网上多媒体信息分析与检索

Web-based Multimedia Information Analysis and Retrieval

庄越挺 潘云鹤 吴 飞 编著

清华大学出版社

网上多媒体信息分析与检索

庄越挺　潘云鹤　吴　飞　编著

清 华 大 学 出 版 社

(京)新登字 158 号

内 容 简 介

本书较系统地讲述了网上多媒体分析与检索技术。全书共 6 部分，分 18 章，分别讲述了基于内容的图像检索、视频结构化与视频检索、基于内容的音频检索、多媒体融合分析与检索、网上多媒体信息检索系统等内容，涉及的媒体类型除文本外，还包括图像、视频、音频及三维图形。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，对广大从事模式识别和多媒体分析等研究、应用和开发的科技人员也有很大的参考价值。

图书在版编目（CIP）数据

网上多媒体信息分析与检索/庄越挺　潘云鹤　吴　飞　编著. —北京：清华大学出版社，2002

ISBN 7-302-05584-X

Ⅰ.网...　Ⅱ.①庄...②潘...③吴...　Ⅲ. ①计算机网络－多媒体－信息－分析

②计算机网络－多媒体－情报检索　Ⅳ.TP202

中国版本图书馆 CIP 数据核字（2002）第 041130 号

出 版 者：清华大学出版社(北京清华大学学研大厦,邮编 100084)

http://www.tup.tsinghua.edu.cn

责任编辑：钟志芳

印 刷 者：北京密云胶印厂

发 行 者：新华书店总店北京发行所

开　　本：787×1092 1/16　**印张**：23.5　**字数**：541 千字

版　　次：2002 年 9 月第 1 版　2002 年 9 月第 1 次印刷

书　　号：ISBN 7-302-05584-X/TP・3299

印　　数：0001～4000

定　　价：37.00 元

前　言

20 世纪人类最伟大的成就之一是发明了计算机，并通过 Internet 使海量的信息自由地驰骋于世界的任何角落。在 Internet 上可获得的信息可谓是浩如烟海，而数字图书馆应用的深入，更使人们足不出户就能阅读世界各国的信息。Internet 使得人们可以在弹指挥手之间不受时间地域的限制获取足够的信息。Internet 所发挥的作用就像古代神话中的千里眼和顺风耳，神奇无比，而今已是现代人掌中的一项平常的工具，它极大地增强了人们对信息的获取与传递能力，从而深刻地改变着每个人的生活、工作与学习！

但是，人类现在面临的问题是如何从信息汪洋中，快速有效地获得所需要的资料。Yahoo 的成功之处，在于为人们在纷杂的互联网中找寻信息提供了一条途径。与 Yahoo 相类似的著名的搜索引擎还有 Google、Infoseek、Lycos、Excite 等，这些搜索引擎都是基于用户输入的关键字进行信息查询的。然而，随着多媒体技术的飞速发展、网络通信能力的提高和计算机处理速度的不断增长，人们认识到 Internet 上的信息除了文本之外，还有大量的图像、视频、音频、动画和图形等，对这些媒体类型的信息进行快速准确的检索已经成为人们的迫切需要。这样，多媒体信息分析与检索的研究应运而生。

应该强调，多媒体信息分析与检索是一项理论与实践紧密结合的技术，它涉及到计算机视觉、模式识别、人机交互、信号处理和认知科学等诸多研究领域。

作者在国内较早从事这一方向的研究。在几年的研究中，发现了这样一个现象：没有对多媒体检索领域进行深入研究的学者，只把信息检索看作一个纯粹应用型的研究方向，即认为多媒体检索只是把计算机视觉、图像（视频）处理、语音识别、信息理论、计算机网络、数据库和人机交互等理论直接拿过来应用而已，而忽略了对多媒体检索自身理论的研究探索。

另一方面，对多媒体检索进行深入研究的学者，会发现多媒体检索不仅只是上述几个研究领域的应用与发展，其自身也存在着相对的独立性。

上述现象使得目前尚没有一本对多媒体检索进行全面阐述的著作。这样，每当有人立志于多媒体方面的理论研究时，却难以找到一本合适的书去引导他们入门。作为教授，在为研究生讲授多媒体检索课程时，要花费很大力气从不同研究领域寻找相关文献。

如上的感受，再加上在多媒体研究方面的多年经验积累，使作者萌发了撰写一本在理论上和实践上将多媒体信息分析与检索涉及的问题讲述清楚的学术专著的念头，而且认为这项工作十分迫切需要。

本书是作者长期研究工作的积累，全书共分为 6 大部分：第 1 部分对多媒体检索研究在每个历史阶段的发展与面临的挑战做了介绍。由于图像、视频和音频是多媒体数据中主要的信息载体形式，本书的第 2 部分、第 3 部分和第 4 部分分别讲述了基于内容的图像检索、基于内容的视频检索、基于内容的音频检索理论。人们常用“眼观六路、耳听八方”形容人通过视觉、听觉等器官来处理数据信息的过程，它也表明人们对事物的认识综合考

虑了多方面信息，缺少任何一个方面的信息，都将使我们或多或少地对事物的认识不全面。因此，在前面三个部分的基础上，第 5 部分介绍了融合视觉和听觉等不同信息进行多媒体检索的有关理论和应用。由于互联网的飞速发展，在数字化信息海洋中构建一个智能网络搜索引擎，是当今网络多媒体计算研究的目标之一，多媒体检索日后的应用平台必然是网络平台。在本书最后，作者对网上多媒体信息检索以及信息检索的未来趋势进行了阐述。

本书附录部分列出了常用多媒体处理工具和与多媒体研究领域有关的著名会议、期刊、研究所和组织情况介绍。

应该讲，在传统文本检索中，用户提交几个关键字后，检索系统可以基于提交的关键字将检索结果反馈回来。这样的检索过程是基于语义和概念的，但是目前多媒体检索离语义检索还有相当长的一段距离，很多多媒体检索原型或是基于视觉（听觉）相似特征，或是基于图像（视频）和视频例子。

不论怎样，通过计算机自动识别并标注多媒体信息，从而使人们能对网上的海量多媒体信息应用自如，是多媒体检索领域每位研究者的心愿，希望这本书能够为这个心愿的早日实现贡献一些力量。

作者之一于 1997 年—1998 年在美国伊利诺斯大学（University of Illinois at Urbana-Champaign，UIUC）留学访问期间，参与了 Thomas S. Huang 教授和 Sharad Mehrotra 教授合作从事的关于多媒体信息分析与检索系统 MARS 等研究项目，收获很大，为本书的写作打下了良好的铺垫。在本书完成之际，谨向他们二位教授表示衷心的感谢，也特别感谢与目前在美国微软研究院工作的芮勇（Rui Yong）博士曾经在 UIUC 的愉快合作。香港城市大学的李青博士对本书的写作给予了很大的支持，在此深表感谢。

课题组的研究生吴翌、杨骏、刘骏伟、赵雪雁、毛祎、郑科等同学为本书的撰写工作提供了很多帮助，在此一并表示感谢。

衷心感谢国家自然科学基金项目（编号：69803009），教育部优秀年轻教师基金，以及教育部博士点基金（No.20010335049）的资助。可以说，没有这些项目的支持，就不会有书中丰富的研究成果，从而也绝没有本书的出版。

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，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。

作者于求是园

2001 年 12 月

目　录

第 1 部分　绪论

第 1 章　基于文本方式的信息检索 4
1.1　布尔模型 6
1.2　聚类模型 6
1.3　向量模型 7
1.4　概率模型 8
第 2 章　基于内容的多媒体检索技术 10
2.1　多媒体分析步骤 11
2.2　多媒体特征提取 12
2.3　多媒体数据流分割 17
2.4　多媒体识别分类 19
第 3 章　WWW 多媒体信息检索 21

第 2 部分　基于内容的图像检索

第 4 章　图像特征的提取与表达 28
4.1　图像颜色特征 29
4.1.1　颜色直方图 29
4.1.2　颜色矩 31
4.1.3　颜色集 32
4.1.4　颜色聚合向量 32
4.1.5　颜色相图 32
4.2　图像纹理特征 33
4.2.1　Tamura 纹理特征 33
4.2.2　自回归纹理模型 34
4.2.3　基于小波变换的纹理特征 35
4.2.4　其他纹理特征 35
4.3　图像形状特征 36
4.3.1　傅立叶形状描述符 36
4.3.2　形状无关矩 37
4.3.3　基于内角的形状特征 38

4.3.4 其他形状特征.....39
4.4 图像空间关系特征.....40
4.4.1 基于图像分割的方法.....40
4.4.2 基于图像子块的方法.....41
4.5 图像高维特征约减和索引.....42
4.5.1 图像高维特征缩减.....42
4.5.2 图像高维特征索引.....43
第5章 图像相似度比较方法.....44
5.1 图像特征相似度比较.....44
5.1.1 欧拉距离.....44
5.1.2 直方图相交.....45
5.1.3 二次式距离.....45
5.1.4 马氏距离.....45
5.1.5 非几何的相似度方法.....46
5.2 图像特征性能评价.....46
5.2.1 颜色特征评价.....47
5.2.2 纹理特征评价.....47
第6章 图像检索中的相关反馈机制.....49
6.1 相关反馈技术分类.....50
6.2 查询向量相关反馈.....51
6.2.1 文本检索中的相关反馈.....52
6.2.2 图像检索相关反馈模型.....53
6.3 特征权重相关反馈.....54
6.3.1 特征权重相关反馈结构.....55
6.3.2 图像特征归一化.....57
6.3.3 图像特征权重调整.....59
6.4 其他图像相关反馈技术.....61
第7章 图像检索的现状和未来.....64
7.1 现有图像检索系统.....64
7.2 图像检索未来发展趋势.....66

第3部分 视频结构化与视频检索

第8章 视频内容结构化.....73
8.1 视频镜头边缘检测.....76
8.1.1 绝对帧间差法.....76
8.1.2 图像像素差法.....77
8.1.3 图像数值差法.....77

8.1.4　颜色直方图法……78
8.1.5　压缩域差法……80
8.1.6　矩不变量法……80
8.1.7　边界跟踪法……81
8.1.8　运动矢量法……81
8.2　镜头边缘阈值确定……82
8.2.1　像素点变化阈值……82
8.2.2　镜头切分阈值……82
8.2.3　镜头渐变阈值……83
8.3　视频关键帧提取……83
8.3.1　基于镜头边界法……84
8.3.2　基于颜色特征法……84
8.3.3　基于运动分析法……85
8.3.4　基于聚类的关键帧提取……85
8.4　视频场景构造……86
8.5　新闻类视频结构化……95
8.5.1　视频新闻内容分析……96
8.5.2　现有新闻类分析系统……101
第 9 章　视频检索和视频反馈……104
9.1　视频检索……104
9.2　视频相关反馈……109
9.2.1　视频层次反馈……110
9.2.2　镜头层次的反馈……111
第 10 章　视频检索技术的现状和未来……113
10.1　视频检索的应用前景……113
10.2　现有视频检索系统……114
10.3　视频检索发展趋势……115

第 4 部分　基于内容的音频检索

第 11 章　音频信号特征提取与表达……122
11.1　音频时域特征提取……123
11.1.1　短时平均能量……124
11.1.2　过零率……125
11.1.3　线性预测系数……125
11.2　音频频域特征提取……126
11.2.1　傅立叶级数……128
11.2.2　复数形式傅立叶级数……129

11.2.3 傅立叶积分与连续频谱.....130
11.2.4 抽样定理.....132
11.2.5 连续信号的滤波与卷积.....132
11.2.6 能谱特征.....133
11.2.7 平均功率与功率谱特征.....134
11.2.8 倒谱特征分析.....134
11.2.9 LPC 倒谱和 Mel 系数.....136
11.2.10 其他频域特征.....138
11.3 音频时频特征提取.....138
11.3.1 短时傅立叶变换.....139
11.3.2 小波变换.....140
11.3.3 连续小波变换.....140
11.3.4 离散小波变换.....142
11.3.5 小波特征系数提取.....143
11.4 音频例子特征提取.....143
第 12 章 音频分割与识别.....145
12.1 音频分割算法.....146
12.1.1 音频分层分割.....147
12.1.2 基于压缩域特征音频分割.....149
12.1.3 基于模板的音频分割.....154
12.2 音频例子识别模型.....156
12.2.1 基于隐马尔可夫链音频例子识别.....159
12.2.2 基于增量支持向量机的音频例子识别.....177
12.2.3 基于最近特征线法的音频例子识别.....190
12.2.4 音频例子混合识别模型.....191
第 13 章 基于内容的音频检索技术.....193
13.1 相似音频例子检索.....196
13.1.1 基于分类模型的音频例子检索.....196
13.1.2 基于模糊聚类音频例子检索与音频相关反馈.....197
13.2 广播新闻结构化.....205
13.3 音乐检索.....209
第 14 章 音频检索的现状与未来.....211
14.1 音频检索的类别.....211
14.2 音频检索未来与挑战.....213

第 5 部分 多媒体融合分析与检索

第 15 章 多媒体融合分析.....221

15.1 多媒质特征融合.....222
15.2 单媒质交叉索引.....224
15.3 单媒质结果融合.....226
第 16 章 多媒体融合检索系统.....229
16.1 文本与视觉信息融合检索.....229
16.2 结合文本和视觉的图像检索与反馈.....232
16.3 基于多模态融合的视频结构化.....235
16.3.1 多模态信息结构化新闻类视频.....235
16.3.2 音频分析技术.....237
16.3.3 视频中的文本分析.....240
16.4 基于压缩域音频特征的足球比赛精彩场景识别提取.....241
16.4.1 足球比赛的声音特征.....242
16.4.2 思路和实现方法.....243
16.4.3 结果分析和比较.....247
16.5 基于支持向量机的视频字幕提取.....249
16.6 基于人脸对象的多媒体内容分析.....254
16.7 基于多模态融合的视频场景分析.....272

第 6 部分 网上多媒体信息检索系统

第 17 章 面向 WWW 多媒体检索系统.....279
17.1 网络信息收集 Web Crawler.....280
17.2 面向 WWW 的多媒体检索系统 Webscope-CBIR.....288
17.3 网络智能检索界面.....291
17.4 个性化 WWW 检索.....295
第 18 章 发展与挑战.....298
18.1 数字化图书馆.....298
18.2 特征维数约减与变换.....303
18.3 三维多媒体检索.....312
18.4 基于关键块的图像检索.....318
18.5 检索复杂性度量.....324
18.6 新一代媒体表示对软件和硬件的影响.....329
18.7 感知界面.....331
18.8 多媒体推理.....333
18.9 结论.....340
附录 1 音频处理工具 HTK.....341
附录 2 多媒体研究领域资料汇总.....347
参考文献.....349

第1部分 绪 论

1951年，商人与学者Calvin Moores首次使用“信息检索（Information Retrieval，IR）”这个单词去描述如下过程：客户提交一个找寻信息的请求，然后通过某种转换或计算，得到与客户请求相似或相关的资料。

他写到，“Information retrieval embraces the intellectual aspects of the description of information and its specification for search, and also whatever systems, techniques, or machines that are employed to carry out the operation.”[167]。

可以看出，对信息检索这个新概念，Moores至少强调了三点：（1）用户对要找寻信息的内容进行高度抽象概括，形成语义描述；（2）使用一个相似度量函数，从信息仓库中得到与用户请求相似的信息集合，并且将它们反馈给用户；（3）用何种系统和何种技术自动实现上面两个目标。其中，第一点就是设计一个检索界面，方便用户的查询；第二点就是将用户请求与所建立的索引进行相似度比较，得到“最相似”的反馈信息；而第三点则是如何保证信息检索系统的实用化。

应该讲，虽然那时提出的信息检索概念仅仅是指对纯文本文件内容进行操作，而没有考虑到信息载体将会是以图像、声音、文字和光电等形式表现的多媒体数据，但是它抓住了信息检索系统就是帮助用户通过检索请求去查找相似信息的本质，这个查找过程要能够顺利完成，需要模式识别、统计理论、人工智能、数据库、网络通信和人机交互等知识的综合应用。

从那时开始，人类能够通过更加广泛的渠道接触到更多信息，每个人接触信息的器官“眼”和“耳”在无形中被延伸，特别是随着现代计算机和通信技术的飞速发展，以超大规模集成电路为核心的自动计算技术、光纤、微波通信和卫星传输使得全球的电脑网络可以相互连接起来，形成了世界上最大的信息存储和传送平台互联网，使人们能够很容易访问到文字、视频（图像）和音频等数字化资源。在人们欢呼雀跃一个巨大的“信息海洋”可以通过键盘和鼠标轻易访问到，为可以通过桌面看世界而激动不已时，人们不禁发现，这些半结构（semi-structure）或无结构化的数字多媒体资源却难以进行基于语义内容的检索，如果只是使用目前纯粹的文本检索工具，人们很难从浩如烟海的信息中找出相似的图像、图形、虚拟三维世界、电影、歌曲和视频镜头，真有望“洋”兴叹的无奈。

“图灵机”宣告了人类对信息进行自动计算时代的开始，“超大规模集成电路”则让知识的自动计算能够走进每个人，“互联网（Internet）”的出现使知识共享并能够快速自动处理。摩尔定律指出：在计算机性能方面，每隔18个月，计算机性能将增长2倍，平均每年增长60%左右，也就是说，每隔10年，计算机性能增加100倍，未来18个月的增长等于以前所有增长的总和（由于计算机性能未来18月的增长与其以前增长无关，所以摩尔

定律表现出无记忆增长曲线）；在计算机数字化信息存储能力方面，现在计算机存储能力也大概按照每 18 个月翻一番的速度在增加，18 个月后计算机新的存储能力等于以往所有存储量的总和。现在可以使用 1G（1Gigabyte=2^{30}字节）空间存储 100~3000 本书，1T（1Terabyte=2^{40}字节）空间存储人一生中所说的任何东西，1P（1Petabyte=2^{50}字节）空间存储人一生中所看见的所有东西。其实，人一生能够接触到的信息大概只有 4T 左右。可以设想，未来技术会提供足够的存储空间来存放数字化进程中所产生的任何信息，并且理论上让这些信息永远“在线”；在数字通信方面，广域网和局域网的传输速度每年平均增长率为 26%~60%，数字化信息传播将更快。

上面是有关数字化信息处理与传播的增长速度，其实，知识本身增长速度将远远高于摩尔定律，在 21 世纪中叶，人类知识将每隔 73 天增长两倍，从而进入一个真正意义上的“知识爆炸”时代！

在这样的条件下，由于传统和现有搜索引擎只能完成基于纯文字的检索任务，使得信息检索面临很大挑战：知识与信息表示形式日趋丰富，如以多媒体形式存在，传统文字信息检索将基本失去用武之地。可以说，每天都会不断更新的庞大数字资源（如美国航空航天局的数字行星项目每天要产生 1000G 新数据）虽然被互联网链接起来，但是由于缺乏有效的网络搜索引擎，这些数字资源显得杂乱无章，难以达到资源共享的目的。人们知道巨大的网络信息海洋中有自己需要的歌曲和电影等信息，却不知道它们到底在哪里，人类利用网络数字化信息资源的能力与产生和传输它们的能力出现了巨大反差。

新技术产生、发展和不断完善的推动力来自于现实生活需要，在人们发出“互联网到底是存储信息的金矿还是埋葬信息的沼泽地”困惑时，多媒体信息检索技术于 1990 年开始蓬勃发展起来，成为网络多媒体计算领域的一个重要分支，其目的就是在分布式网络环境下，为 Web 多媒体数据仓库构建一个快速方便的搜索引擎。

与其他计算机研究方向一样，学科交融和交叉在多媒体检索研究中也表现得十分明显，多媒体检索研究涉及到人工智能、心理学、脑科学、计算机视觉、信号处理、统计方法学、模式识别、数据库、计算机网络、视频通信和人机交互等诸多方面的理论。并且，多媒体信息检索研究的目的就是帮助人们更快捷、更方便和更准确地找到需要的多媒体信息，所以多媒体检索的本质是理论与实践紧密结合的一项研究。

斗转星移，从 1951 年信息检索这个概念的提出，到现在人们更多强调网络多媒体检索的环境，可以发现在不同历史阶段，由于信息来源与信息编码方式不同，人们对信息检索需求有着细微差别。的确，信息检索这个方向的研究随着相关学科的发展，已经经历了很大的发展与变化：从基于文本的信息检索，到现在基于内容的视频（图像）、音频检索和多媒体检索，以及未来面向 WWW 的多媒体检索。

有关学者喜欢用路、车和货物的关系来形容 Internet 实体世界：即互联网这个信息高速公路以光纤、电缆和卫星等通信设备为路，以集电脑、电视和电话为一体的多媒体设备平台为车，以网络上在线的图形、图像、视频、三维虚拟空间和音频等多媒体信息为货物，形成遍布全球的数字资源网。而一个强大的网络搜索引擎就是这个高速公路上面的路标地图，通过它可以方便地找到自己需要的信息。

早期信息检索处理的对象只是纯文本，因为在内存和速度都有限的情况下，计算机能

够对音频和视频信息进行快速处理根本是不可想象的，基于文本的信息检索技术便应运而生，要强调的是，这种基于文本的信息检索技术更多的是对文本信息基于领域进行粗分（如分成娱乐、法律和军事等类别）或提取关键字进行标注，然后按照关键字匹配程度查找相似文本信息。

随着计算机数据处理能力的提高和多媒体编码技术的进步，多媒体逐渐成为人们经常使用的信息载体，于是查找相似的视频和音频成为人们需要，基于内容的视频（图像）、音频和多媒体检索技术在 20 世纪 90 年代开始兴起。多媒体检索技术主要是通过分析多媒体信息中的视觉和音频特征，以达到查找视觉和听觉上相似内容的目的。

网络的发展不仅使之成为信息传送和存放平台，而且也必须考虑网络分布、动态和异构等特性，所以在进行网络多媒体检索时，需要考虑网络计算环境，使网络也成为多媒体检索计算过程的一部分。面向 WWW 的多媒体检索就意味着将多媒体检索技术应用到互联网上，让人们可以方便地检索出网络上分布着的相似多媒体信息。

在绪论这一部分的各个章节，本书将介绍信息检索引擎的概念，历史演变；力求勾划出这个研究领域的发展轨迹以及介绍信息检索未来将面临的挑战。

第 1 章　基于文本方式的信息检索

对相似的纯文本信息进行检索是人们查询信息的第一次努力，虽然那时信息媒质只是纯粹的文本数据，但是对文本信息检索所形成的检索方法，很多被用到日后的多媒体数据检索中。

文本信息检索主要的工作是将用户提交的查找请求与数据库中的信息进行相似度比较，然后将“最相似”的信息反馈给用户，完成信息检索任务。当然，所谓“最相似”的标准由检索系统定义，一般可以将关键字匹配程度作为相似度的衡量标准。

仔细考虑上述过程，会发现如下现象：用户提交的检索需求大多是紧凑的几个单词而已，而被检索出的信息或者是条新闻，或者是篇文章，甚至是本书（多媒体数据还可以是图像、视频、歌曲和三维物体等），于是在将用户查询请求与被检索信息进行“最相似”匹配之前，需要对被检索的信息建立起索引，如提取关键字或形成抽象语义描述等。

根据索引方式的不同，基于文本方式的信息检索大致可以分为三类：（1）对文章里的纯文本内容进行全文分析，形成关键字或文章结构等索引信息。（2）使用文本（一般是关键字）对图像、视频、音频和三维物体等多媒体信息进行标注，标注文字是对多媒体语义内容的精练描述。早期多媒体检索即采用这一方式，就是将多媒体语义内容使用文字信息描述出来，结果多媒体的检索就变成了文本的检索。由于文本描述多媒体数据存在主观性和开销太大等局限性，这种早期基于文本标注的多媒体检索方式已经很少采用了。（3）超文本链接索引方式。超文本链接是 Web 对信息的一种空间组织方式，其实，每个超链接就是对所链接内容的描述，也就是所指向内容的索引。

在这一章中，主要介绍纯文本信息检索理论与方法，多媒体信息文本标注检索与超链接索引可以使用纯文本信息检索的同样方法达到。

为被检索的纯文本信息建立索引，也就是提取文本数据特征，不像视频（图像）和音频数据可以定义诸如颜色、运动和音调等的特征，文本数据的特征就是关键字。只要被检索文本信息提取出了关键字集合，就能够用关键字集合去描述这些文本信息，从而建立索引。基于这个索引结构，通过比较用户查询请求和被检索信息关键字匹配程度，完成相似内容的查找。

在生成索引过程中，如果对全部纯文本信息进行分析，得到关键字索引，那么就叫做全文检索（Full-text Search）；如果只是对小部分纯文本信息进行分析，得到关键字索引，那么叫做非全文检索。所以，文本检索可以分为全文检索和非全文检索两类。

在全文检索出现之前，在线的文本信息数量不是太大太多，况且这些小文本信息比较规范，有明显的题目和摘要，结构清楚。对它们提取特征关键字时，只是分析题目和摘要，然后将得到关键字的集合作为索引，可以完全忽略文章其他部分的内容分析。

随着存储容量和自动处理技术的提高，一篇文章变得越来越长，对于会含有成千上万

不同单词的文本信息，如果只是分析提取其题目和摘要中的关键字特征，显然会导致文本信息丢失，也就是说，题目和摘要的关键字不能反映整个文本信息所介绍的内容。

因为在检索时，为了考察用户提交的检索请求和所检索文本相似度，不仅要知道两者之间存在多少共同的关键字，而且需要知道这些共同关键字各自在被检索文本中出现的次数是多少，它们之间分布如何等情况。只有这样，才能返回较正确的检索结果。

显然，要知道这些全局信息，就要对被检索的全部文本信息进行分析，全文检索技术便开始产生[77]。

值得指出的是，由于被检索的有些文本信息是海量的（如一本大小为 100MB 的英文文章会包含一千七百万左右单词），并且这些数字化文本信息普遍缺少明显结构，提取这些海量信息的关键字或结构等特征时（所谓结构特征是指把整个文本信息分成不同主题类别子段），采用了数据挖掘（Data Mining）或文本挖掘（Text Mining）技术：信息检索中的数据（文本）挖掘主要是为相似度比较建立索引，帮助用户更好找到查询请求所描述的相似信息。而联机分析处理（OLAP）（由关系数据库之父 E.F.Codd 于 1993 年提出）等应用中使用数据挖掘的主要目的是从数据中获取知识和模式，分析数据未来趋势，为决策提供依据。

可以讲，全文检索是真正意义上基于文本的信息检索，其所应用的技术也被后来发展的多媒体检索技术继承和改进。

全文检索通过对整个文本信息的分析，将全部文本划分为主题紧凑的不同子段，用不同的关键字特征标注各个子段，从而为整个文本建立形如 (P_i,F_j,W_j) 的索引结构。其中，P_i 表示第 i 个主题子段，F_j 表示从这个子段中提取的 j 个关键字，W_j 表示每个关键字在子段中出现的频率和分布等特征。

将一个很长的文本信息划分为不同的主题子段方法，是出于这样的观察：当一段文字从一个话题 T_m 转移到另外一个话题 T_n 时，其属于话题 T_m 的关键字会逐渐减少，而属于话题 T_n 的关键字会逐渐增多，这叫做主题转换（Topic Shift）。在整个文本信息中发生主题转换的地方进行切分，就可以把全文分成不同的主题子段，整个文本信息所有的 (P_i,F_j,W_j) 就构成了文章特征向量集合，也就是为文本建立了索引结构。

从文本中提取什么关键字（也就是文本的特征）需要定义，一般来讲，专业词汇或者短语被定义为关键字（如“关键帧”和“刑法条列”，因为看到这两个单词就可以知道它们可能在视频处理和法律文件的文本信息中出现），而“桌子”和“矿泉水”这样的单词就让人不能明确它们会出现在怎样主题的文本信息中。

在机器翻译中，不同主题类别信息中出现的标志性单词按权值大小被集中到不同的语料库（corpus）中，可以有法律语料库，也可以有信息检索语料库。法律语料库中都是表示法律类别的标志性单词，而信息检索语料库中都是表示信息检索内容的标志性单词。这样，可以通过分析哪些语料库中的标志性单词在被检索文本出现、它们出现的频率是多少以及每类标志性单词是聚集出现还是分散出现等特点，将被检索文本切分成不同的主题子段，每个主题子段对应一个或几个语料库，而所提取的每个主题子段特征就是所对应语料库中标志性单词在这个子段中出现的频率和分布等数值。于是，一篇即使很长的文章可以

通过主题转换分成不同部分，然后对每个部分建立索引，完成整篇文章的索引。如果一篇文章被分为 3 个不同的主题域，则这篇文章的索引结构可能就是形如 $\{(P_1,F_1,W_1),(P_2,F_2,W_2),(P_3,F_3,W_3)\}$，其中 $P_i(1\le i\le 3)$ 表示不同主题子段，它们分别对应文本中某些段落，$F_j(1\le j\le 3)$ 表示从每个主题域中提取的关键字集合，$W_j(1\le j\le 3)$ 表示关键字分布和个数等信息。上面这种结构又称为特征向量。

提取了文本信息的特征向量，为文本建立索引结构后，剩下就是如何将用户的查询请求与被检索文本的特征向量进行相似度匹配，找到相似文本信息。也就是说，采用什么衡量方法去评判用户查询请求和被检索文章两者之间的相似性。

采取怎样的相似度匹配算法模型，对检索结果影响很大。在文本检索研究中，主要有布尔、聚类、矢量和概率四类相似度匹配模型，这些模型也对后来多媒体检索匹配模型的产生起了很大影响。

1.1 布尔模型

假设被检索数据库中所有文本信息已经被分成了 s 个主题子段，每个子段索引结构为 (P_i,F_j,W_j)，如果 P_i 提取了 k 个特征（也就是 k 个关键字），那么 F_j 可以表示为 $(f_{j1},\cdots,f_{jk})$，这 k 个特征对应的值为 $(w_{j1},\cdots,w_{jk})$。如果用户查询请求表示为 $\Re(q_1,\cdots,q_l)$，其中 $\Re$ 表示 and、or 或者 not 等逻辑运算，$q_p(1\le i\le l)$ 表示 l 个特征。那么所谓布尔匹配模型的意思是：按照 $\Re$ 形成的逻辑关系，在被检索主题子段中找寻匹配文本。如果 $\Re$ 是 and 运算，表示用户的查询条件是要求特征 $q_p(1\le i\le l)$ 都出现。于是在所有主题子段 (P_i,F_j,W_j) 中，判断每个 $q_p(1\le i\le l)$ 是否出现在 $(f_{j1},\cdots,f_{jk})$ 中，如果所有的特征 $q_p(1\le i\le l)$ 都出现，表示子段 (P_i,F_j,W_j) 满足检索要求，是“相似”信息，则把整个文章作为检索结果反馈给用户；如果在所有主题子段 (P_i,F_j,W_j) 中，每个 $q_p(1\le i\le l)$ 均不出现，表示整篇文章与用户检索请求无关，不满足检索要求，此文章是无关信息。

在这个匹配过程中，还可以根据 $q_p(1\le i\le l)$ 出现在 $(f_{j1},\cdots,f_{jk})$ 中次数多少，对被检索文章进行排序，返回排序结果。

当然，上面的逻辑方式比较简单，“扩展布尔模型（extended boolean）”[66]提供子段分级方式，对简单的布尔检索模型进行了增强。

1.2 聚类模型

聚类检索模型建立在聚类假设的基础上，该假设认为不同主题子段文本信息可以聚集到某几个关键字，使用最少关键字来表示子段文本。

聚类模型大量的工作在于聚类索引结构的生成，一旦生成了聚类索引结构，其相似度匹配将很快完成。

假设对于每个(P_i,F_j,W_j)子段，由于每个子段关键字个数不一样，想通过聚类使每个子段关键字一样多，那么可以应用K平均聚类、模糊聚类和混合高斯聚类等算法完成，由于K远远小于先前每个子段关键字个数，所以聚类算法也起到了特征约减作用。

通过聚类得到的关键字特征，被称为“聚类质心”。这样，每个查询请求和事先形成的子段聚类质心进行匹配。

由于聚类算法主要目的还在于特征约减和质心形成，所以具体的匹配方式可以选择其他任意匹配模型，如距离匹配和布尔匹配等。

1.3　向 量 模 型

在向量模型中，被检索信息和用户提交的检索请求均被表示为V维向量模型（Vector Space Model，VSM），V表示索引数目，每一维其实是每个索引关键字的权重。

如果所有被检索文本集合用D表示，每篇被检索的文章表示为$D_i(D_i \in D)$。为了简单起见，假设D_i只有一个主题子段（即D_i中所有段落都属于一个主题），那么可以如下表示D_i，$D_i=(T_1,T_2,T_3)$。这意味着从D_i中提取了三个关键字，每个关键字用$T_i(1 \le i \le 3)$表示。用户提交的检索请求Q表示为$Q=(T_1,T_3)$，意味着用户想查找与关键字T_1和T_3。

基于上面向量模型，就可以很方便使用关键字是否出现等匹配算法来衡量被检索文本与用户查找请求之间的相似度，来得到相似信息。

只用关键字是否出现来为文本建立索引不能保证检索正确率，后来使用每个关键的权重值来为被检索文本建立索引。仍假设某个文本D_i只有一个主题子段，D_i表示为$D_i=(W_{i1},W_{i2},W_{i3})$，这里$W_{ij}$表示文本$D_i$中第$j$个关键字的权重。

W_{ij}的值可以通过$TF \times IDF$得到，其中TF（Term Frequency）和IDF（Inverse Document Frequency）分别表示术语频率和逆文档频率。文本D_i中第j个关键字术语频率$TF_{ij}=\dfrac{|T_j|}{|D_i|}$，其中$|T_j|$表示在文本$D_i$中关键字$T_j$出现的次数，$|D_i|$表示$D_i$中所有单词的个数；$D_i$中第$j$个关键字的倒文本率$IDF_{ij}=\lg(Num/df(j))$，其中$Num$表示被检索文本总数，$df(j)$表示在所有被检索的文本中，包含了关键字$T_j$的文本数目。

在向量模型中，把查询请求和子段特征都表示成向量，那么相似度比较就转换成计算向量之间距离长短了。如果查询请求向量和某个子段特征向量之间的距离很短，则这个子段和查询请求相似，这是使用向量模型的优势。比如，假设用户查询请求是$Q=(Q_1,Q_2,\cdots,Q_m)$，某个文本D_i第j个子段的向量模型是$(w_{i1},w_{i2},\cdots,w_{im})$，$Q_i(1 \le i \le m)$和$w_{ij}(1 \le j \le m)$分别表示对应关键字通过$TF \times IDF$计算得到的权重。

为了计算用户请求与子段j之间的相似度，实际中可以使用欧拉距离、内积距离和余弦距离等度量。如两个向量之间的余弦相似性计算公式为：