1. 绪论
   1. 论文的研究背景及意义

二十一世纪是一个信息爆炸的世纪。随着互联网技术的普及应用，各种各样的内容充斥于网络之上，文档、图片、视频影片、web页面等等，令人目不暇接。海量的信息内容极大的丰富了人们的生活，但随着内容的进一步丰富，我们发现，想要找到自己需要的内容信息却变得越来越困难。对于企业而言亦是如此, 调查表明,大中型企业平均会有至少一个主要的内容库每天都在运行中[1]。一个互联网企业，在自身发展过程中，往往会产生大量的图片，文档等素材，如不能对其实现有效的管理，则会造成资源的浪费，重复开发等问题。

中国的互联网企业，往往缺乏内容管理的传统。开发者本身往往缺乏编写文档意识，而即使形成了简单的接口文档，也往往是保留于开发者手中，在有人需要的时候在通过近似“口耳相传”的方式传播，这对系统的后期维护造成了极大的困难，人员变更时的工作交接更是成为了一个噩梦。同时，开发时实用的美工素材等往往也得不到有效保存，导致不同部门甚至同一部门下的不同产品风格各异，重复的开发工作也加重了美工，UI设计人员的工作量。而内容管理系统则是解决企业这一困难的有效工具。

内容管理，即Content Management(CM)，是指一系列用以支持对内容的收集，管理，发布的过程与技术的集合，其关注的重点放在各种非结构化或半结构化的数字资源的采集、管理、使用方案、传递和增值上，并可以集成到结构化数据的环境中[2]。这里的内容广义上讲，可以是任何形式的信息；狭义上讲，则是指通过电脑存储的数字信息。具体而言，内容可以是文本，图片，音频、视频等多媒体信息，或任何需要被管理的文件类型。

内容管理技术已经被广泛应用于各种领域之中，例如新闻媒体，电商网站等，都对内容管理有着天然的需求。同时，用户对于内容管理系统的需求也是多种多样的。有些使用者可能更关心海量内容的存储，而另一些则关心如何快速准确的搜索到需要的内容信息。因此，对很多企业而言，通用化的内容管理系统可能并非不能很好的适应其需求，其所需要的是功能更加灵活并且订制化的内容管理系统。

普通的内容管理系统的实现，在存储，版本控制，文本搜索等基本功能上已经日趋完善，然而基于文本内容检索图片则往往不能取得很好的效果。在目前多数企业应用的内容管理系统之中，起搜索功能往往是基于关键字，或是TAG（标签）的。对于文本类内容而言，这样的搜索形式已经足够满足大部分使用需求，而对于图片等难以结构化存储的内容数据而言，其效果往往较差。图片类资源的关键字，TAG等，往往需要用户在上传资源时主动添加。在系统中资源数量积累到一定程度后，此种搜索方式只能模糊定位，用户还需要在搜索结果中进行二次过滤，才能找到自己需要的资源。同时，也容易造成资源重复上传的问题，浪费存储空间。在这种背景下，让用户可以通过“以图搜图”的形式，例如使用分辨率较低的缩略图获取高清的图片资源，或是通过手绘的简略内容，获取类似的图片素材文件，就具有十分现实的意义。如何将图像识别技术应用于内容管理系统之中，为用户带来更好的使用体验，将是本文的重点研究内容之一。本文将在实现内容管理基本功能的同时，致力于将图片检索技术应用于内容管理平台之中，实现“以图搜图”的功能，让美工人员，UI开发人员可以更加方便的寻找到自己需要的图片素材，提高工作效率，具有良好的现实意义与应用前景。

* 1. 国内外研究现状
     1. 内容管理技术的国内外研究现状

从内容产生的时刻起，内容管理也就伴随着应运而生了。而“内容”本身的历史也可以说是人类认识的历史[3] 。例如图书馆书籍的管理，家谱族谱等等，都可以认为是内容管理的范畴。而数字化的内容管理技术被广泛认知则是在互联网应用流行以后[4]。IBM曾把信息管理定义为数据管理(DM Data Management)和内容管理(CM Content Management)两部分集成，显示出内容管理的重要性[5]。

用于内容管理的数字化软件应用被称为内容管理系统(Content Management System)[6][7][8]。内容管理系统的概念从被提出至今，其内涵随着数字技术、互联网技术的发展不断丰富。在当前的技术条件下，基于Web技术的内容管理系统已经成为了主流的发展方向，而在Web开发中，Java技术则收到相当多开发者的青睐，因此很多内容管理系统都是构建与Java EE之上[9][10]。目前国外从事内容管理系统研发的主要厂商包括Mambo、Vignette、Interwoven、BroadVision、Openmarket、ATG、Allaire、Documentum、Hummingbird 等[11]，这些公司的内容管理产品与解决方案大多面向企业级用户，具有很强的专业性。TRS与TurboCMS则是国内影响力较大的两个内容管理产品[12]。

* + 1. 图像检索技术的国内外研究现状

“图像检索”，顾名思义，即指对图像进行搜索的过程。在过去数十年所产生的信息爆炸中，内容信息的数量呈几何倍数增长，而其中很大一部分是非文本类的数据。这同时也意味着这些非文本类的数据的搜索需求也有了惊人的增长。

传统的图像检索技术是基于描述的。即用户人为地对图片进行描述，或给图片打上关键字标签（TAG），本质上仍是文本的检索。这种方式有如下几个严重的缺陷：

1. 必须对图片进行描述才能搜索。这个过程可能需要人为进行，自动化的效果不高，并没有实际意义。
2. 检索精度不够。例如用户输入“动物”，“鸟类”这样较为模糊的关键词时，很有可能还需要花费大量精力对结果进行二次筛选。
3. 对于复杂的图片描述困难。对于内容很丰富的图片，只能使用简单的关键词描述，会使检索的结果大打折扣。

因为存在着上述困难，基于内容的图像检索(Content Based Image Retrieval, CBIR)[13]就应运而生了。CBIR并不需要对图片进行描述，而是依据图片本身的内容，根据用户提供的相似图片进行检索，这就大大提高了检索的精度。CBIR的实现依赖于两个关键步骤：图像的特征提取和匹配。

图像特征提取：图像的特征是多方面的。上文所述的“关键字”本质上也是对图像特征的一种描述。限于目前的技术发展水平，计算机只能自动化的识别图像的一些低阶特征，如纹理，颜色，形状等。而在图像检索算法中，检索的基础则正是这些低阶视觉特征。提取这些视觉特征，并将其转化为数字特征以便被计算机识别，就成为了图像检索算法的重中之重，也是一大难点。

图像特征的匹配：在正确提取了图像特征之后，就可以对数据库中的图像进行索引以供检索。图像特征的匹配指的是在用户给出了目标图像之后，如何快速准确地通过索引定位到相似图片的过程。

CBIR从1992年被提出至今，已经经过了二十多个年头的发展，并取得了一定的成果，目前已经有一些较为成熟的算法被提出。同时，CBIR也被应用于相当多的领域之中，如人脸识别[14]，医疗[15]等。

选择合适的图像检索方法，并将其应用于内容管理系统之中，为用户检索内容素材提供更加便捷的途径，是本文的重点研究内容之一。

* 1. 论文的研究目的和组织结构
     1. 论文的研究目的

研究分析和研究网站内容管理系统的软件架构、工作原理和功能应用。将内容管理系统与企业的具体需求相结合，针对企业文档、图片素材多，人员权限需要细致管理的特点，对系统以及各个模块进行详细设计。

研究图像检索的相关技术。查阅，学习相关研究论文，了解掌握目前主流的几种图像检索算法的原理，并根据系统的需求以及企业软硬件条件的限制，从其中选取几种合适的算法。分析，比较其作为内容管理平台图像搜索工具的优劣，从中选取最合适的算法，将其集成至内容管理系统之中。

编码实现内容管理系统。选取合适的开发框架，实现内容管理系统，并对其进行测试验证，最终达到可交付的水平。

* + 1. 论文的组织结构

本文的章节安排如下：

第一章为绪论，描述了课题产生的背景，对内容管理系统和图像检索技术作了简单的介绍，并阐释了本文的研究内容和研究意义。

第二章为相关技术及算法简介，前半部分针对内容管理技术，Java EE相关的框架开发技术作了简介。后半部分重点针对目前主流的图像检索算法作了简介，同时详细介绍了本文所选择的算法。

第三章为内容管理系统的分析与设计，主要进行了系统的需求分析，并在此基础上对系统进行了整体设计，将系统划分为各功能模块，简述了每个模块所起的作用。

第四章为内容管理系统的实现，在第三章的基础上，对系统的各个模块进行详细设计，对系统的关键模块，如文档管理，分类管理，尤其是图片内容检索模块进行了详细的阐述并给出了实现方式。

第五章为系统测试，从人工测试以及自动测试两个方面对系统测试过程进行了阐述，重点介绍了图像检索，内容分类管理等模块的测试过程。

第六章为总结与展望，首先总结了本文的工作及意义，然后对未来可进行的进一步研究工作做了展望。

1. 2. 相关技术及算法简介

本文的研究目的是构建一个基于CBIR的内容管理系统，而其中基于内容的图像检索技术(CBIR)是本文的研究重点。本章将从三个方面对本文所使用到的关键技术进行介绍：首先是内容管理相关的技术，其后对图像检索技术，特别是本文所使用的感知哈希算法进行了重点阐述。

* 1. 内容管理关键技术简介

内容管理的含义十分广泛，对于内容管理系统应该如何构建，也缺乏统一的标准，各个公司实施的内容管理系统均具有各自的特点，但仍不失一些共性。一般认为，一个完整的内容管理系统至少需要包含以下几方面的内容：

1. 收集系统：即负责进行内容的收集、获取、分发、编辑、整合及转换的工作，同时应该能够满足用户对内容的查询需求。
2. 管理系统：负责对内容的存储进行管理，确保内容存取过程的正确性。同时应该具有对内容版本更替、分支的控制、记录功能。针对系统内不同人员的角色控制也可归入管理系统之中。
3. 工作流系统：即保证内容管理流程可以正确运行的系统。内容管理的流程包括了内容从收集起，经历存储，展现，一直到内容的发布。

内容管理始终是围绕着内容进行的，而以下几个方面则是尤为重要的：内容存储，内容查询。

内容存储：对任何一个内容管理平台而言，其所存储的内容都是相当广泛的。从简单的文本内容，到图片、音频、视频等多媒体内容，以至于压缩归档，可执行程序等等，都可以作为需要被管理的内容。而对内容的存储则是内容管理系统需要解决的首要问题。随着云技术的发展，传统的存储方式已经逐渐被新的分布式存储解决方案所替代。内容管理平台往往需要管理海量的内容信息，因此分布式存储解决方案将是最好的选择。本文所选择的存储方案是百度公司内部开发的NFS存储系统。

内容查询：内容管理系统需要将内容正确的呈现给用户，同时也需要给用户提供快速准确定位内容资源的手段，因此需要合适的内容搜索方法。目前内容管理平台的搜索方法主要有以下几种：

1. 基于标题的搜索：任何内容文件都具有标题，可以根据标题建立索引以供内容检索。
2. 基于标签的检索：用户在发布，编辑内容时，可以根据内容的特征为内容打上标签(TAG)，并据此建立索引以供内容检索。
3. 基于内容的检索：目前主要是针对文本内容，可以直接根据文本的内容信息进行内容检索。

对于图片内容信息，也可以根据内容进行检索。依据内容的图像检索(CBIR)可以满足用户某些特殊的检索需求，如根据缩略图搜索高清原图，根据手绘简图进行相似图像检索等。有关CBIR技术将在下节详细介绍。

* 1. 图像检索技术主流算法简介

CBIR从二十世纪八十年代被提出以后，一直是计算机领域一个热门的研究内容。很多学者，互联网公司纷纷投入其中，美国的谷歌公司，微软公司，国内的百度公司等都纷纷推出了“以图搜图”的引擎。

CBIR的两个核心研究内容是图像特征的提取和图像特征的匹配，两者都是针对“图像特征”而言的。

图像特征，即可以被用来描述图像的一系列内容，可分为图像的全局特征与局部特征，而全局特征又可进一步划分为颜色特征，纹理特征，形状特征等。

颜色特征[16]：颜色特征是图像最显著的特征，具有稳定，旋转、评议、尺度变化无关性，且颜色特征计算简单，表现出很强的鲁棒性[17]。颜色特征的提取一般和颜色空间相关，常见的颜色空间包括RGB空间，HSV，YCrCb[18]等。可以通过颜色直方图[19]，颜色矩[20]，颜色协方差矩阵[21]，颜色布局等来表征图像的颜色特征用以进行图像检索。

纹理特征­[22]：纹理特征主要描述的图像的相随灰度集火颜色的某种规律性变化。各种物体都有各自的纹理特征。比较著名的纹理特征是T.Ojala于2002年提出LBP方法[23]等。

形状特征：相比图像特征和纹理特征，形状特征承载了较多的语义信息，因而在图像检索中可提供比较强的描述能力。形状特征的表示方法包括了傅立叶描述子[24]，多边形逼近，不变矩，B样条，形变模板和曲率尺度空间等。

除了上述颜色、纹理、形状等全局特征，还可以使用图像的局部特征表征图像用以进行检索。相比全局特征，局部特征对于很多尺度变换，摄影变换都是鲁棒的。

在基于图像局部特征的检索算法中，SIFT算法应用比较广泛。SIFT，即尺度不变特征变换（Scale-invariant feature transform，SIFT），是一种针对图像局部特征对图像进行特征提取的算法，由David Lowe于1999年提出[25]。

SIFT具有位置、尺度、旋转不变性，即无论图片进行缩放，旋转，改变亮度或拍摄视角，都不影响算法对图片的特征识别，用于图像的检索和匹配时具有一定的优势。

SIFT算法对图像的尺度，位置均具有不变的特性，图像的旋转亦不能对算法结果造成影响；同时，改变图像视角，对图像进行仿射变换也不会大规模改变算法结果。SIFT的结果蕴含十分丰富的信息内涵，碰撞的可能性亦较小。基于这些特征，SIFT作为图像检索算法，具有检索快速，结果准确的特点。

近年来，随着技术的不断发展，一些新的技术也逐渐被应用于CBIR中，如卷积神经网络[26]，Contourlet变换[27]等。在一些特殊领域，如高光谱图像分析[28]，也有CBIR技术的介入。

* 1. 感知哈希算法

哈希函数(Hash Functions)，是一类能够将变长数据映射为定长数据的函数，被广泛应用于计算机的各个领域中，哈希表就是哈希函数的一种最典型应用。除了数据查询外，因为哈希函数具有不可逆的特点，也被应用于信息安全领域，数字签名技术就是哈希函数在信息安全领域的一个典型应用。

传统的哈希方法往往是直接针对数字内容进行运算的。例如一幅图片，可以直接根据其二进制表示通过一系列复杂的运算计算出哈希值。这样的方法运算简单，但由于其与图像本身的内容无关，因而产生了信息的丢失。同一幅图片，经过放缩，裁剪等操作，或者改变压缩方式，都会导致其二进制表示产生巨大的变化，从而彻底的改变哈希值。而从人类认知的角度而言，这些操作并没有改变图片的本质信息，他们是“相似”的图片，也因此应该具有相近的哈希值。而传统的哈希方法并不能满足这一需求，因此针对多媒体文件的感知哈希方法被适时提出。

感知哈希函数，是基于认知心理学的信息加工理论，由多媒体数据集到多媒体感知摘要集的一类单向映射，将具有相同感知内容的多媒体数字表示唯一地映射为一段数字摘要满足感知安全性要求[29]。所谓感知特征，即基于人类认知所总结出的特征信息，如图片的颜色，布局等信息。对于图像的放缩，压缩等操作几乎不会改变这些感知特征。感知哈希函数正是基于对媒体内容的感知特征计算，其显著特点是具有相似感知特征的对象也具有相似的感知哈希值。

感知哈希方法计算迅速，占用存储空间小，在检索时，可以通过欧式距离或汉明距离确定图像是否相似，非常适合用于系统资源有限的系统中作为图像检索算法[30][31]。

**[1]** Dukart J R. Content repositories: One or many?[J]. E-doc, 2004(3).

**[2]** 石雪松. 内容管理的真正内涵[J]. 中国经济和信息化, 2003(4):24-24.

**[3]** Winters J. What Is a Content Management System?[J]. Aall Spectrum, 2009.

**[4]** 陈亚南. 网站内容管理系统的设计与实现[D]. 山东大学, 2007.

**[5]** IBM Content Manager. IBM - Content Manager[J]. Ibm Corporation.

**[6]** Managing Enterprise Content: A Unified Content Strategy. Ann Rockley, Pamela Kostur, Steve Manning. New Riders, 2003

**[7]** The content management handbook. Martin White. Facet Publishing, 2005.

**[8]** Content Management Bible, Bob Boiko. John Wiley & Sons, 2005.

[9]刘峻伯, 王昊. 基于Java EE网站内容管理系统的设计与实现[J]. 数字技术与应用, 2016(2):173-173.

[10]杨军. 基于JAVA平台的网站内容管理系统设计与实现[J]. 科技创新与应用, 2016(22).

**[11]** 孔佳, 李昀. 内容管理系统的产生与发展[J]. 农业网络信息, 2008(3).

**[12]** 陈洪波. 基于门户技术的内容管理系统研究[D]. 中国科学院计算技术研究所, 2006.

**[13]** 夏德深. 现代图像处理技术与应用[M]. 东南大学出版社, 1997.

[14] Alvarez C B, Cortez J V, Alonso G R, et al. Face Classification by Local Texture Analisys through CBIR and SURF Points[J]. IEEE Latin America Transactions, 2016, 14(5):2418-2424.

[15] Sanghavi J, Kayande D, Sanghavi J, et al. Content Based Image Retrieval (CBIR) System for Diagnosis of Blood Related Diseases[C]// 2013:11-15.

[16] 马凌蛟. 基于颜色和形状特征的图像检索技术及其应用[D]. 太原理工大学, 2014.

**[17]** 付海燕. 基于图像哈希的大规模图像检索方法研究[D]. 大连理工大学, 2014.

**[18]** Bensaali F, Amira A. Design and Efficient FPGA Implementation of an RGB to YCrCb Color Space Converter Using Distributed Arithmetic[J]. Lecture Notes in Computer Science, 2004, 3203:991-995.

**[19]** 王向阳, 杨红颖, 郑宏亮,等. 基于视觉权值的分块颜色直方图图像检索算法[J]. 自动化学报, 2010, 36(10):1489-1492.

**[20]** 杨红菊, 张艳, 曹付元. 一种基于颜色矩和多尺度纹理特征的彩色图像检索方法[J]. 计算机科学, 2009, 36(09):274-277.

**[21]** 钟智彦, 文志强, 叶德刚. 基于改进协方差矩阵的半色调图像分类研究[J]. 微型机与应用, 2015(19):54-57.

[22]高金金. 基于双线性映射与哈希方法的图像检索研究[D]. 中北大学, 2014.

**[23]** Ojala T, Pietik&#, Inen M, et al. Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2002, 24(7):971-987.

**[24]** 刘梅, 刘伟东, 许荣庆,等. 基于小波变换及傅里叶描述子的图像检索[J]. 系统工程与电子技术, 2003, 25(8):1000-1002.

**[25]** Object Recognition from Local Scale Invariant Features[C]// Proceedings of the International Conference on Computer Vision, Greece. 1999.

[26] 刘兵, 张鸿. 基于卷积神经网络和流形排序的图像检索算法[J]. 计算机应用, 2016, 36(2):531-534.

[27] 杨舒, 王玉德. 基于Contourlet变换和Hu不变矩的图像检索算法[J]. 红外与激光工程, 2014, 43(1):306-310.

[28] 谌湘倩, 马绍惠, 王宏伟. 基于差异空间相关反馈过程的高光谱图像CBIR系统[J]. 计算机应用研究, 2016(7).

**[29]** 牛夏牧, 焦玉华. 感知哈希综述[J]. 电子学报, 2008, 36(7):1405-1411.

[30] Tang J, Li Z, Wang M, et al. Neighborhood Discriminant Hashing for Large-Scale Image Retrieval.[J]. IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society, 2015, 24(9):2827-40.

[31] Zhang X, Liu W, Dundar M, et al. Towards large-scale histopathological image analysis: hashing-based image retrieval.[J]. IEEE Transactions on Medical Imaging, 2015, 34(2):496-506.

**[21]** 袁梅冷, 黄烟波, 黄家林,等. J2EE应用模型中MVC软件体系结构的研究与应用[J]. 计算机应用研究, 2003, 20(3):147-149.

**[22]** Johnson R, Hoeller J, Arendsen A, et al. Professional Java Development with the Spring Framework[J]. Apc, 2007:195-237.

**[23]** 薄奇, 许林英. Spring框架中IoC的实现[J]. 微处理机, 2008, 29(1):147-149.