学校编码：10384　　分类号____密级____
学号：23020111153059　　UDC___

厦门大学

硕　士　学　位　论　文

改进 K-Means 聚类算法在基于 Hadoop 平台的图像检索系统中的研究与实现

A Research and Implementation with Improved K-Means Clustering algorithm To Image Retrieval System Based On Hadoop Platform

黎光谱

指导教师姓名：郑建德　　教授
专　业　名　称：计算机应用技术
论文提交日期：2014 年　　月
论文答辩时间：2014 年　　月
学位授予日期：2014 年　　月

答辩委员会主席：_____
评　　阅　　人：_____

2014 年　月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下，独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

摘 要

现代人们的生活已经进入了移动互联网时代，各种移动互联网设备的普及和广泛应用极大的方便了人们的生活学习等各个方面。与此同时，来自各行各业的大量信息正以多媒体信息的方式数字化并不断累积。其中图像作为最为基本的多媒体信息之一易于理解和使用，人们对图像检索的需求也从开始的根据文本描述来检索图像发展到根据图像内容来检索相似图像。

图像检索早已成为计算机领域的一个研究热点，它可以按照检索内容划分为基于文本的图像检索和基于内容的图像检索。本文主要的内容是如何应用大数据技术进行基于内容的海量图像检索技术的研究和实现。

从数据层面分析，一个基于内容的图像检索系统要解决大量图像数据的存储和快速处理两个最主要的问题，本文将使用专门用于大数据存储和处理的 Hadoop 技术来存储大量的图像数据并进行离线地分布式计算；从检索技术层面分析，要进行特征提取和处理，本文将提取图像的 SIFT 特征，然后对这些特征进行 K-Means 聚类，将聚类结果作为 Bag-of-Words 模型的视觉词袋对所有图像的 SIFT 特征点进行量化处理，从而用一个固定维数的特征向量表示每一幅图像，此外再用 TF-IDF 加权技术处理这些特征向量，最终计算这些图像与检索图像特征向量之间的相似度，返回相似度最小的一些图像。

本文使用并修改 HIPI-Hadoop 图像处理接口实现在 Hadoop 上进行图像类型的存储处理和检索，提出了一种改进的并行 K-Means 算法并应用于特征点的聚类，使用一种基于面积的相似度计算方法计算图像特征向量间的相似度。改进了部分 Mahout 源码适应大数据的处理需求。

图像检索应用广泛，对基于 Hadoop 的图像检索系统的研究将对大数据时代图像检索技术的发展起到一定的指导作用。

关键词：图像检索，大数据，存储，分布式计算，Hadoop，HIPI，K-Means

ABSTRACT

Contemporary people's life has entered The Mobile Internet era. People's life and study and some other aspects have benefited a lot from the popularization and widespread application of various mobile internet devices. At the same time， lots of information from every walk of life are digitized and accumulating in the form of mutimedia information. As one of the most basic multimedia information，The image is easy to be understood and used，People's demands for The Image Retrieval are also developed from the begining of Retrievaling according to the text description to Retrievaling similar images according the image content.

The Image Retrieval has been a research hotspot in the field of computer science，it can be divided into Text-Based Image Retrieval and Content-Based Image Retrieval. The primary content of this paper is how to do research on The Content-Based Image Retrieval and implementation with huge amounts of images by using Big Data Technology.

From the aspect of data analysis， a Content-Based Image Retrieval system must figures out two principal problems which are The Storage and Rapidly Processing of a large number of image data. We will use Hadoop technology dedicated to The Storage and Processing of Big Data to store huge amounts of image data and proceed off-line distributed computing; from the aspect of Retrieval technology analysis， we need to proceed feature extraction and processing， in this paper， we would extract the images' SIFT features， and then cluster these traits with K-Means clustering， next that， quantify all the SIFT features by using Bag-of-Words Model with the bag-of-visual-words which is the clustered result forward，so we can present a image with a fixed dimension feature vector， besides， dispose these vectors using TF-IDF weighted technology， Finally， calculate the similarity between these images' vectors and retrieval image vector， return several images with smallest similarity.

This article would use and modify HIPI-Hadoop Image Processing Interface to calculate with image type on Hadoop and store them， proposed a revised parallel

K-Means algorithm and applied it to the clustering of feature points， A similarity calculation method based on area would be used to calculate the degree of similarity between image feature vectors. In order to adapt to the requirements of The Big Data Processing we have the source code of Mahout improved.

The Image Retrieval has a widespread application， the research to The Image Retrieval System based on Hadoop will play a guiding role for the development of Image Retrieval Technology in Big Data Era.

Keywords: Image Retrieval， Big Data， Storage， Distributed Computing， Hadoop， HIPI， K-Means

目 录

Contents

第一章 绪 论

1.1 研究背景及意义

当代，人们的生活已经慢慢从过去的 PC 时代步入了移动互联网时代，各行各业的信息正在不断的累积，例如纽约证券交易所每天产生 1TB 的交易数据、Facebook 存储着约 100 亿张照片共约 1PB 存储容量、瑞士日内瓦附近的大型强子对撞机每年产生约 15PB 的数据[1]，此外还有像京东和淘宝这样的互联网金融网站的交易数据、国家气象数据、天体运动观测数据、后台日志数据等。因此数据规模飞速增长，如何有效的管理、高效分析这些数据成为了新的研究热点。

图像作为一种广泛使用的基本的多媒体信息，不论是在科学技术还是日常生活等各个方面都涉及到图像数据。与此同时人们感到要在海量的图像中寻找到自己所需要的图像变得越来越困难，理论源于实践，于是人们开始对图像检索进行研究，目前图像检索已经成为了当今计算机研究领域的一个研究热点[2]。

图像检索就是根据对图像内容的描述，在图像数据库中查找具有描述特征或者包含与描述特征最相似的特征的图像[3]。图像内容主要包括视觉内容和信息内容。视觉特征比如形状、颜色、纹理等属于图像的物理表示；信息内容即图像的语义，如主题、场景、尺寸、年代等。

传统的基于单节点架构的图像检索系统已经无法应对大数据时代海量图像的存储和计算需求，此时，Google 主要采用 GFS[4]分布式文件系统来存储海量数据，并用 MapReduce[5]模型进行分布式计算，用 BigTable[6]替代了传统的关系数据库，它是一种基于键值对型的非关系数据库。

Hadoop 是 Apache 软件基金会组织下的一个开源项目，它采纳 Google 的 GFS 和 MapReduce 思想实现了 HDFS（Hadoop Distributed File System）和 MapReduce 编程框架[7]，可以在廉价的机器上部署分布式集群，HDFS 用于分布式存储大数据并用 MapReduce 编写并行计算程序，其实 Hadoop 可以看作是一个任务调度和管理工具，好比当某个人有很多任务要做的时候，如果时间宽裕，他可以选择一个人一件一件的顺序完成，也可以选择安排几个人同时做，他要做的就是分配任务并统计综合，而 Hadoop 就提供了自己的语言 MapReduce 来分配和调度并最终

总结处理结果。

目前图像检索的发展受到图像处理和人工智能等方面技术的限制. 为了实现自动化、智能化、通用的图像检索系统，将 Hadoop 的分布式框架应用于图像检索领域可以实现一个高效、稳定易扩展的图像检索系统。图像检索应用广泛，例如外观专利图像检索、医学图像检索、指纹人脸识别等。此外，图像检索还是视频检索的基础。

1.2 国内外研究现状

有关图像检索的研究从上个世纪 70 年代就已经开始，那时主要是通过文本信息来检索图像，比如图像的拍摄日期、作者、图像的类别和尺寸等，我们把这类图像检索方式称之为基于文本的图像检索技术。经过将近 20 年的发展，出现了基于内容的图像检索技术，该种检索方式主要是通过图像的视觉内容如颜色、形状、纹理等特征来检索相似图像，其目的是为了避免使用文字上的描述而是以视觉相似性为基底通过用户提供查询的图像或是用户指定的图像特征来检索相似图像[8]。

目前，基于文本的图像检索技术已经非常成熟，但是也有其缺点和不足：首先，对于图像的文本和关键字标注需要人工完成。再者，由于人工标注存在着观察角度的不同和主观意的念差异等问题，会产生标注歧义，也就是说人工标注很难完全表达一张包含了多个目标甚至涵盖了一些感情色彩的图像。由于基于文本的图像检索的不足，随着市场需求的提高和研究的深入，最近不久就相继出现了百度识图，谷歌识图，搜狗识图等基于视觉内容的图像检索应用，虽然准确度还有待提高，但它已经为图像检索领域的研究指明了一个明确的方向。基于内容的图像检索技术，能够自动提取图像特征，避免了歧义性[9]。

基于内容的图像检索系统是对图像处理、模式识别、信息检索等技术的集成综合。经历了理论的反复论证和技术的不断创新，出现了一个个成熟的图像检索系统，主要代表有[10]：

1. 第一个以微计算机为基底开发的图像数据库检索系统，是由 80 年代麻省理工学院的 Banireddy Prasad、Amar Gupta、Hoo-min Toong、and Stuart Madnick 所共开发出来的。这是记载于 1987 年 2 月发布的 IEEE Transactions on Industrial

Electronics[11]。

2. 最早出现并商用的基于内容的图像检索系统是 IBM 开发的 QBIC 系统[12]，它支持基于 WEB 的图像检索服务，用户在线提出查询要求，系统经过查询后按相似性顺序给出查询结果。

3. VisualSeek 是由美国哥伦比亚大学研究的基于内容的图像检索系统[13]。

4. 业界检索效果比较好的就是 Google 图像搜索，它采用了机器学习算法，综合考虑的文本和图像的视觉特性。

5. 国内关于基于内容的图像检索的研究是从 20 世纪 90 年代后期开始的，在卫生安全和机器制造方面得到广泛应用，还对农业生产和国防研究等领域带来重要的影响[14]。如国防科技大学设计的 NewsVideoCAR 系统、中科院的“Mires 系统”[15]、浙江大学的 WebscopeCBR 系统。

Hadoop 是主要用于海量数据的存储和分析的计算平台，其在数据挖掘领域已经大显身手，其开源项目提供了对文本数据进行处理的各种接口，被各种大型企业用来处理日志文件，交易数据等，从中进行数据挖掘做出市场预测、产品推荐、问题反馈等用途。但是在 Hadoop 上面进行的有关图像的处理却不是很多，主要是因为图像数据有别于基本的文本数据，Hadoop 没有提供专用的类型用于处理图像，需要开发研究人员自行设计。不过近年来也相继出现了基于 Hadoop 的图像处理的研究和应用，比如美国弗吉尼亚大学就开发了 HIPI(Hadoop Image Process Interface)，即 Hadoop 图像处理接口用于在 Hadoop 上进行图像处理。Facebook 用 Hadoop 进行 PB 级别图像的存储和分析，创业公司 Skybox Imaging 也使用 Hadoop 来存储并处理从卫星中拍摄的高清图像数据。

1.3 本文内容与结构

本文使用开源 Hadoop 项目解决大规模图像的存储和并行计算问题，来缩减图像检索系统的后台计算时间，主要利用到了 HDFS 的大数据存储的特性和 MapReduce 并行计算的能力。

首先，搭建了一个三个节点的小型 Hadoop 集群，作为研究的实验平台；其次，通过对并行计算框架 MapReduce 的学习，设计实现了满足图像检索系统需求的 MapReudce 框架上的图像类型；最后，根据具体检索方案，设计后台数据