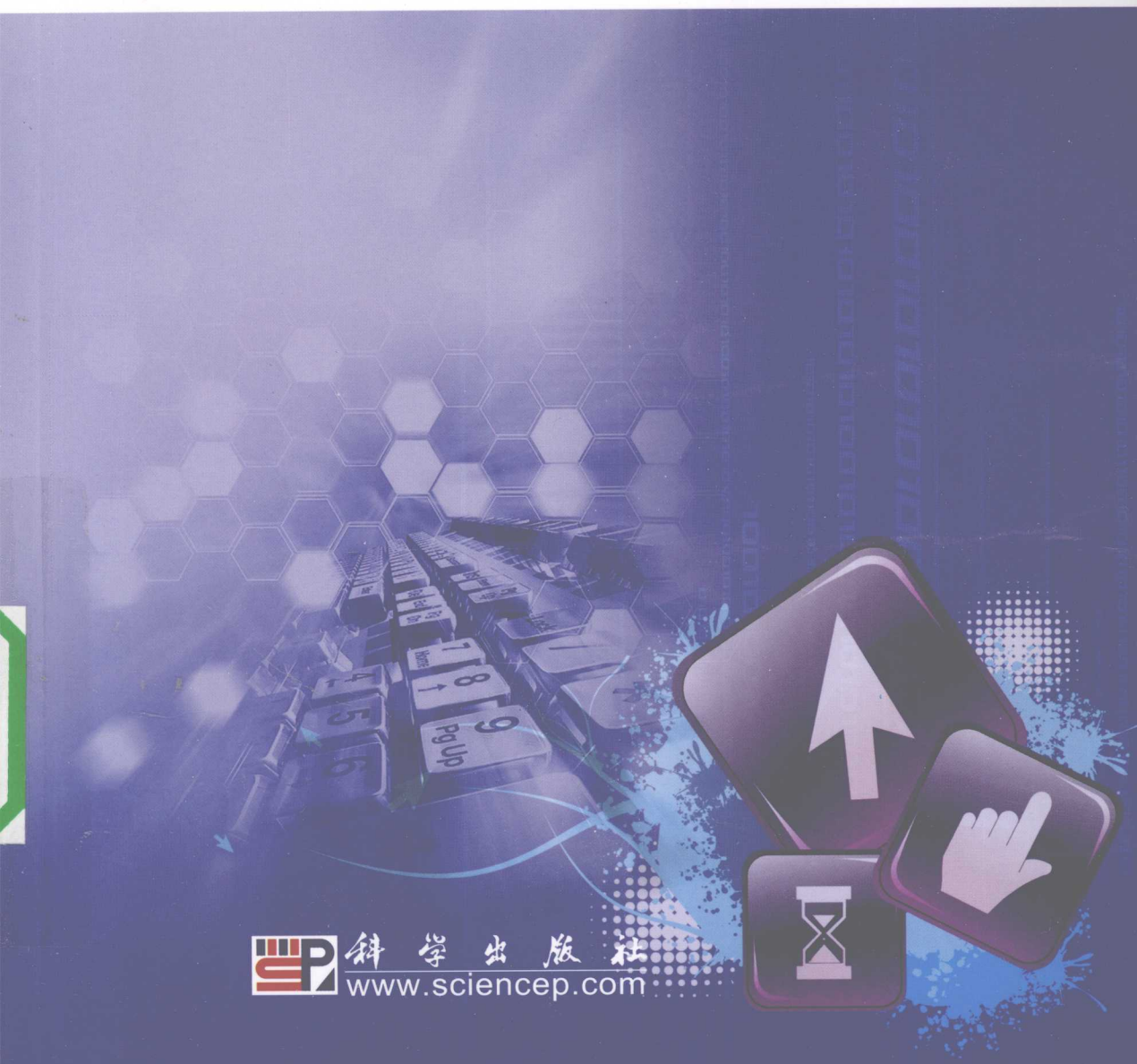
智能科学技术著作丛书

智能检索技术

陆建江　张亚非　徐伟光　苗 壮 编著



科学出版社
www.sciencep.com

智能科学技术著作丛书

智能检索技术

陆建江　张亚非　徐伟光　苗　壮　编著

科学出版社
北京

内 容 简 介

面对海量信息，信息的精确检索就像大海捞针一样困难。智能检索技术吸取多个学科的研究成果，力图通过对文本、图像和视频信息的智能处理，实现信息的精确检索。本书系统地阐述了文本、图像和视频检索的理论方法和实现技术，并重点突出了本领域的最新研究成果。

本书可作为高等院校计算机科学与技术、模式识别与智能系统等学科方向高年级本科生和研究生的教材，也可作为相关领域学生的参考书。

图书在版编目(CIP)数据

智能检索技术/陆建江，张亚非，徐伟光，苗壮编著. —北京：科学出版社，2009

（智能科学技术著作丛书）

ISBN 978-7-03-025328-6

Ⅰ.智… Ⅱ.①陆…②张…③徐…④苗… Ⅲ.计算机网络-情报检索 Ⅳ.G354.4

中国版本图书馆 CIP 数据核字(2009)第 148432 号

责任编辑：张海娜 / 责任校对：钟 洋

责任印制：赵 博 / 封面设计：陈 敬

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

新蕾印刷厂 印刷

科学出版社发行 各地新华书店经销

*

2009 年 8 月第 一 版 开本：B5(720×1000)

2009 年 8 月第一次印刷 印张：16

印数：1—3 000 字数：302 000

定价：48.00 元

（如有印装质量问题，我社负责调换）

《智能科学技术著作丛书》序

“智能”是“信息”的精彩结晶，“智能科学技术”是“信息科学技术”的辉煌篇章，“智能化”是“信息化”发展的新动向、新阶段。

“智能科学技术”(intelligence science&technology，IST)是关于“广义智能”的理论方法和应用技术的综合性科学技术领域，其研究对象包括：

· “自然智能”(natural intelligence，NI)，包括：“人的智能”(human intelligence，HI)及其他“生物智能”(biologicai intelligence，BI)。

· “人工智能”(artificial intelligence，AI)，包括：“机器智能”(machine intelligence，MI)与“智能机器”(intelligent machine，IM)。

· “集成智能”(integrated intelligence，II)，即：“人的智能”与“机器智能”人机互补的集成智能。

· “协同智能”(cooperative intelligence，CI)，指：“个体智能”相互协调共生的群体协同智能。

· “分布智能”(distributed intelligence，DI)，如：广域信息网，分散大系统的分布式智能。

1956年，“人工智能”学科诞生，50年来，在起伏、曲折的科学征途上不断前进、发展，从狭义人工智能走向广义人工智能，从个体人工智能到群体人工智能，从集中式人工智能到分布式人工智能，在理论方法研究和应用技术开发方面都取得了重大进展。如果说，当年“人工智能”学科的诞生是生物科学技术与信息科学技术、系统科学技术的一次成功的结合，那么，可以认为，现在“智能科学技术”领域的兴起是在信息化、网络化时代又一次新的多学科交融。

1981年，“中国人工智能学会”(Chinese Association for Artificial Intelligence，CAAI)正式成立，25年来，从艰苦创业到成长壮大，从学习跟踪到自主研发，团结我国广大学者，在“人工智能”的研究开发及应用方面取得了显著的进展，促进了“智能科学技术”的发展。在华夏文化与东方哲学影响下，我国智能科学技术的研究、开发及应用，在学术思想与科学方法上，具有综合性、整体性、协调性的特色，在理论方法研究与应用技术开发方面，取得了具有创新性、开拓性的成果。“智能化”已成为当前新技术、新产品的发展方向和显著标志。

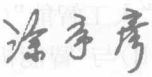
为了适时总结、交流、宣传我国学者在“智能科学技术”领域的研究开发及应用成果，中国人工智能学会与科学出版社合作编辑出版《智能科学技术著作丛书》。需要强调的是，这套丛书将优先出版那些有助于将科学技术转化为生产力以及对社会和国民经济建设有重大作用和应用前景的著作。

我们相信，有广大智能科学技术工作者的积极参与和大力支持，以及编委们的

共同努力，《智能科学技术著作丛书》将为繁荣我国智能科学技术事业、增强自主创新能力、建设创新型国家做出应有的贡献。

祝《智能科学技术著作丛书》出版，特赋贺诗一首：

智能科技领域广
人机集成智能强
群体智能协同好
智能创新更辉煌



中国人工智能学会荣誉理事长

2005 年 12 月 18 日

前　言

面对海量信息，信息的精确检索就像大海捞针一样困难。智能检索技术吸取多个学科的研究成果，力图通过对文本、图像和视频信息的智能处理，实现信息的精确检索。本书系统地阐述了文本、图像和视频检索的理论方法和实现技术，并重点突出了本领域的最新研究成果。

本书涵盖智能检索技术的主要内容，全书共分 14 章：第 1～4 章介绍文本的智能检索技术，包括文本检索技术、文本自动分词、概念语义空间、基于本体的文本检索技术等；第 5～10 章介绍图像的智能检索技术，包括 MPEG-7 标准中图像的视觉特征、图像的局部特征、基于视觉特征的图像检索技术、基于语义的图像检索技术、Web 图像的检索技术等；第 11～14 章介绍视频的智能检索技术，包括视频的结构化技术、语音识别技术、视频的标注技术等。

本书的成果是集体智慧的结晶，由陆建江、张亚非、徐伟光和苗壮负责撰稿。另外，感谢赵天忠、李阳、肖琪、谢正辉、周波、李冉、李言辉、康达周、王进鹏、王家宝、田豫龙等同学为本书撰写工作付出的辛勤工作，这些同学参与了全书的校对工作，在此深表感谢。全书每一章的内容组织和细节都经过多次讨论和修改才定稿，力求深入浅出，让读者轻松掌握相关的知识。尽管每一节、每一句、每篇参考文献，甚至每个标点我们都精心检查，但难免还存在一些缺点和遗漏，殷切希望广大读者批评指正。希望本书的出版能够对智能检索技术相关领域的研究人员有所裨益，并希望通过阅读本书，读者能够很快进行相关领域的研究工作。

目　录

第 1 章　文本检索技术

20 世纪 90 年代，当基于 Web 的超文本应用兴起的时候，越来越多的人开始对文本检索技术感兴趣。随着时间的推移，基于 Web 的信息越来越多，如何在海量的信息中获取自己真正需要的信息成为一个巨大挑战。强大的搜索引擎 Google 能搜索上百亿之多的 Web 页面，主要原因是引擎采用了基于索引的检索技术，减少了搜索的响应时间。

1.1　基于索引的检索技术

在海量信息中获取自己真正需要的信息，顺序搜索的响应时间将变得不可忍受。解决搜索响应时间的办法是对文本文档库中的文本进行预处理，为文本文档库建立一种便于搜索的数据结构——索引。基于索引的检索技术（见图 1.1）非常适用于大规模、稳定的或周期性变化的文本文档库，如今绝大部分搜索引擎采用的都是基于索引的检索技术。

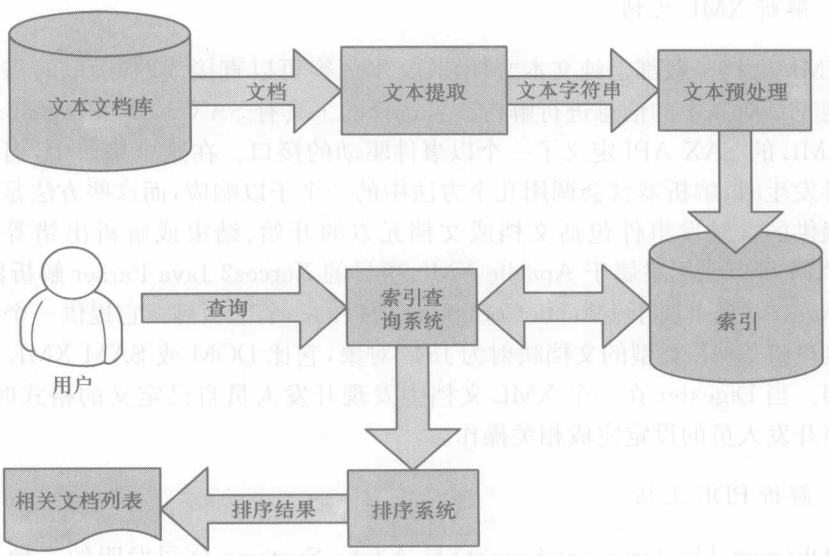


图 1.1　基于索引的检索技术

首先检索系统将所有的检索对象收集起来，构建集中的本地文本文档库。例如对于 Web 搜索引擎，其检索对象主要是 Web 网页，因此搜索引擎需要从互联网

上抓取尽可能多的网页保存到本地文本文档库中，一般这个过程由程序自动完成，本书不关注这个过程的细节，有兴趣的读者可以参考相关资料[1]。本地文本文档库构建完成之后，检索系统提取文本文档库中文档的文本字符串，并进行文本预处理。在有了文本预处理结果后，需要建立文档的索引。利用文档索引可以大大提高信息检索的速度。目前有多种建立文档索引的方法，但是对于大规模的文本文档库来说，用得最多的是倒排索引。在信息库的文档建立索引后，就可以对其进行检索。用户提交查询后，检索系统将直接访问索引。由于索引是一个可以便于搜索的数据结构，检索系统可以通过索引快速获得与查询相关的文档集合。在获取相关的文档后，由排序系统评价相关文档与查询的相关程度并对其排序，最后返回给用户。本章接下来的内容将对基于索引的检索技术的各个模块进行详细介绍。

1.2 文本提取

文本提取过程主要是提取各种格式文档中的字符串。文本检索系统不仅面向互联网的 Web 网页，还面向各种文档类型，例如 XML、PDF、Microsoft Word 或者 Excel 等类型的文档，下面以 XML、PDF 和 Microsoft Word 格式信息为例，介绍如何从多种常用格式的文档中提取文本内容。

1. 解析 XML 文档

XML 文档一般都是纯文本文件，其文本内容可以直接读取，读取时需借助一些其他工具对其中的信息进行解析。可选择的工具有 SAX API 和 Digester。

XML 的 SAX API 定义了一个以事件驱动的接口。在这个接口中，当某个分析事件发生时，解析器就会调用几个方法中的一个予以响应，而这些方法是由调用程序提供的。触发事件包括文档或文档元素的开始、结束或解析出错等。采用 SAX API 可以使用隶属于 Apache XML 项目的 Xerces2 Java Parser 解析器。

Digester 是隶属于 Jakarta Commons 项目的一个子项目。它提供一个简单的上层接口将 XML 类型的文档映射为 Java 对象，它比 DOM 或 SAM XML 解析器更好用。当 Digester 在一个 XML 文档中发现开发人员自己定义的格式时，就可以按照开发人员的设定完成相关操作。

2. 解析 PDF 文档

PDF(portable document format)是 Adobe Systems 公司发明的一种文档格式。这个格式打破了简单的文本内容的局限，它允许作者插入图片、超链接、颜色等。PDFBox 是由 Ben Litchfield 编写的一个免费开源库。使用该开源库，从 PDF 中获取文本内容变得非常简单，只需如下几条 Java 语句即可得到一篇 PDF 文档

中的文本内容：

```
PDFParser parser=new PDFParser(new InputStream(path));
Parser.parse();
CosDocument cos=parser.getDocument();
PDFTextStripper stripper=new PDFTextStripper();
String text=stripper.getText(new PDDocument(cosDoc));
```

3. 解析 Microsoft Word 文档

与其他格式文档不同，Microsoft Word 文档的格式是保密的。Microsoft 公司对这个格式进行了保密处理，这使得其他人很难编写应用程序去读写 Microsoft Word 格式的文档。但是一些开源的项目解决了这个问题，例如 Jakarta POI 和 TextMining.org 项目。

POI 是 Jakarta 的一个子项目，它为操作基于微软的 OLE2 Compound Document 格式的各种文件格式提供了一个 Java API 包，使用 POI 可以从 Microsoft Word 文档中提取文本内容，同样也可以对 Excel 或其他 OLE2 Compound Document 格式的文档进行操作。从 Microsoft Word 中提取文本内容，只需如下几条 Java 语句即可：

```
WordDocument wd=new WordDocument(new InputStream(path));
StringWriter docTextWriter=new StringWriter();
Wd.writeAllText(new PrintWriter(docTextWriter));
docTextWriter.close();
String text=docTextWriter.toString();
```

TextMining 包提供了类似于 Jakarta POI API 的接口，并对 Microsoft Word 文档的接口进行了优化。用 TextMining 只需一条语句即可获得 Word 的文本内容：

```
String t= new WordExtractor().extractText(new InputStream(path));
```

1.3　文本预处理

提取出文本字符串后，还需对文本字符串进行预处理以选择合适的词来建立索引。文本预处理首先将文本中包含的词分析出来，即分词。英文单词天然地被空格隔开，很容易切分，而中文的词是由连续的单个字符组成，因此中文分词相对困难一些，本书将在第 2 章重点介绍中文分词的相关技术。在语义表达方面并不是所有词的表达能力都是同等的，因此除分词之外，文本预处理还包括停用词删除、词干提取、索引词选择和建立词典等操作。

1.3.1 停用词删除

我们知道如果一个词在某个文本中多次出现，那么这个词就很有可能与文本的主题密切相关。然而如果一个词在多个文本中出现，而且频率过高，那么它对文本的区别能力就非常低。一般地，在文档库的文本中出现频率超过80%的词对检索过程根本起不到作用。这部分词被称为停用词(stopword)。在选择构建索引的词时，停用词需要被过滤，以提高索引效率。一般地，冠词、介词、连词等都是停用词，实际检索系统都会设置一个停用词表。

删除停用词可以大大缩小索引空间的大小，一般可以缩小40%左右。删除停用词的缺点是可能会影响检索系统的查准率，有的文本检索系统为了克服这一缺点采用全文索引，并不剔除停用词，对所有的词都建立索引。

1.3.2 词干提取

词干提取是为了解决英文检索中存在的问题而采取的操作。词干是指将词的词缀(前缀和后缀)删除后剩下的部分。例如单词“compete”是它的变形“competes”、“competitor”、“competition”、“competing”和“competed”的词干。在英文检索中，如果用户输入的词是信息库中某个相关文本中词的一种变形，词的变形可以是该词的复数、动名词或者过去分词形式等，那么这些相关文本将被视作与查询无关的文本，这将大大影响召回率。为解决这个问题，在构建索引时，用词干来代替词干的所有变形。

词干提取不仅在很大程度上提高召回率，改善信息检索的性能，同时由于词干的众多变形都由词干代替，用于构建索引的词数量也大大减少，索引空间也进一步缩小。

目前，词干提取技术可以分为：词缀删除、表格查询、后续变形、*N*-连字。词缀删除技术比较直观、简单、有效。在词缀删除中，最重要的就是对词中后缀的删除，因为大多数词的变形是通过后缀来实现的。目前已经有多种关于词缀删除的算法，其中，Porter算法以其简单性和有效性而得到广泛应用。表格查询技术通过在表格中查找某个词的词干来实现，表格中的信息依赖于整个语言中词的词干，因此通常需要相当大的存储空间来存放表格，这就制约了表格查询技术的应用。后续变形技术主要是通过结构化语言的知识来确定词素的边界，这种技术比词缀删除技术复杂。*N*-连字技术判断单词中的字母是否连在一起，这一过程实际上是词条聚类的过程。

1.3.3 索引词选择

如果采用全文索引，那么文档库中的所有词都要建立索引，而对有些语义表达