Computer System Networking and Telecommunications
计算机系统网络和电信，2019,1(2)
ISSN:2661-3719 (Print); 2661-3727 (Online)

Research on the Massive Image Retrieval Method

Zilan TANG Zhengde BAO
School of Computer and Software, Jincheng College, Sichuan University, Chengdu, 611731

Abstract

Image retrieval is one of the foundations in the field of computer image. In order to adapt to the huge data scale, the retrieval of massive images has made many improvements compared with the traditional retrieval mode. This paper introduces in detail the different combination of massive image retrieval and the main frame technology used to make image retrieval and big data background fusion, and analyzes the efficiency and accuracy of different retrieval methods combined with existing literature, and proposes improvement direction for future massive image retrieval.

Key Words

Massive Image, Image Retrieval, Hadoop, Spark, HDFS

DOI:10.18686/jsjxt.v1i2.687

海量图像检索方式的研究

唐子岚 鲍正德
四川大学锦城学院计算机与软件学院，四川成都，611731

摘 要

图像检索是现今计算机图像领域的一大基础，为适应现今庞大的数据规模，针对海量图像的检索相较于传统检索模式也已经做出了诸多改进。本文详细介绍了海量图像检索的不同组合实现方式和主要使用的框架技术，使图像检索与大数据背景融合，并结合现有文献分析了不同检索方式间的效率及准确度，同时对未来海量图像检索提出改进方向。

关键字

海量图像；图像检索；Hadoop；Spark；HDFS

1.引言

随着计算机视觉在近年的迅速发展和将计算机应用在图像处理方式中的不断探索，图像检索技术也随之拥有光明的应用前景。图像检索可应用于多媒体、医疗、公安、设计、教育及自动化技术等广泛领域，且随着信息量的扩展，图像数据的体量也越来越庞大。怎样处理海量图像成为图像检索技术的重要突破口，本文便对海量图像的检索方式进行研究，并提出现阶段海量图像处理的不足和未来发展方向。

2.图像检索的原理

图像检索的过程主要分为三个阶段：用户需求的分析和转化，明确向数据库发起检索请求的方式；提取图像特征；根据相似度算法计算图像匹配度，从而获取检索结果。

2.1 图像检索类型

图像检索主要分为基于文本的图像检索（TBIR）和基于内容的图像检索（CBIR）两种类型。基于文本的检索方式主要应用在图像数字化之前，需要人为标注图像信息，不仅受制于图像信息需要人工标注从而易产生歧义，更不能适应如今海量图像的更新迭代。现今的海量图像检索技术主要基于图像内容进行检索对图像进行特征提取，或者结合文本与内容进行综合性分析实

现更高层次的准确检索。

2.2 基于内容的图像检索方法

基于内容的图像检索普遍使用特征输入的查找方式，图像特征包括颜色、纹理、平面空间对应关系、外形等，在过去的实验中，常用于检索图像的特征有颜色直方图、颜色布局、Tamura 纹理特征、边缘直方图等特征，且在实际检索中可以根据需求为不同特征赋予不同的权重[1]。在处理图像时，首先使用 SIFT 或 SURF 算法进行特征提取，提取的图像特征通常还要使用 K-Means 算法进行聚类分析以得到统一的特征向量。

3.海量图像检索方式

海量图像检索主要可以划分为三个模块：海量图像的存储、图像特征提取、检索计算，其具体流程如图 1 所示。研究者们在如何优化存储、提高计算效率等方便做出了多方面的研究，并用实验验证了所提出的检索框架的合理有效性。

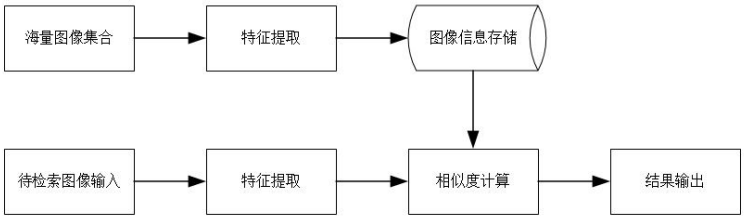

图 1 海量图像检索流程

在早期的研究中，研究人员已经对基于内容的图像检索做了大量的研究，而在近十年内，研究者们逐渐开始对海量图像的检索方式进行探索。文献[2]的图像特征选取了颜色直方图，颜色布局，颜色相关图，Gabor 过滤器和 Tamura 纹理，提取出的图像信息由数据库 Hbase 中的一张表存储。其索引过程由 MapReduce 实现，并采用了 Lucene 的增量索引实现索引的并行化，而 Lucene 创建的索引文件都存储在 HDFS 上。C. Gu 等人在实验时建立了一个 B/S 的搜索页面，可以直观地检索图像，但他们采用的图像集数量较少，没有体现出真正意义上的大数据处理的优势。文献[3]在图像特征的提取方面也采用了分布式的处理方法，使用 MapReduce 输出图像特征点，再通过分布式的 K-means 算法聚类构建单词表。张学浪等人的实验也是较早使用 BOW 模型为图像特征建立词库的，大大减少了计算量。在输入图像进行匹配时，该实验采用了 KNN 算法得到图像直方图，通过直方图月特征库进行匹配。文献[4]使用的存储方式也是 Hbase 数据库，但在表的设计与文献[2]略有不同。林文煜等人在图像的检索方式及算法等方面没有太多的创新，但他们设计了海量图像检索的一个清晰的框架，包括图像的输入模块、分布式集群处理模块和检索结果的显示模块。文献[5]在图像的特征提取部分除了常规的用 K-Means 算法对特征向量聚类外，还采用了 TF-IDF 数据挖掘技术对特征向量实时了量化。在建立索引和搜索时都用到了分布式处理方法，同时也基于 Lucene 索引文件实现海量图像的搜索。在进行试验验证时，王立等人的图像数据规模达到了上亿级别，且分别对图像的不同特征在检索时的效率进行了对比，其检索效率和检索的准确性的提升较为有说服力。

除了使用 Hadoop 集群和 MapReduce 的组合外，近年来研究者们也对使用 Spark 检索海量图像进行了一些探索。文献[6]便提出了基于 Spark 的图像检索方法，使用 Spark 代替 MapReduce 提取图像特征，然后使用聚类算法将图像特征训练出一个视觉词袋模型（BoVW）。该系统利用了 Spark 抽象弹性数据集 RDD 的优势，并通过实验验证了 Spark 集群在图像数据处理中的高效性。文献[7]也是近期基于 Spark 大数据平台进行海量图像检索的研究，同样验证了 Spark 相较于传统检索方式在效率上的优势，但他们的研究都没有将分布式数据库 Hbase 整合到系统框架中，对图像数据的管理高效性有提升的空间。文献[8]对 Hadoop MapReduce 和 Spark 在图像处理效率上进行了实验比对，结果表明，Spark 在处理迭代和实时处理作业时的效率更高，然而在索引中，没有中间数据任务的情况下，Hadoop M/R 依然占有优势。

4.海量图像检索的未来发展方向

在过去的研究中，研究者们对海量图像的检索主要关注点在于分布式的检索计算方式上，但提高检索效率也可以多从图像特征的提取角度出发。在面对不同需求时，可以选取不同的特征进行匹配，针对不同特征可以使用不同的算法，且聚类算法也可以进行优化改进。此外，研究者们往往把重点放在了检索效率的提升上，对于检索的准确度关注度还不算太高。

在 Hadoop 和 Spark 大数据处理平台的对比上，国内研究者所做的实验也并不太多，在统一的数据量条件下检索运算的时间等变量还没有准确可靠的数据，多数效率的比对都是基于传统检索方式和分布式处理的比对。未来研究者们可以探索基于 HDFS 分布式存储的更高效的存取加载方式，提高海量图像处理的系统稳定性，在进一步提高检索效率的同时提升检索准确度或者与图像识别等领域的算法结合，探索更为普适、易用的海量图像检索框架。

5.总结

本文对海量图像的检索方式和技术架构进行了分析和总结，目前这类方法在设计实现上已经较为成熟，许多研究人员也采用了不同的数据集、不同的算法和分布式平台等进行了实验验证。但总体检索效率和准确度仍有提升的空间，有待进一步改进。

参考文献

[1]张学浪. 基于 Hadoop 的海量图像检索关键技术研究[D].西北农林科技大学,2013.

[2] C. Gu and Y. Gao, "A Content-Based Image Retrieval System Based on Hadoop and Lucene," 2012 Second International Conference on Cloud and Green Computing, Xiangtan, 2012, pp. 684-687.

[3]张学浪,耿楠.基于云计算的图像并行检索关键技术研究[J].计算机应用与软件,2013,30(05):220-222.

[4]林文煜,戴青云,曹江中,何小明,李能.一种基于内容的海量图像检索框架的设计与实现[J].电脑知识与技术,2016,12(09):212-215.

[5]王立,陈军峰.Hadoop 分布式的海量图像检索[J].现代电子技术,2018,41(09):62-67.

[6]王迅,冯瑞.基于 Spark 的海量图像检索系统设计[J].微型电脑应用,2015,31(11):11-13+17+2.

[7]曹健,张俊杰,李海生,蔡强.基于 Apache Spark 的海量图像并行检索[J].计算机应用,2018,38(S2):183-186+230.

[8] M. A. Hedjazi, I. Kourbane, Y. Genc and B. Ali, "A comparison of Hadoop, Spark and Storm for the task of large scale image classification," 2018 26th Signal Processing and Communications Applications Conference (SIU), Izmir, 2018, pp. 1-4.

[9] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta and R. Jain, "Content-based image retrieval at the end of the early years," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 12, pp. 1349-1380, Dec. 2000.

作者简介

第一作者：唐子岚（1998-），女，汉，四川省成都市，本科，四川大学锦城学院，研究方向：大数据、图像处理。

第二作者（通讯作者）：鲍正德（1989-），男，汉，黑龙江哈尔滨，研究生，四川大学锦城学院，研究方向：电子商务。