音频信息检索理论与技术

韩纪庆　郑铁然　郑贵滨　著

科学出版社

音频信息检索理论与技术

韩纪庆　郑铁然　郑贵滨　著

科学出版社
北京

内 容 简 介

本书系统地介绍音频信息检索研究的相关理论、技术与方法，以及该学科领域的最新进展。内容包括音频信息检索的基本理论、表示级和语义级的音频信息检索技术等。在表示级的检索中，重点介绍基于直接特征匹配的音频样例检索方法，内容涉及基于分段的实时检索、基于索引的检索，以及基于硬件实现的快速检索。在语义级的检索中，分别介绍语音文档检索、说话人检索、音乐检索等内容。针对语音文档检索，介绍直接利用语音识别最优候选结果的检索、基于音节网格搜索的检索、基于音节倒排索引的检索、基于后验概率邻接音节矩阵的检索，以及语音文档检索中的容错技术。针对说话人检索，介绍直接利用说话人识别进行检索的方法，以及基于说话人索引的间接检索方法。针对音乐检索，介绍音乐语义信息获取方法——音乐自动标注，以及哼唱检索、拍打检索、基于节拍谱的检索等方法。

本书可作为高等院校计算机应用、信号与信息处理、通信与电子系统等专业及学科的研究生教材，也可供该领域的科研及工程技术人员参考。

图书在版编目(CIP)数据

音频信息检索理论与技术/韩纪庆，郑铁然，郑贵滨著. —北京：科学出版社，2011

ISBN 978-7-03-030372-1

Ⅰ.①音… Ⅱ.①韩…②郑…③郑… Ⅲ.①语言信号处理 Ⅳ.①TN912.3

中国版本图书馆 CIP 数据核字(2011)第 028529 号

责任编辑：张海娜 杨 然 责任校对：李 影
责任印制：赵 博 / 封面设计：耕者设计工作室

科学出版社出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com

源海印刷有限责任公司印刷

科学出版社出版 各地新华书店经销

*

2011 年 3 月第 一 版 开本：B5(720×1000)
2011 年 3 月第一次印刷 印张：16
印数：1—3 000 字数：307 000

定价：48.00 元

(如有印装质量问题，我社负责调换)

前　言

社会的需求是推动理论和技术发展的原动力。近年来，计算技术、网络技术的迅猛发展，使人类积累了大量的多媒体数据。如何有效地利用好这些宝贵的数据资源就显得越来越重要。音频是多媒体数据的重要类型，它广泛存在于互联网和个人计算机中。几乎所有的视频节目中都包含音频内容，此外，还存在大量的歌曲、会议录音、语音通信等纯音频形式的数据。因此，迫切需要能够对音频信息进行有效组织、管理和使用的技术。这些应用需求有力地推动了音频信息检索技术的发展。

理论和技术的积累也为应用的实现提供了可能。对音频信息检索而言，伴随着文本信息检索、语音信号处理、音频信息处理等理论和技术的日臻完善，积累了诸多的经验、沉淀了丰富的理论，这为音频信息检索的研究提供了借鉴和理论上的保证。

纵观音频信息检索研究的发展历史，它是随着两类相关的研究工作的不断深入而逐步发展起来的：一为传统的语音识别研究，二为利用计算机对音乐等进行处理的研究。语音识别研究尽管经过了几十年的努力，并没有完全达到令人满意的程度，但 20 世纪 90 年代在广播新闻语料上的积极探索，为对语音类节目的检索提供了可能。而在基于计算的音乐处理研究方面，早期对音乐结构分析上的工作，也为后来音乐检索的研究奠定了基础。

正是由于音频信息检索具有广阔的应用前景，因此，近年来国内外学者广泛开展了此方面的研究工作。我国也在国家自然科学基金委和科技部等资助的项目中，对此方向的研究给予了大力支持。本书的大部分内容即是作者在承担两项国家自然科学基金项目和一项国家 973 项目课题中，所取得的研究成果的总结和凝练。在此，对国家自然科学基金委和科技部对音频信息检索方面的大力支持表示衷心的感谢！

本书分为五部分、共 11 章，分别介绍音频信息检索的发展与理论基础、表示级的音频检索，以及语义级的语音文档检索、说话人检索和音乐检索。其目的不仅让读者对音频信息检索理论和技术有一个系统的了解，而且努力将本领域的新动态介绍给读者，希望读者能在学术思想上受到启发。书中的第 1、2 章由韩纪庆编写，

第 6～9 章由郑铁然编写，第 3～5、10、11 章由郑贵滨编写。韩纪庆负责全书的总体安排和定稿。

由于音频信息检索研究尚处于不断发展的过程中，有许多理论和技术尚待探讨，加之作者水平有限，疏漏和不足之处在所难免，敬请读者批评指正。

作　者

2011 年 1 月

于哈尔滨工业大学

目　录

第三部分　语义级语音文档检索

第四部分 语义级的说话人检索

第五部分 语义级的音乐检索

第一部分

音频信息检索的发展与理论基础

第1章 绪 论

1.1 信息检索技术的分类及进展

1.1.1 概述

信息检索(information retrieval)技术的历史最早可以追溯到图书资料的手工检索时期,主要应用于图书馆等场所,从大量的图书资料中找到用户所需要的书目。随着现代技术的发展,一方面人类积累的图书、资料、文件越来越多,且多以电子化的形态存在,采用传统的手工查找方法难以满足实际要求,如何有效管理和高效查找相应的内容变得越来越迫切;另一方面计算机技术在信息处理领域的快速发展,也为高效地实现自动信息检索提供了可能,由此产生了现代信息检索技术。

现代信息检索是指针对用户的检索需求,利用一定的检索算法,从结构化或非结构化的数据中获取相关有用信息的过程。这一概念的提出最早可以追溯到1945年Bush的论文[1]。在该文中,作者第一次提出了设计自动的、能在大规模存储数据中进行查找的机器的构想。这篇论著被认为是现代信息检索技术的开山之作。

现实世界中存在着大量的数据文件,它们保存了历史上多种多样的信息。这些文件既有文本类型的,如各种报刊、图书资料和科技文献等,也有音、视频多媒体类型的,如影视节目、音乐、图片等。由于面对的数据对象的类型不同,其所要查找的内容及所采用的方法也有所不同,因此通常可以将信息检索技术分为文本信息检索和多媒体信息检索两大类。

从20世纪40年代信息检索概念的提出,到50年代文本信息检索的逐步兴起,再到90年代蓬勃发展起来的多媒体信息检索技术,时至今日信息检索这一研究方向经历了巨大的变化,从早期基于文本的信息检索发展到当前基于内容的多媒体信息检索,检索源的数据类型越来越复杂,检索策略和技术手段也越来越丰富。下面我们将分别介绍文本信息检索与多媒体信息检索各自的相关概念,以及它们主要的进展情况。

1.1.2 文本信息检索

文本信息检索是指利用一组与查询相关的关键词组成的查询请求来搜索需要

的文本文档，即定位文档中的查询关键词来发现匹配的文档。如果在文档中找到了相关的关键词，即为匹配成功。对于大规模的语料库，任何查询请求都有可能返回数量很多的结果，因此就需要对若干个检索结果进行排序，以便将用户可能最感兴趣的结果排在前面。通常，如果在一个文档中找到了较多的与查询相关的词，则认为该文档比其他有较少相关词或没有相关词的文档更相关。因此，可以将若干个候选文档按所包含相关词的多少降序排列来作为检索结果。

文本信息检索研究始于20世纪50年代，早期代表性的工作是IBM公司Luhn的工作，他提出了利用词对文档构建索引，并利用查询请求与文档中词的匹配程度进行检索的方法[2]，这是经典的文本信息检索中倒排文档检索的雏形。60年代出现了一些优秀的系统及评价指标。在系统方面，Salton开发的SMART系统构建了一个很好的研究平台[3]。在此平台上，研究者可以定义自己的文档相关性测度，以评测和改进检索性能。在评价指标方面，由Cranfield的研究团队组织的评测，提出了许多目前仍然被广泛采用的评价指标。70～80年代，研究人员提出了许多信息检索的理论与模型，并且证明对当时所能获得的数据集是有效的，其中最为著名的是Salton提出的向量空间模型(vector space model，VSM)。该模型至今仍是信息检索领域最常用的模型之一。需要指出的是，当时的研究大多是针对数千篇文档组成的集合，数据的规模偏小。文本数据的缺乏使得当时的技术在海量文本上的可靠性无法得到验证。到了90年代，美国国家标准技术研究院(National Institute of Standards and Technology，NIST)开始组织文本检索会议(Text Retrieval Conference，TREC)。它是一个评测性质的会议，既为研究者公开评测其工作提供了平台，也为参评者提供了大规模的文本语料，从而大大推动了文本信息检索技术的快速发展。第一次TREC会议是在1992年召开的，其后不久，互联网的兴起为文本信息检索技术提供了一个巨大的实验场。在研究人员的不懈努力下，出现了大量实用的文本信息检索系统，并在雅虎、谷歌、百度等著名的互联网搜索引擎中得到了广泛的应用。

目前在文本信息检索领域，简单的信息检索已经开始向更加复杂且人性化的垂直搜索演化，同时更多地引入了信息抽取技术以提取文档中的结构化信息，从而提高检索性能。

1.1.3 多媒体信息检索

随着互联网技术的发展，多媒体的数据量急剧增加。面对海量的多媒体信息，如何有效地组织和管理这些信息，以实现快速准确的检索变得越来越迫切。而伴随着文本信息检索技术的日臻成熟，人们也逐渐把目光更多地投向了多媒体信息检索。早期的工作直接使用了文本信息检索中的方法来实现对多媒体数据文件的检索。其常用的方法是使用文本对图像、音视频等多媒体信息进行标注，标注的文

字是对多媒体语义内容的精练描述。这样通过标注的文字，借助文本信息检索技术就可以进行多媒体信息的检索。由于采用文本描述多媒体信息需要人工标注完成，显然随着多媒体数据的急剧增长，这种人工文本标注的方法代价巨大、难以实现；同时，多媒体文件的人工标注方式存在着主观歧义性问题，同一文件由不同人在不同情况下的标注结果可能存在明显的差异。因此，基于文本标注的多媒体信息检索有其明显的局限性，目前已很少采用。

要有效解决多媒体信息检索问题，就必须从多媒体自身的特点出发，借鉴已有的方法开展相关的研究工作。从方法上看，首先容易想到的是文本信息检索中成熟的技术，但显然基于词匹配的文本信息检索方法不适合直接用于图像及音、视频等多媒体文件的检索，原因在于多媒体文件中不存在确切的“词”，尤其是对那些以音乐为代表的没有语音内容的音频数据。因此，对多媒体信息的检索有其特殊的困难，需要研究特殊的方法。

要实现多媒体信息检索，首先需要进行多媒体信息分割，它包括两层含义[4]：一是多媒体信息中不同类型媒体的分割与剥离，典型的情况是根据多媒体封装格式将其中的音频和视频信息进行分离，以便对每种媒体类型进行单独处理；二是同一类型多媒体信息按语义内容的分割。尽管多媒体信息中不存在明确的关键词，但要实现多媒体信息检索也需要像文本信息检索中进行分词处理一样，将多媒体信息流按照其内容切分为小的语义单位，之后才能进行检索。目前的信息流分割方法大体可分为两类：一类是基于特征值突变检测或信息流局部相似性分析。这类方法可用于分割任意多媒体流数据，但在实际应用中，由于分割对象往往很复杂，并且方法不具有针对性，因而很难取得理想的效果。另一类方法是基于多媒体信息片段的监督分类。由于这类方法具有较强的针对性，因而当应用环境发生较大变化或应用于新的分割问题时，需要重新采集样本数据训练并更新相应的分类器，因而灵活性相对较差。

鉴于多媒体信息具有图像、视频、音频等不同的表现形态，对它们进行检索的研究进展情况也不尽相同，因此下面针对这些不同的表现形态，分别介绍图像信息检索、视频信息检索、音频信息检索的相关情况。

1. 图像信息检索

图像信息检索是指利用图像中包含的颜色、纹理、布局等丰富的特征，通过对这些特征的分析，构建图像数据库的索引，用户在检索时从其查询请求中提取相同的特征，依据某种相似度准则计算查询请求与索引中各个记录的相似性大小，并按相似性降序的方式给出检索结果。

图像信息检索的研究起源于20世纪90年代，1994年美国启动了数字图书馆项目，支持了若干研究计划，有力推动了图像信息检索研究。在此之后，出现了一

些图像信息检索系统[5]，如 IBM 公司开发的最早商业化的 QBIC 系统，它支持基于图像样例的检索，以及草图、轮廓和选定的色彩与纹理样式的检索。该系统中采用的相关技术对后来的图像信息检索系统产生了重要的影响。美国哥伦比亚大学研发的 WebSeek 系统是一种面向 WWW 的文本和图像搜索引擎，其特点是采用了图像区域间的空间关系特征及从压缩域中提取的视觉特征，能同时支持基于视觉特征和空间关系的检索。美国麻省理工学院研发的 Photobook 系统，可分别实现基于形状、纹理和人脸面部特征的检索。

由于图像特征可以从各个方面来描述图像的特点，因此图像信息检索都是基于图像特征来进行的。早期的研究主要以单一视觉特征作为计算相似性的依据，这种单一特征的方法只能描述图像的部分特点，明显有其片面性。为此，后续的研究大都采用多种特征融合的检索方法。以往的图像信息检索通常考虑图像的全局信息，这样检索时计算简单，对图像的平移和旋转不敏感，可以在一定程度上满足用户的检索要求。然而，很多时候用户并非关心图像的全局相似性，而只是关注图像中某一特定的感兴趣区域。此时，使用全局特征将不能很好地突显特定区域的特点。因此，近年来研究者开展了基于感兴趣区域的局部特征来进行图像检索的研究。

图像信息检索在技术上的难点主要表现在[5]，图像的低层视觉特征所代表的视觉信息与图像的高层语义之间存在着较大的差异，这使得检索结果不能很好地满足用户的要求。为此，研究者将相关反馈(relevance feedback)技术引入图像信息检索领域，在检索过程中引入了人的参与，通过用户与检索系统的交互来提高检索系统的性能。在交互过程中，只要求用户根据其信息需求对系统当前的检索结果给出是否相关或者相关程度的信息，系统便可根据用户的反馈进行调整或学习以给出更好的检索结果。

目前图像信息检索研究的热点在如下几个方面：①以往的研究主要集中在低层视觉特征上，而用户对语义的理解要高于低层特征的表达，因此对图像语义特征的表达和基于语义内容的检索是研究的重点之一；②由于图像的规模一般要大于纯粹的文本信息，因此对图像信息检索算法而言，在检索速度和效率上都有很高的要求，高效的检索算法是研究的热点之一；③如何充分考虑不同用户的需求，向其提供简洁友好的界面，也是需要重点研究的问题。

2. 视频信息检索

从图像信息检索到视频信息检索，其明显的差异在于数据从单幅静止图像变成了连续图像帧的形式。应该说，视频是多媒体信息中最复杂的一种，它集图像、声音和文本等于一体，在时间上由连续的一系列帧组成且没有结构，一般是用帧、镜头和场景来加以描述[5]。其中，帧是视频信息的最小单位，对应一幅图像；多个

帧组成镜头对应一段视频，描述的是同一场景中的连续动作；而场景则由多个镜头组成，它是针对同一批对象，通过不断变换拍摄角度而形成的，反映了视频中所蕴涵的高层抽象概念和语义。尽管视频信息具有表现力强、信息量大、形象生动等优点，但其非结构化的数据格式、巨大的数据量等特点，使得对视频信息的检索变得相当困难。

要实现对视频信息的检索首先需要对其进行分析和标注，以便能对视频数据库进行合理的组织及建立有效的索引结构。为了对视频信息进行处理和表征，需要对视频进行分割，将连续的图像帧分割成长短不一的视频镜头，在此基础上实现视频流的表征和相似性测量，即视频的结构化。视频的结构化能够从一段很长的视频中抽象出其内部隐含的情节发展结构，为视频的索引构建和检索提供有效的手段。视频特征数据通常用高维向量表示，因此对视频特征建立索引，就是对高维向量建立索引。由于高维数据自身的无序性和复杂性，使得传统关系数据库中的索引结构不再适用，为此，研究者开展了高维数据索引结构的研究，如典型的树形索引结构与非树形索引结构等。

以往的视频信息检索研究主要集中在两个方面：一是基于视频低层特征的样例检索，它利用用户给出的查询样例，分别提取样例视频和数据库中视频的低层物理特征，并按一定的相似性测度来计算二者之间的相似性，从而得到用户所需的查询结果。由于视频信息是由大量的图像帧组成，如果对每一帧图像都提取特征，显然效率将很低，同时也没有必要。一般是提取代表镜头内容的图像帧，即关键帧来代表镜头的特征。对低层物理特征，主要有关键帧中的颜色、纹理、形状、MPEG-7中定义的视觉特征描述子，以及相邻图像帧之间运动特征的变化信息等。二是基于视频描述信息的语义检索，它通过对视频库中的视频数据进行语义分析以获得高层语义特征，并通过高层语义特征构建视频数据库的索引。检索时提取查询中的同类语义信息，并基于某种相似性测度实现视频内容的检索。其中，高层语义特征主要包括利用文字识别技术提取出的关键帧中的文字符号、利用人脸检测技术提取的人脸特征、利用说话人识别或语音识别技术提取的音频特征等。

类似于图像信息检索，在视频信息检索中研究者也引入了相关的反馈机制，使用户也参与到检索系统中作为检索系统的一部分。这样检索系统按照用户对返回的视频结果所进行的评判来调整检索策略，从而可以提高检索系统的实用性，最终让用户查找到自己所认可的视频信息。

很多国家的科研机构开展了视频信息检索技术的研究，如美国的数字图书馆项目，在视频信息检索方面主要开展视频信息分析、语义信息建模等研究。已有的很多视频信息检索系统都充分利用了视频中包含的图像、语音、文本等媒体信息，构造多种检索模型用于视频信息检索[5]，如美国卡内基梅隆大学开发的 Informedia 数字图书馆项目中，集成了语音、图像和自然语言理解等技术；英国剑桥大学的