◎张 鸿 著

多媒体信息的融合分析与综合检索

科学出版社

多媒体信息的融合分析与综合检索

张 鸿 著

科学出版社

北京

内 容 简 介

本书较系统地讲述了多媒体信息的融合分析与综合检索技术。全书共三个部分，十二个章节。首先从基于文本的信息检索、基于内容的多媒体检索两个方面介绍了信息检索的基础性概念和方法，然后详细阐述了如何从内容特征的角度，实现多媒体数据的结构化表达，并在此基础上，进一步分析了异构的多媒体特征的融合分析方法，以及如何综合各种类型的多媒体信息，实现跨媒体检索。

本书层次分明，内容详实，理论分析与算法实践相结合，力求实用。本书可作为高等院校计算机科学、图书情报等专业的研究生或高年级本科生的技术资料或教学用书，对从事模式识别和多媒体分析等研究、应用和开发的广大科技人员也有很大的参考价值。

图书在版编目(CIP)数据

多媒体信息的融合分析与综合检索/张鸿著. —北京：科学出版社，2011.11

ISBN 978-7-03-032738-3

Ⅰ.①多… Ⅱ.①张… Ⅲ.①多媒体检索系统 Ⅳ.①G354.47

中国版本图书馆 CIP 数据核字(2011)第 229512 号

责任编辑：黄金文/责任校对：梅 莹

责任印制：彭 超/封面设计：苏 波

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

武汉中科兴业印务有限公司印刷

科学出版社发行 各地新华书店经销

*

2011 年 11 月第 一 版 开本：787×1092 1/16

2011 年 11 月第一次印刷 印张：10 1/2

印数：1—1 500 字数：230 000

定价：48.00 元

（如有印装质量问题，我社负责调换）

前　言

随着信息技术的高速发展，文本数据已经不再是人们日常生活中的主要信息来源，图像、音频、视频等不同类型的多媒体信息比比皆是，这些多媒体信息从音、形、意等不同方面绘声绘色地表达了丰富的语义信息，并通过 Web 页面、数字图书馆、多媒体百科全书等载体进行共享，信息世界正变得多姿多彩、栩栩如生。多媒体信息的普及不但使信息世界更加丰富和立体，也使人们可以不受地域和时间的限制，看得更远、听得更多。

多媒体信息在数据量上不断膨胀，并且多媒体数据本身又具有非结构化或半结构化的特点，一般都缺乏有效索引或标注。这就造成了人们望"洋"兴叹，面对浩瀚的信息海量，却难以快速、准确地找到所需要的多媒体信息。为此，在机器视觉、多媒体内容分析和检索、机器学习、统计分析、信号处理、模式识别等多个领域，已经有大量的研究人员长期从事多媒体信息的存储、分析、检索等研究工作，以解决日益增长的信息获取需求，多媒体信息的融合分析与综合检索研究也应运而生。

另一方面，从认知神经心理学的角度来看，人脑对外界事物的认知需要跨越视觉、听觉等不同感官传递的信息，以做出综合判断。那么，用计算机来处理图像、音频、视频等多媒体信息，就好比用人脑来处理视觉、听觉等多种感官传递的信号，只不过计算机目前的智能程度还远远不及人脑的智能。这方面的研究工作也显得极为重要且意义深远，已有的研究成果还较为有限，仍需要有大量的、创新性的研究工作来实现进一步的突破和进展。

作者在国内长期从事多媒体信息的融合分析和综合检索研究，并在这一领域内较早地开展了多媒体异构特征的融合分析和跨媒体检索研究。在这些年的研究过程中发现，该研究领域一直是国内外研究中的热点，许多权威的国际期刊、高级别的国际会议，例如 IEEE Transactions on Multimedia，IEEE Transactions on Pattern Analysis and Machine Intelligence，Pattern Recognition，International Conference on Computer Vision (ICCV)，ACM Multimedia Conference 等，都陆续发表了多篇相关论文，我国自然科学基金委、科技部、教育部也常年资助一些与之相关的科研项目。

目前，大量的研究成果的最新进展是以科技论文或专利的形式进行发布或发表的，然而，能够较为全面地介绍这一领域相关知识的专著还较少，现有的相关著作大多偏重于应用型的多媒体素材制作、数字图像(视频)处理、音频信号处理等，而将多媒体特征分析和检索作为基础性的研究，详细地阐述相关的理论、算法和研究进展的专著还较少。另一方面，考虑到国内每年都有大量的研究生和学者投入这一研究领域，而较为合适的、全面的入门和参考读物还为数不多。因此，作者将在这一领域研究中长期以来积累的知识和经验，总结成为本学术专著，力求从理论上清楚地阐述相关的概念和算法，并且从实践上给予一定的入门指导。

本书是作者长期研究工作的积累，共分为三个部分。

第一部分以信息检索为切入点，介绍了一些必备的基础知识，包括基于文本的信息检索和基于内容的多媒体检索，其中基于文本的信息检索是最早被提出也是应用最为广泛的一种信息检索方式，在多媒体数据的检索应用中也取得了较好的效果；而基于内容的多媒体检索是一种结合了人工智能、机器学习、统计分析等多学科综合性的多媒体数据分析技术，也是本书重要的理论背景。

第二部分介绍了图像、音频、视频三种类型的多媒体信息的结构化表达方法，是本书的理论基础。数据表达是数据分析和检索的前提和基础，而多媒体数据通常是非结构化的，这一部分主要介绍了如何通过对底层特征的分析，计算各种类型的特征模型，实现数据结构化。

第三部分主要从特征的角度，介绍了不同类型多媒体信息的融合分析和跨媒体检索方法。这一部分是对前两部分的理论深入和综合应用，首先介绍了跨媒体检索的基本概念，然后分别从统计分析、线性映射、非线性降维等方面进行了详细阐述，还给出了相关的模拟实验结果。

作者在浙江大学攻读博士期间，曾参与了计算机学院长江学者特聘教授庄越挺教授课题组的国家自然科学基金重点项目“跨媒体海量信息的综合检索与智能技术的研究”，有幸得到了庄越挺教授和吴飞教授的指导，收益颇丰，这些为本书的写作打下了良好的基础。在此，向这两位德艺双馨的教授表示衷心的感谢！

本书是在国家自然科学基金资助项目（项目批准号：61003127）和湖北省教育厅科学技术研究项目（Q20091101）的资助下完成的，在此，表示衷心感谢；此外，还感谢武汉科技大学计算机学院对本书的支持！

由于作者水平有限，时间紧迫，再加上多媒体检索是当前的技术前沿，发展迅速，对书中遗漏之处，敬请读者不吝指正，以便本书日后再版时予以更正。

张鸿

2010 年于武汉

目　录

第一部分　信息检索概述

第一章　基于文本的信息检索

随着人类社会的飞速发展，各种类型的信息资源急剧增长，要从茫茫的信息海洋中快速、有效地找到所需的资源，也变得越来越困难。信息检索在人们的日常生活中变得越来越普及和重要。为了适应信息检索的应用需求，信息检索技术应运而生，并且在这些年中不断地变化和更新。该技术最早是在 1949 年由美国学者 C. Mooers 提出并使用，目前已经涉及到自然语言处理、人工智能、机器学习、多媒体检索技术等多个学科领域。信息检索不仅是学术界一直以来的热门话题，而且已经成功地移植到许多商业应用领域中，如 google、百度等 Web 搜索引擎。

对纯文本信息的检索是人们在信息检索领域迈出的第一步，纯文本信息是指文献、论文、专著等文字性的材料；随后，又出现了基于文本的多媒体检索，也称为基于关键字的多媒体检索。这两种技术都是通过提交查询关键字来检索所需的文本或多媒体信息资源。由于文本信息和多媒体信息存在较大的区别，如文本可以表达直接的语义信息，而图形、音频、视频等多媒体源数据需要人工标注或自动标注后，才便于建立起语义索引，因而，这两种技术存在较大的差别。本章将对信息检索的基本概念、纯文本信息的检索和基于文本的多媒体检索进行介绍。

1.1　基本概念

从广义上讲，信息检索是一种有目的性和组织性的信息存取活动，包括信息的“存储”和“检索”两个部分：前者主要研究如何将各种异构、海量、无序的信息进行有效地组织和存储，存储的内容可以是文献的书目信息、文摘或全文等文本信息，也可以是图像、音频或视频等多媒体信息；后者则是面对用户提出的各种检索需求，快速、准确地查找到相关信息，相应地，检索方式可以是文献的作者、提名、关键词等，也可以是图像的颜色和形状、一段乐曲、一个关键帧等。

基于文本的信息检索在存储和检索两个过程中都包括了文本信息，是信息检索领域的一个重要组成部分，主要可分为纯文本信息检索和基于文本的多媒体检索两个方面，分别介绍如下。

(1) 纯文本信息检索较早被提出，从早期的结构化书目信息检索，到当前的无结构或半结构化的自由文本检索，从关键词检索，到基于概念的语义检索，一直都是较为热门的研究方向。

(2) 基于文本的多媒体检索(也称为基于关键字的多媒体检索)是指：用户提交一个或多个查询关键字，就可以检索到与关键字在语义上相关的图像、音频、视频等多媒体信息。例如，提交关键字“爆炸”，可以找到“爆炸”的图像、音频和视频。从本质上来讲，基于关键字的多媒体检索是对存储的多媒体信息建立文字索引，在检索的过程中，根据用户提

交的关键字与文字索引之间的匹配结果，查找相应的多媒体信息，因此，也属于基于文本的信息检索范畴。

不论是何种类型的信息检索，在系统实现时主要包括四个关键步骤，即预处理、建立索引、查询处理、搜索算法，现分别介绍如下。

(1) 数据的种类和来源各不相同。例如，结构化的书目信息、半结构化的网页数据、非结构化的多媒体数据等等。因此，在建立数据库时需要进行预处理。预处理的主要任务是提取出结构化特征、统一编码转换等。如从网页数据集中提取正文、链接信息，从图像数据集中提取颜色、纹理、形状等特征，并形成结构化的视觉特征向量。

(2) 为了快速找到所需信息，可以对数据集建立索引。例如，用关键词对文档建立索引，根据音频特征对视频片段建立索引等；此外，B+树、TRIE 树、哈希表也都是常用的索引方法。

(3) 用户在提交查询请求时可以有多种方式，包括关键词、自然语言形式表述的语句、布尔表达式等，计算机需要对此进行分析和处理，以更准确地理解用户的查询意图。例如，可以以同义词为依据，对用户的查询请求进行扩展，当用户提交“电脑”作为查询请求时，“计算机”也会作为相关结果返回。

(4) 在上述工作的基础上，搜索算法用于从数据库中找到最相关的信息，并返回给用户，其中相似度的计算和检索结果的排序是两大关键技术问题。例如，基于 Web 链接的 PageRank 算法、基于向量空间的距离计算方法，以及基于语义空间的相似度算法等。

1.2 纯文本信息的检索

1.2.1 技术背景

为了与基于关键字的多媒体检索相区分，本章将对文献、论文、著作、文摘等纯文字信息的检索称为纯文本信息检索。对纯文本信息的检索是人们在信息检索领域迈出的第一步。并且，纯文本检索研究中形成的一些经典算法，后来被成功地移植到了多媒体检索领域。

早期的文本检索研究大多是对整个文本数据库进行分析，将其划分为主题不同的子段，并用关键字进行索引，以支持全文检索。用户可以根据自身的信息需求向文本检索系统提交查询，系统则根据一定的相关性算法，在文本数据库中找出与查询条件相关的文本子集，并按照相关性大小的降序输出。在支持相关反馈的系统中，用户还可以在查询结果中标记相关和不相关文本，并反馈给系统，系统再根据优化算法进行求精和二次检索。

因此，纯文本信息检索的核心问题在于，如何计算数据库中存储的文档与用户提交的查询条件之间的相似度。那么，采用什么样的相似度匹配模型，对检索结果将会有较大的影响呢？

自 20 世纪 60 年来以来，大量的文本检索模型被提出，布尔方法、向量空间、贝叶斯统计方法、概率模型等被引入到文本检索的相似度计算中；之后，随着人工智能研究的发展，产生了用户建模、自然语言处理等技术；机器学习中的一些理论也被应用到文本检索中来，如遗传算法、神经网络、贝叶斯推理等；到了 20 世纪 90 年代，随着网络搜索引擎技术的迅速发展，文本检索被成功应用到网络文本搜索领域。如今，google 和百度等网络文

本搜索技术获得了巨大的成功，甚至融入到了人们的日常生活之中。

1.2.2　经典模型

传统的文本信息检索主要有三种经典模型，即布尔模型、向量模型、概率模型。这些模型对后来的多媒体检索技术产生了重要影响，并且，也被改进和应用到一些多媒体检索系统中。下面分别对这三种模型进行介绍。

1. 布尔模型

布尔模型是基于集合论和布尔代数的一种较为简单的检索模型，将数据库中的文本表示成关键字的集合，且要求用户以布尔表达式的形式，将提交的查询关键字用“与”、“或”、“非”组合起来。例如，$q=w_1 \vee (w_2 \wedge w_3)$，如果数据库中的文本满足表达式 q 时，就作为与查询相关的结果被检索出来。

可见，布尔模型的优点在于清楚、简单，且使用率较高。同时，其缺点在于布尔模型在检索时实现的是二元判定，即相似或不相似，对于用户提交的查询条件，数据库中的文本被简单地分为“相似”和“不相似”两个类别，而无法计算相似度的大小，因此，不利于检索结果的排序，限制了检索功能。此外，在查询条件的表达方面，很多用户难以将检索需求精确地转换成布尔表达式。

2. 向量模型

在该模型中，文本数据以向量的形式进行表达。例如，数据库中的一篇文档表示成一个 m 维的向量 $D=(d_1,d_2,\cdots,d_m)$，其中向量的每一维 d_i 分别代表这篇文档在特征 i 上的权重值。对于文本检索而言，特征可以是字、词、词组或其他文本信息，一般而言，以词作为特征的检索效果最好。因此，通常采用对文本切词，形成文本词集，并将常用的词集合并成为词典，词典中的每个词即作为特征向量的一个维度，从而可以将数据库中的每个文档都表示成向量模型，即：

$$D=(d_1,d_2,\cdots,d_m)=(W_1 \cdot t_1, W_2 \cdot t_2, \cdots, W_m \cdot t_m)$$

其中，$W_i(i \in [1,m])$ 表示词典中的关键词，$W_i \cdot t_i$ 是这个词的权重，m 表示词典中词的数目。目前流行的权重计算方法是基于词频的 TF * IDF(Term Frequency * Inverse Document Frequency)方法，TF 和 IDF 的值可以分别采用下列公式进行计算：

$$\mathrm{TF}_i=\frac{|W_i|}{|D|}$$

$$\mathrm{IDF}_i=\lg(N/df(i))$$

式中，TF_i 表示关键词 W_i 在文档 D 中出现的频率，$|W_i|$ 表示在文档 D 中关键词 W_i 出现的次数，$|D|$ 表示文档 D 中所有关键词的个数；IDF_i 表示文档 D 中关键词 W_i 的倒文本频率，N 表示数据库中文档的总数，$df(i)$ 表示在所有被检索的文档中，包含了关键词 W_i 的文档数目。可见：

· TF 反映了某个关键词在某一篇文档中的重要性，TF 越大，则一篇文档中某个词出现的频率越大，表示这个关键词越能反映文档的内容，与文档主题的关联度也就越大。

· IDF 反映了某个关键词在整个数据库中的重要性，IDF 越大，则出现这个关键词

的文章数目越少，表示该词越特殊。

在基于向量模型的文本检索系统中，查询请求和数据库文档都表示成多维向量，相似度计算则可以通过向量间的距离公式来度量，如欧氏距离、内积距离、余弦距离等，然后，根据相似度大小实现检索结果的自动排序。因此，向量模型是一种代数模型，而布尔模型则是将文档和查询条件用关键词集合来表示，也称为集合论模型。

3. 概率模型

所谓概率模型是指，用户提交的查询条件和数据库中的文档之间采用概率方法计算相似度的值，即文档与查询条件在多大的概率意义上是相似的，概率越大就越相似，查询结果相应地按照概率值递减的次序返回给用户。为了实现概率计算，数据库文档和查询请求都是采用矢量的形式表示，其中的每个分量代表一种特征的取值。查询条件 r 和数据库中某一篇文章 d 之间的相似性概率 $P(r,d)$ 可以依据贝叶斯定理、文档中不同关键词之间的相关性和依赖性，以及特征分布的独立性假设进行计算。

这种方法的优点是以严格的数学理论为依据，并且简单、直观，在检索过程中充分利用了文档特征之间的依赖性和相互关系，缺点在于相应的存储和计算开销较大。

1.3 基于关键字的多媒体检索

由于图像、音频、视频等多媒体数据具有非结构化和半结构化的特点，难以像文本检索那样提取出能够反映语义的关键词。因此，为了能够快速、准确地访问多媒体数据，研究人员在过去的十几年中开展了大量的研究工作，其技术路线主要可分为两类：基于关键字的多媒体检索方法和基于内容的多媒体检索方法。基于关键字的多媒体检索即以用户提交的关键词为查询条件，从数据库中找到语义上相关的各种多媒体信息，包括基于关键字的图像检索、基于关键字的视频检索、基于关键字的音频检索等。

图 1-1 为基于关键字的图像检索的例子，用户在百度图片搜索页面上输入关键字“老虎”，系统返回了与之相关的“老虎”图像。

基于关键字的多媒体检索技术早期受到了文本检索的启发，20 世纪 70 年代末，文本检索技术首次被应用于图像检索中。首先，用人工标注的形式对数据库中的所有图像进行关键字标注，然后，计算用户提交的查询关键字和数据库中的图像标注之间的相似度，并按照相似度大小的降序输出相似图像，形成了基于关键字的图像检索。

在图像的关键字标注过程中，往往会根据图像的采集条件，采用纯手工方式或半人工干预等方式。一般而言，如果采集的图像是独立的，则关键字完全来源于标注者；若图像周围有伴随性文本，则往往采用文本分析技术，从伴随文本中提取关键字作为图像标注，例如：对网页上图像周围的新闻内容进行文本语义分析，得到图像的关键字标注。

这种方法使得检索对象不再局限于单一的文本，而可以是各种类型的多媒体数据，同时，也存在一定的局限性。如图 1-2 所示，不同的人对这幅图像进行标注，很可能会得到不完全一样的结果，并且，标注的详尽程度也不尽相同。

上述局限性可以归纳为以下几个方面。

(1) 所谓“一图胜千言”，图像描述了丰富的语义信息，音频、视频等多媒体数据也同

图 1-1　基于关键字的图像检索例子

可能标注的关键字：丛林、树木、草丛、石头、溪水等

图 1-2　图像的人工标注例子

样具有语义丰富的特点，这就使得人工标注的详尽程度难以统一，从而直接影响了查询的复杂程度。

（2）存在人为理解的主观性和偏差性问题，即使对于同一幅图像或同一段视频，不同的人很可能会做出不同的理解和标注，将对图像标注结果造成极大的影响，甚至造成检索过程中的不精确匹配和错误匹配。

（3）人工标注费时费力，尤其是对于大规模的多媒体数据集，需要花费大量的人力物