS、CMY、YUV 等等。目前大部分图像是用 RGB 模型进行压缩存储的。但是 RGB 模型与色调、饱和度和亮度等都没有直接联系,所以一般在提取特征时会将其转换到其他颜色空间。颜色布局特征(Color Layout Feature)是 MPEG-7 中提到的一种颜色描述符,以一种非常紧凑的形式,能有效地表达图像颜色空间分布信息[5-6]。

1.2　网络爬虫系统的设计

网络爬虫是一个自动下载网页的程序,它根据抓取目标,有选择地访问万维网上的网页与相关的链接,获取所需信息。当网络爬虫程序通过打开某个 HTML 页面时,它会分析 HTML 表及结构来获取信息,并指向其他页面的超链接,然后通过既定的搜索策略选择下一个要访问的站点[7]。影响爬虫效率的因素有很多,其关键因素如下:

(1)URL 去重问题。

URL 通常是一串很长的字符串,当采集的数据中含有大量超链接时,URL 去重是确保新加入的 URL 没有重复。为了减少索引文件的存储空间,不能将 URL 作为索引字段,文中采用将 URL 编码为一个整数作为其索引字段,即 URL 哈希。文字采用 CRC32 校验码算法将数据进行哈希[8],然后将数据库表进行拆分,拆分之后需对数据库表的查询操作采用联合查询。文中先利用 MD5 算法加密 URL,获得校验字符串,取其第一位(十六进制表示),即将原先的表拆分成 16 个。在查询或者录入 URL 之前,先按照上述方法处理后,根据计算出的第一位,获得相应的查询子表,由此可以将数据表分散到各个服务器中,提升查询效率。

(2)磁盘网络以及 CPU 的瓶颈问题。

网络爬虫最主要的效率瓶颈在于网络带宽利用率低、适应性差;功能模块设计不够完善;各功能模块协同工作效率欠缺等。目前主流爬虫系统采用并发工作流的设计,以充分利用网络带宽。由于基于进程的并发代价较基于线程的并发而言相对较高,故大部分网络爬虫都是多线程架构设计[9]。

(3)采集中断的恢复问题。

由于不可预料的事件发生,导致网络中断或者退出爬虫系统,系统必须具有中断恢复能力,继续上一次的采集。

文中根据具体实验室网络环境、CPU、磁盘、内存的情况,经过实验分析发现采用 4 个爬虫同时进行采集时速度最快。为了实现多线程的并发[10],设立一个固定大小的缓冲区,在实验中性能较好的缓冲区大小为 5 000 ~ 20 000。同时采用同步锁实现缓冲区的并发存取。

1.3　索引的建立

文中采用位置敏感哈希[9],具体方案如下:

(1)将每个取到的哈希特征的每一维作为一个关键词加到文档中,所有的同一特征取得的哈希值属于同一域,然后进行文档倒排索引。

(2)查询的时候,先计算查询图片的特征,然后构造 N 个哈希函数,映射成 N 维向量之后,将其作为一组具有 150 个“关键字”的“句子”进行查询。打分机制使用的是 TF-IDF,取得打分靠前的结果,之后进行特征重排[11]。

相似的特征通过位置敏感哈希之后映射到同一个桶的概率 p_1 比映射到不同的桶的概率 p_2 要大。文中使用了 150 个哈希函数,容错率远远大于单个哈希。

文中对比了 LSH 和线性搜索的效率,其结果如表 1 所示。

表 1　线性搜索和基于 LSH 搜索的效率对比

数据集数量/万张	线性搜索消耗时间/ms	LSH 消耗时间/ms
5	$1.987 * 10^3$	31
10	$3.833 * 10^3$	46
20	$7.674 * 10^3$	94
40	$15.241 * 10^3$	172
80	$32.576 * 10^3$	325
120	$56.896 * 10^3$	435

2　基于移动平台上的图像检索系统

文中系统的总体框架如图 1 所示。客户端发送各种请求,服务器接收到请求之后开始处理,如果是查询图像,则进行相应的查询操作,之后封装数据返回给程序[12]。

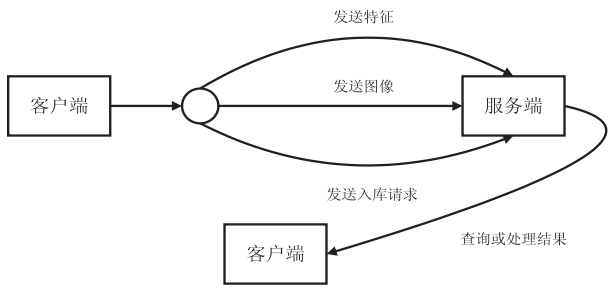

图 1　系统总体框架图

客户端基于 Android 系统,其功能结构如图 2 所示。首先客户端启动先要求用户选择是用已经存在的图片还是使用拍摄的图像。两个选择都可以得到一个图像,然后客户端选择是否利用 Grabcut 算法来提取图像中的主要部分[13-14],之后对图像按照用户选择的特征进行特征提取,当然用户可以直接选择发送图片而不提取特征来节省客户端的计算时间。当服务器接收到查询请求后,如果该查询是直接用特征进行查询

的,则在服务端计算特征的 LSH 值,然后查询之后将结果返回给客户端;如果查询的是图像,则直接调用图像特征提取接口,然后将其封装成文档 Document,再进行查询,接着返回查询结果。客户端在接收到结果之后,如果用户对查询结果满意,则可以进行下次查询或者直接退出程序;如果不满意,用户可以填写相关的图像描述信息,之后将其发送到服务器,服务器按获取到的信息将其存入索引中。

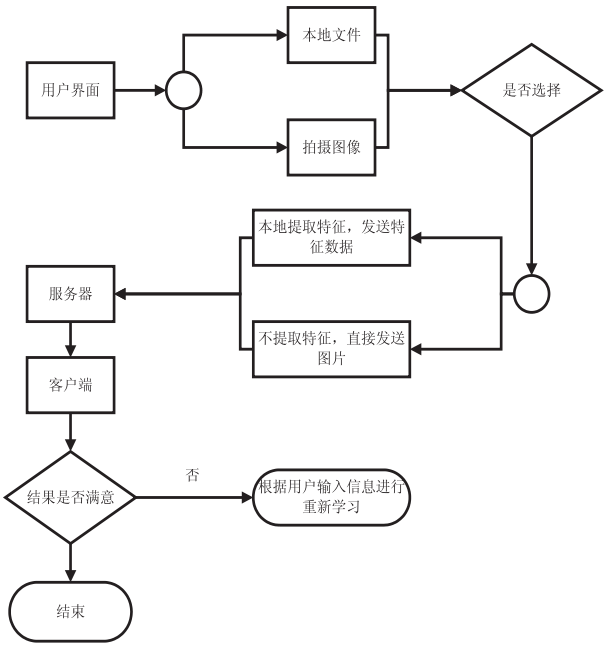

图 2 客户端功能流程图

文中采用 C/S(客户端/服务器)模式。服务端采用 Java 语言编写,数据库使用的是 Mysql 数据库,客户端的实现需要安装 Android SDK, Opencv4Android, NDK, Cygwin。文中将特征提取算法从 PC 平台移植到 Android 平台。而在客户端,引入了 Grabcut 图像分割算法,提取图像中的主要部分。

图 3 为系统移动终端手机上的主界面。

图 3 手机端主界面

图 4 为设置界面,包括设置提取哪种图像特征,以及选择是在手机端提取图像特征发给服务器,还是直接将图片发给服务器。

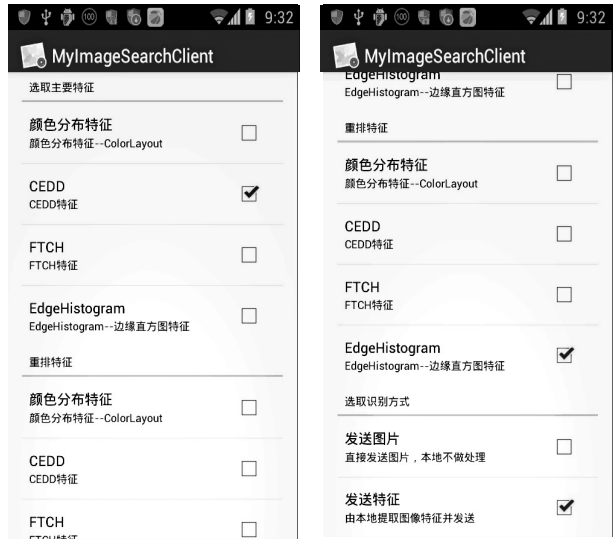

图 4 手机端设置界面

图 5 为客户端图像处理和识别的界面。

图 5 客户端图像处理、识别界面

图 6 为服务器返回结果显示在客户端的界面。

3 结束语

文中设计并实现了一个基于移动平台的图像检索系统,该系统包括两大模块(移动客户端和服务端),是一种典型的 C/S 构架。移动端实现了特征提取算法,并使用 OpenCV 提供的 Grabcut 算法,使得搜索方式更加丰富。而服务器分成了好几个模块,每个模块都有明确的分工:数据采集模块,此模块是检索系统必备的一个模块,检索系统需要大量的数据作为支撑,而大量的数据,人工采集基本不可能完成,数据采集模块旨在构建一个自动采集数据的程序,能够高效不间断地采集资源。文中采用了数据库分表和哈希的技术,以及数据库缓存、多线程并发等手段来提高采集效率。特征提取和索引模块,是将采集到的数据进行进一步

图 6　结果展示界面

的处理,如图像检索系统需要对采集来的图像进行特征提取,并存储图像的描述信息,以及能找到该图像的网址等。

参考文献:

[1] Gudivada V N, Raghavan J V. Special issue on content-based image retrieval systems[J]. IEEE Computer Magazine, 1995, 28(9): 119-120.

[2] Deselaers T, Keysers D, Ney H. Features for image retrieval: an experimental comparison[J]. Information Retrieval, 2008, 11(2): 77-107.

[3] 余黎青. 移动平台上基于内容的图像检索系统的研究与实现[D]. 合肥:安徽大学,2014.

[4] Grubinger M, Clough P, Hanbury A, et al. Overview of the ImageCLEFphoto 2007 photographic retrieval task[C]//Workshop of the cross-language evaluation forum for European languages. Berlin: Springer, 2007: 433-444.

[5] 陈　杰. 主题搜索引擎中网络蜘蛛搜索策略研究[D]. 杭州:浙江大学,2006.

[6] Martinez J M. MPEG-7 Overview (version 8). ISO/IEC JTC1/ SC29/ WG11[EB/OL]. 2002. http://mpeg.telecomitalialab.com/stan-dards/MPEG-7/MPEG-7.htm.

[7] 王海霞,覃团发. 综合 MPEG-7 中颜色特征的图像检索方法[J]. 计算机应用研究,2005,22(3):164-165.

[8] 王　栋. 基于 CRC 的多比特纠错算法研究与实现[D]. 西安:西安电子科技大学,2013.

[9] 周立柱,林　玲. 聚焦爬虫技术研究综述[J]. 计算机应用,2005,25(9):1965-1969.

[10] 杨开杰,刘秋菊,徐汀荣. 线程池的多线程并发控制技术研究[J]. 计算机应用与软件,2010,27(1):168-170.

[11] TF-IDF[EB/OL]. 2015-09-27. http://zh.wikipedia.org/wiki/TF-IDF.

[12] 李东阳. Android 手机上图像分类技术的研究[D]. 北京:北京邮电大学,2012.

[13] Khattab D, Ebied H M, Hussein A S, et al. Multi-label automatic GrabCut for image segmentation[C]//14th international conference on hybrid intelligent systems. [s.l.]: IEEE, 2014: 152-157.

[14] 周良芬,何建农. 基于 GrabCut 改进的图像分割算法[J]. 计算机应用,2013,33(1):49-52.