

基于灰度分类的图像搜索引擎

魏正曦, 邱玲, 赵攀

(四川理工学院计算机学院, 四川 自贡 643000)

摘要: 图像搜索是下一代搜索引擎迫切需要解决的课题,在对图像搜索引擎的主要功能和关键技术进行了分析和讨论的基础上,详细剖析了图像搜索引擎设计中任务分析、解决方案、感受哈希算法、实现等关键过程,并实现了一个基于灰度值分类的图像搜索引擎。实际试验表明,本图像搜索引擎的搜索速度较快、性能稳定,具有较好的搜索效果。

关键词: 图像搜索引擎;网络爬虫;图像识别;感知哈希算法

中图分类号: TP391

文献标志码: A

随着网络进入 Web2.0 时代,人们已不满足于仅对文本信息的搜索,还希望能够从已知图像中找到更多的相关图像,图像搜索引擎今后将成为用户检索网络图像的主要工具^[1]。以图像分析与图像识别技术为支撑的图像搜索技术正在成为人们关注和研究的热点。

谷歌、百度、搜狗作为用户搜索信息时最常用的三个综合性搜索引擎,已经推出了用关键词进行图像搜索的服务,但使用图像搜索图像的功能仍处在测试阶段^[2]。目前国内外还没有形成一个成熟的图像搜索引擎产品,主要原因在于图像比起文本其内容更加的丰富、复杂,所包含的信息量是文本无法比拟的。另外,数字图像处理技术还有很多难点没有解决,文本本身能够表达一定的语义含义,而图像只能通过本身的内容特征来表达。因此,基于内容的图像检索比起基于文本的检索难度要大的多。

1 设计任务

1.1 解决方案

与文本搜索引擎相对比,图像搜索引擎需要完成以下四项工作:自动收集网络图像、建立和维护图像索引数据库、计算图像相似性返回检索结果、设计友好的人

机交互界面^[3]。相应地,图像搜索引擎的开发涉及以下四项基本内容:

(1) 网络图像的获取。在完成网络图像搜索任务之前,搜索引擎所在主机要储备大量的图片以备图像比对。可以使用网络爬虫程序来搜索和下载 Internet 图像,该程序从某个网站的顶层开始,按照广度或深度优先的遍历策略下载网站图片,并按照一定的格式存储到本地文件系统中^[4]。

(2) 图像特征提取算法。搜索引擎对下载的图像文件首先进行统一格式化处理,然后提取图像特征、建立特征库。使用图像处理算法为每张图片生成索引,它可以用来比较、计算不同图片的相似度。将大量的图像索引都存入到一个数据库,即图像特征库。在建立图像特征库时,需要保存源图像在 Internet 中的 URL 地址以及其他属性以备用户将来查询,帮助用户获取该图像所在的原始页面。

(3) 图像搜索策略。用户在浏览器端上传想要搜索的示例图片,示例图片经过上载和格式化转换后送至搜索引擎软件的图像处理模块,根据图像比对算法,本地主机为上载图片生成图像索引,然后以索引为范本在图像特征库中检索相似图片;然后将匹配度超过某个临

收稿日期:2013-06-07

基金项目:人工智能四川省重点实验室项目(2011RYY03)

作者简介:魏正曦(1976-),男,四川自贡人,副教授,硕士,主要从事计算机应用方面的研究,(E-mail) 413789256@qq.com

界值的一组图片按一定的顺序排列^[5]、以缩略图的形式返回至用户浏览器界面,从而完成图像搜索任务。

(4) 编码实现。按照上述解决方案合理划分软件功能模块,选取相应的开发环境、数据库、软件工具包,逐一完成软件的编码和测试工作。

1.2 工作流程设计

图像搜索引擎的工作流程如下:用户在客户端上载待检索的图片,图片经过预处理后发送至服务器端,服务器端自动识别图像特征,在图像特征库中检索相似图片,最后在页面显示图像检索结果^[6]。整个步骤如图1所示。

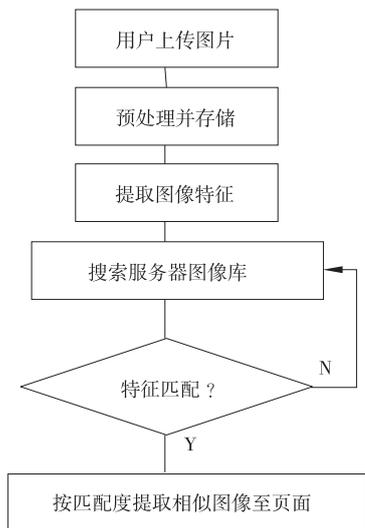


图1 工作流程图

2 图像处理算法

本设计的核心内容是图像匹配算法,即由已知的模板图中搜索相匹配的子图像。一般地,图像匹配算法有基于灰度值的方法如感知哈希算法、序列相似性检测算法等等,另一种是基于特征提取的方法,如基于颜色特征、纹理特征、形状特征、空间位置等特征的匹配。本文选用基于灰度值的算法用于建立图像特征索引。

基于灰度值的图像处理算法^[7]的基本思想是将图像看成是二维信号,采用统计相关的方法寻找信号间的相关匹配。利用两个信号的相关函数,评价它们的相似性以确定相同点。

灰度匹配可以利用某种相似性度量,如相关函数、协方差函数、差平方和、差绝对值和等测度极值,判定两幅图像中的对应关系。感知哈希算法^[8]是最具有代表性的算法之一。

2.1 感知哈希算法

感知哈希函数能在多媒体数据集与多媒体感知摘

要集之间建立起单向映射关系,也即,它能将具有相同感知内容的多媒体数字映射表示为一段唯一的数字摘要。

感知哈希函数被引入到图像识别领域是因其有如下的一系列优良特性:

(1) 唯一性:不可逆的提取原始数据的数字摘要,内容映射具有单向性。

(2) 区分性:感知内容不同的数字图像表示不会映射为相同的感知哈希值。

(3) 鲁棒性:感知内容不同的不同多媒体数字表示仍映射为同一哈希值。

(4) 摘要性:在满足以上基本性质的情况下,感知哈希算法处理所得的数据量所占的数据容量较小。

本设计即是根据感知哈希算法的上述特性,将其用于图像内容识别。

2.2 图像处理过程

利用感知哈希算法处理图像的工作过程如下:

(1) 图像格式化:将图片缩小到 $n \times n$ 的分辨率 ($n \leq 8$), 总共 n^2 个像素。这样可以去除图片的细节,只保留结构、明暗等基本信息,从而排除图像因为分辨率、亮度等属性带来的差异。

(2) 灰度降阶:将图像统一降阶为 T 级灰度,也就是,把所有像素点简化成总共只有 T 种颜色, $T \leq 64$ 级。这样处理旨在排除图像颜色数的差异,将其灰度值放在同一个范围内进行计算。

(3) 计算图像的灰度平均值^[9]:用公式(1)计算图片中所有 n^2 个像素的灰度平均值 u 。式中 x_i 表示某一个像素的灰度值, p_i 表示它在图像中出现的概率值。

$$u = \sum x_i \cdot p_i \quad x_i \in [0, T] \quad (1)$$

(4) 像素的二进制映射:按公式(2)把像素映射成一个二进制数。将每个像素的灰度与平均值进行比较,大于或等于平均值,记为1;小于等于平均值,记为0。

$$f(x_i) = \begin{cases} 1 & x_i > u \\ 0 & x_i \leq u \end{cases} \quad (2)$$

(5) 构造哈希序列值:将上一步的比较结果组合在一起,构成一个 n^2 位的二进制整数序列,例如 $\{1, 0, 0, 1, \dots, 0, 1\}$, 它就是每张图片的图像指纹。

(6) 图像比对:比对算法就是看不同图像的 n^2 位二进制整数中有多少位不相同,这相当于计算图像间的汉明距。一般而言,如果不相同的数据位不超过5,就说明两张图片很相似;如果大于10,就说明这是两张不同的图片。

执行算法的前 5 步可以从原始图片中计算图片的指纹值,并将图片指纹、原始图片路径、图片 URL 等相关信息写入到数据库。

在写入数据库的时候,需要存入一个关键字所对应的多个属性记录,这是因为同样的一张图片可能来自不同的网站链接。后续的工作可以通过对比不同图片的指纹,从而算出图片之间的相似度。

3 设计与实现

3.1 开发工具

Heritrix 是一个专门为互联网上的网页进行存档而开发的网页检索器^[10]。它由 Java 编写而成并且完全开源,并支持多线程,现在已经成为一个成熟的开源爬虫,并被广泛使用。图像搜索引擎采用 Heritrix 自动获取网络图像。

Berkeley DB 是一个高性能的嵌入式数据库,它为应用程序提供可伸缩的、高性能的数据库管理服务,支持数千的并发线程同时访问操作数据库,数据流量可达 TB 级别。它不需要对某种查询语言进行解析,也不用生成执行计划,这就大大提高了运行效率。基于这些特点,搜索引擎的图像特征数据库采用 Berkeley DB。

软件选用 MyEclipse 作为开发编译环境。该工具包的功能非常强大,支持的编程语言类型十分广泛,尤其是对各种开源产品的开发提供了多种便利。利用它可以在数据库和 Java 应用程序服务器的整合方面极大地提高工作效率。

3.2 功能模块

图像搜索引擎软件主要由两大功能模块组成,如图 2 所示。

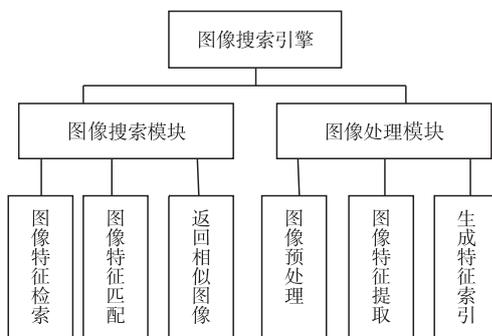


图 2 功能模块图

图像处理模块对客户端上传的图片进行预处理,包括安全检查、格式转换、尺寸处理等,其中格式转换、尺寸处理将根据实际需求进行调整;图片上传后存放于服务器的临时目录下;通过图像处理算法自动对图片内容

进行分析、识别,得到此图片的相关特性信息,生成图像索引,待下一步调用图像搜索模块进行检索。

图像搜索模块在用户检索之前,搜集并存储大量的图片,将图片按颜色、形状、直方图等特征进行分类,建立图像特征索引库;把经过内容识别的上传图片在图像特征库中进行检索,比对并将一组相似图片及其来源的 URL 返回至用户搜索页面。

本图像搜索引擎还提供了不同图像之间的匹配度,如公式(3)所示。算法首先找出不同图像的哈希序列值(n^2 位二进制整数序列)中有多少相同的二进制位 m ,然后计算 m 与 n^2 的比值 λ , λ 能反映不同图像的相似程度,比值越大表明图像也就越相近,如果两幅图像完全相同,则匹配度为 100%。

$$\lambda = \frac{m}{n^2} \times 100\% \quad m \leq n^2 \quad (3)$$

当相似图像作为搜索结果返回到页面时,图像搜索结果按匹配度大小降序排列呈现给用户。

3.3 测试结果

测试选取一台主机作为服务器,搭建和配置运行搜索引擎所需的硬件和软件环境,连接图像特征数据库,保证程序对图像特征库的有效访问。

图像搜索引擎的用户界面如图 3 所示。用户点击“浏览文件”按钮选择图像搜索的源图片。如果上传的图片符合要求,经过客户端的预处理,浏览器上面会出现一个进度条,上面有上传图片的名称以及图片容量。此时,“开始上传”按钮从最初的灰色变为可用;点击该按钮,图像搜索引擎便自动为用户搜索相似的图像。



图 3 用户上传图片

图像搜索结果如图 4 所示。在搜索结果页面中,点击图片或者图片下方的超链接地址则可查看图片在源网址页面。在 URL 下面,引擎还给出了相似图像的匹配相似度。

本测试中,图片库中的图片数量约为 5 万张。搜索损耗时间数据表明,即使图片量较大的情况下,本图像搜索引擎也能在 1 秒钟以内返回图像的搜索结果,并且给出相似图像的匹配度。这说明本搜索引擎的检索速度较快,采用的算法执行效率高。



图4 搜索结果图

4 结束语

基于内容的图像搜索是下一代搜索引擎迫切需要解决的问题,本文对其主要功能和关键技术进行了分析和讨论,设计并实现了一个基于灰度值分类的图像搜索引擎。测试表明本图像搜索引擎的总体性能稳定,基本达到了预期的设计要求。

需要指出的是,搜索结果中有时会出现匹配度适中而内容与上载图片不一致的图像。原因在于图像匹配度是按照整幅图像内容而非仅仅是对图像局部(如脸部)进行计算,这样处理有时会造成图像的语义误差。因此,本搜索引擎在图像识别算法方面还有待于进一步地研究和改进。

参考文献:

- [1] 何海地.互联网搜索引擎变革给图书馆服务的启示[J].图书馆杂志,2013,32(10):29-34.
- [2] 高 瑞.百度、Google、Sogou 三种图像搜索引擎功能的比较[J].中国科技信息,2010,41(18):91-92.
- [3] 高云辉,刘春双.搜索引擎技术在图像检索中的应用研究[J].计算机光盘软件与应用,2013(4):123-125.
- [4] 吕志花.网络信息挖掘及其在搜索引擎方面的应用[J].微计算机信息,2008,2(3):173-174.
- [5] 谢 辉,陆月明.搜索引擎中基于内容的图像重排序[J].计算机应用,2012,33(2):460-462.
- [6] 吴文超,汪 彦,舒 会,等.有关内容图像搜索引擎技术的探索与实践[J].计算机与数字工程,2009,37(10):124-127.
- [7] 魏正曦.图像阈值自动选取算法的 C++ 实现[J].四川理工学院学报:自然科学版,2010,23(4):420-422.
- [8] 孙 锐,闫晓星,丁志中.基于图像正则化的抗几何变换的感知哈希算法[J].工程图学学报,2010,31(2):116-122.
- [9] 梁金明,魏正曦.Ostu 算法的改进研究[J].四川理工学院学报:自然科学版,2010,23(5):543-545.
- [10] 张 敏,孙 敏.基于 Heritrix 限定爬虫的设计与实现[J].计算机应用与软件,2013,30(4):33-35.

Image Search Engine Based on Grey-Classification

WEI Zhengxi, QIU Ling, ZHAO Pan

(School of Computer Science, Sichuan University of Science & Engineering, Zigong 643000, China)

Abstract: The image search is an urgent problem of the next generation of search engines. The main functions and key technologies of the image search engine are analyzed and discussed in the article. Next, the main contents including task analysis, solutions, perception hash algorithm, as well as implementation methods and so on, are discussed. As a result, an image search engine based on gray-classification is designed. The tests show that the image search engine has very fast speed and stable performance, and basically achieves the desired design requirements.

Key words: image search engine; web crawler; image recognition; perception hash algorithm