STIC 情报学论丛

文本信息分析与全文检索技术

化柏林 编著

WENBEN XINXI FENXI YU
QUANWEN JIANSUO JISHU

科学技术文献出版社

文本信息分析与全文检索技术

化柏林　编著

科学技术文献出版社

Scientific and Technical Documents Publishing House

北　京

图书在版编目(CIP)数据

文本信息分析与全文检索技术 / 化柏林编著. -北京:科学技术文献出版社,2008.8
ISBN 978-7-5023-6180-8

Ⅰ.文… Ⅱ.化… Ⅲ.①文字-信息处理 ②情报检索-检索方法 Ⅳ.H087
G252.7

中国版本图书馆 CIP 数据核字(2008)第 158097 号

出　　版　　者　科学技术文献出版社
地　　　　址　北京市复兴路 15 号(中央电视台西侧)/100038
图书编务部电话　(010)51501739
图书发行部电话　(010)51501720,(010)51501722(传真)
邮 购 部 电 话　(010)51501729
网　　　　址　http://www.stdph.com
E-mail: stdph@istic.ac.cn
策 划 编 辑　周国臻
责 任 编 辑　周国臻
责 任 校 对　唐　炜
责 任 出 版　王杰馨
发　　行　　者　科学技术文献出版社发行　全国各地新华书店经销
印　　刷　　者　富华印刷包装有限公司
版 (印) 次　2008 年 8 月第 1 版第 1 次印刷
开　　　　本　787×1092　16 开
字　　　　数　445 千
印　　　　张　20
印　　　　数　1～1500 册
定　　　　价　42.00 元

(京)新登字 130 号

内容简介

本书主要从核心算法、关键技术、技术实例、发展趋势等方面对文本信息分析技术及全文检索技术进行了剖析与探讨。主要内容包括中文分词与语法分析、文献计量分析技术、网络搜索引擎原理与实例、全文检索技术与实例等，对一些前沿专题——信息抽取、自动问答、列表搜索、知识抽取、辅助审稿、对联生成等进行了探讨。

本书内容新颖、观点独特、案例翔实，注重用实践阐释理论，可以作为情报学研究生的教材，也可供情报研究人员和信息搜索企业参考。

前言

1. 标题诠释及内容的组织

文本信息分析技术，就是运用系统分析的思想，借助智能计算模型的支撑，适当涉及语言处理的各个层面，对处理对象（包括数据、信息、知识）进行形式、结构及语义等要素的分析以实现抽取、挖掘、发现、创新等操作。对于每一种操作，分析出其系统架构、关键技术、主要难点、资源支撑、应用前景以及发展趋势等。

信息检索技术是情报技术中的核心，而信息检索技术的核心是全文检索技术，全文检索技术的核心是倒排索引，倒排索引的前提是分词。无论是搜索引擎（如 google、百度）还是全文数据库商（如中国知网、万方数据），都充分利用倒排索引实现全文检索。可是这些系统还存在一些问题，或者说有些方面尚不能满足用户需求，因此，一些特别的检索或相关技术便应运而生，如信息抽取、自动问答等。发现问题、分析原因并试图提出可行的改进方案也是一种研究。

全文检索可以算是文本信息分析的一种应用，但又不仅限于文本信息分析。对于全文检索，除了文本信息之外，还有数据库技术、用户接口等方面。文本信息分析除了应用于全文检索外，还可用于计量分析、文献处理、知识工程等领域。因此不能把文本信息分析归入全文检索，也不能把全文检索归于文本信息分析，它们之间存在内容的交叉与互包含。因此，只能把文本信息分析与全文检索放在一起讨论，而且把全文检索提到书名中，也是为了彰显全文检索的重要性。本书结构安排上包括概论与通用技术、文献计量分析技术、全文检索技术、文本信息分析专题四大部分，具体内容如图 F-1 所示。

第一部分为文本信息分析基础，包括第 1 章至第 3 章。第 1 章是概论，从总体上介绍文本信息分析技术，探讨情报技术的研究范畴、主要内容、分析过程与方法等，确定文本信息分析技术与情报技术的关系。第 2 章与第 3 章介绍文本信息分析的通用处理技术，包括以词法分析为表征的形态分析，与以句法分析为表征的语法分析。

第二部分为结构化信息分析，以文献计量分析为例进行详细探讨，包括第 4 章与第 5 章。第 4 章介绍文献计量分析研究的系统流程与关键技术，第 5 章以图书情报学核心期刊为例展示计量分析报告的撰写。

第三部分为全文检索技术，包括第 6 章至第 8 章。第 6 章介绍搜索引擎工作原理，展示爬行程序实例，并探讨搜索引擎的发展趋势。第 7 章以常用的全文数据库为例，剖析全文检索技术，并展示全文检索系统的实例。第 8 章以信息抽取、自动问答、统计搜索等专题介绍一些新

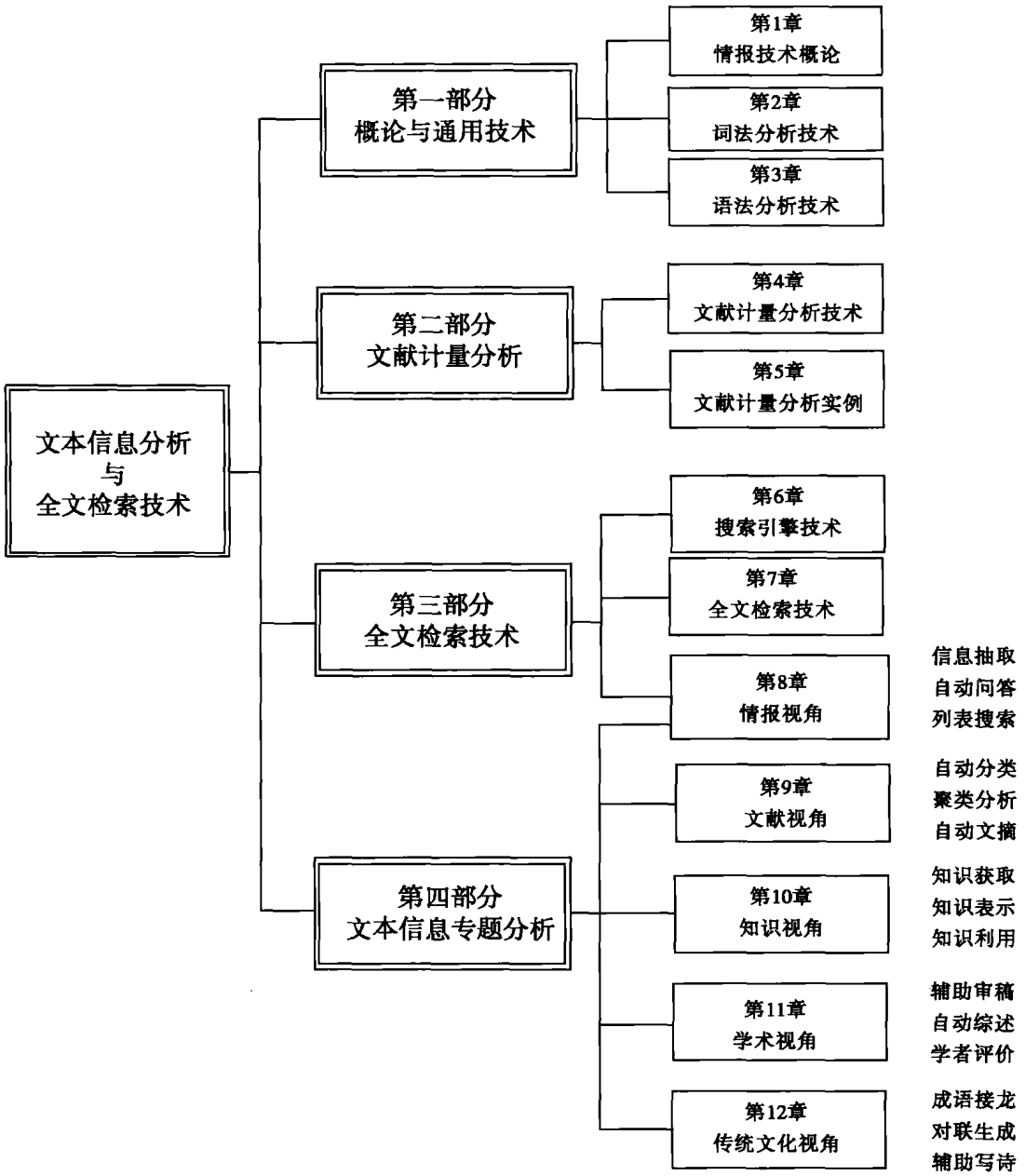

图 F-1　本书内容框架图

型检索及相关技术。这部分所占篇幅虽然不到本书的一半，却是本书的核心内容之一。

第四部分为文本信息分析应用，主要针对结构化信息，分别从情报检索（跨第三部分与第四部分）、信息管理、知识工程、学术问题及传统文化 5 个方面进行论述与探讨，这 5 个方面分别对应着情报学、图书馆学、计算机科学、科学学以及文学等 5 个学科。

2. 内容难易说明

从内容的性质来讲，本书内容分为补充型、学习型、研讨型(图 F-2)。补充型主要为知识做一些铺垫与支撑。学习型内容主要是已经成熟的技术、算法等，是学习的重点。研讨型主要探讨最新进展以及发展趋势，旨在找到研究的切入点以及突破口。

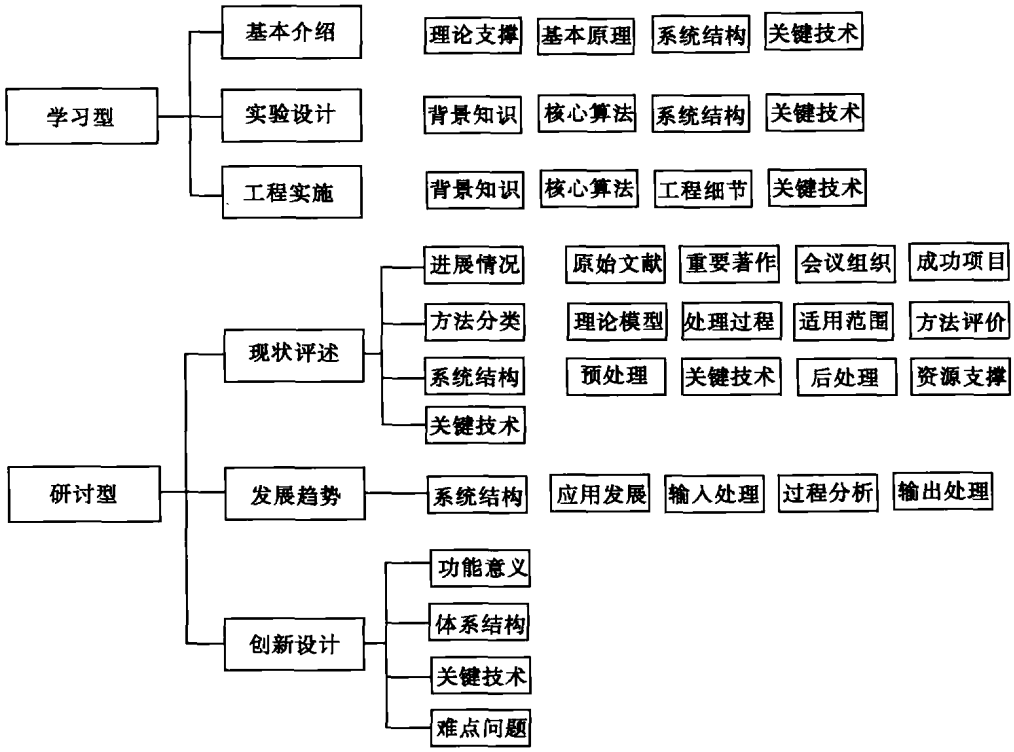

图 F-2 本书内容的属性构成

补充型知识的掌握程度以识记为标准，这部分知识是本书的预备知识与基础条件，主要包括计算机程序设计方面的知识与人工智能基础方面的知识。这些内容不属于本书的论述范畴。程序设计方面的知识包括程序设计语言及开发工具。本书涉及的程序语言包括 VBA、SQL、HTML、JAVA 语言，开发工具包括 Oracle、JDK、Eclipse、Tomcat 等工具。VBA 是 Visual Basic 的一个子集，一般嵌入 office 软件使用，在 excel 中用得最多，更像结构化程序设计语言。Java 是完全面向对象的程序设计语言，网络编程功能较强。HTML，即超文本标记语言，是一种标准的标记语言，主要用于传递与展示网页。SQL，即结构化查询语言，是一种数据库操作语言，可以单独在数据库管理环境中使用，也可以嵌入其他高级语言使用。人工智能方面的知识主要指人工智能的模型与算法，包括向量空间模型、互信息、支持向量机、潜在语义索引、神经网络、遗传算法、决策树、粗糙集、模糊集、隐马尔科夫模型。

学习型知识是本书的重要组成部分，其掌握程度要达到理解与运用。按照介绍程度的不同又分为介绍型、工程型与实验型。介绍型内容主要介绍其核心思想与基本原理，如向量分词算法、搜索引擎工作原理等。工程型指已进行完整的工程开发，成熟完善的功能实现，如文献

信息计量分析。实验型指已进行基本功能实现(亦称小规模实验)部分,通过实验揭示基本原理,如爬行程序、向量分词、倒排索引等。工程型需要考虑各种复杂的情况,包括边界错误等,能够经受得住测试。实验型只是实现基本功能,不考虑性能与健壮性等指标。

研究型内容主要指前沿领域的现状评述以及创新性研究的分析与预测,属于启发学习与自由探讨,包括现状评述型、发展探讨型、创新设计型。现状评述型,主要分析当前研究现状(包括原始文献、会议组织、重要著作、成功项目等),包括专题分类、系统架构、关键技术等。通过对研究的评述旨在找到研究的切入点,如自动问答、信息抽取、自动分类、聚类分析等。发展探讨型即根据当前现状与发展规律提出未来的发展趋势,例如,计量分析的发展趋势、搜索引擎的发展趋势、全文数据库商的发展趋势。创新设计型针对当前问题及需求,创新性地提出一些新的系统,设计其系统结构,剖析关键技术与主要难点,包括知识抽取、列表式搜索、文献自动综述、文献创新性自动评价、对联自动生成系统、辅助写诗系统、学者谱系构建与导师自动评价系统。

3. 创新之处

本书的自写部分包括实验与程序(自己完成但不一定是创新)、想法与创新、发展与展望3类(表F-1)。此外还有案例分析与举例。本书的绝大部分例子属于笔者自行设计,尽量使用专业领域的例子,与本书内容保持一致。

表F-1 本书的实验与程序等内容分布

实验与程序	向量分词程序	第2章第2、第3节
	语法开发平台	第3章第3和第4节
	文献计量分析系统	第4章第1、第2和第3节
	爬行程序	第6章第2节
	全文检索实例	第7章第4节
	学术抄袭与科学引用自动判定系统	第11章第1节
想法与创新	统计搜索	第8章第2节
	自动综述	第11章第2节
	观点搜索	第11章第2节
	基于学位论文致谢的导师评价系统	第11章第3节
	对联自动评测与生成系统	第12章第2节
发展与展望	文献计量分析	第4章第4节
	搜索引擎	第6章第3节
	全文数据库	第7章第5节

(1)实验与程序。本书已进行的实验或系统有向量分词、爬行程序、全文检索、计量分析系统、语法开发平台、学术抄袭与科学引用自动判定系统。中文分词程序在第2章第2和第3节,爬行程序在第6章第2节,全文检索在第7章第4节,计量分析系统在第4章第1、第2和

第3节，语法开发平台在第3章第3节和第4节，学术抄袭与科学引用自动判定系统在第11章第1节。这些内容都通过程序进行实验，而且除学术抄袭与科学引用自动判定系统以外，核心代码皆公布于本书中。

(2)想法与创新。本书就一些问题提出了新的应用，或者说是一些想法。有些内容是受前人或同行启发，有些内容是自己在研究工作中顿悟出来。例如，统计搜索、自动综述、观点搜索、基于学位论文致谢的导师评价系统、对联自动评测与生成系统等，分布在第8章第2节，第11章第1、第2和第3节，第12章第1和第2节。其实，有些内容后来发现已经有人在做了，如成语接龙游戏、文档复制检测软件等，主要缘于笔者视野有限，前期调研不充分，但处理过程与方法并不相同。

(3)发展与展望。本书就某一专题提出发展趋势的有文献计量分析的发展趋势、搜索引擎的发展趋势、全文数据库商的发展趋势，分别位于第4章第4节，第6章第3节，第7章第5节。

本书论述以下十大思想：

1	分类何其难，分类思想无处不在	见第9章第1节
2	检索过程与爬行无关	见第6章第1节
3	检索词短了，结果未必多	见第7章第1节
4	无处不在IPO	见第1章第1节
5	本来无法达到百分之百的事情，那就别期望百分之百	见第2、第8、第10～12章
6	资源的建设比算法更重要	见第6章第3节、第11章第1节、第12章第2节
7	智能系统在受限领域要比开放领域更容易成功	见第8章第8节
8	智能是相对的，知之为知识，不知为智能	见第8～12章
9	任何智能系统归根到底无非知识库与搜索	见第11和第12章
10	万事万物皆检索	见第2、第3、第6～12章

除了着重论述上述思想外，还涉及8组概念的辨析，具体为：

(1)自然语言处理、自然语言理解、文本信息分析、计算语言学之间的辨析。

(2)信息查询、信息检索、信息搜索之间的辨析。

(3)信息检索与信息抽取之间的辨析。

(4)知识抽取与知识获取、知识发现之间的辨析。

(5)数据挖掘与知识发现之间的辨析。

(6)知识管理与知识工程之间的辨析。

(7)自动问答、自动摘要、文献自动综述之间的辨析。

(8)词法分析、语法分析、语义分析、语用分析、形态分析、句法分析之间的辨析。

除此之外，本书涉及4个案例分析。在这4个案例中，有些问题并不新颖，但本书的分析过程及结果说明较为独特，具体如下：

(1)美国情报局破译密码问题,见第 1 章第 3 节。

(2)四人过桥问题,见第 9 章第 1 节。

(3)非相关文献知识发现问题(婆媳关系问题),见第 9 章第 2 节、第 10 章第 3 节。

(4)导师评聘问题(理发师悖论问题),见第 10 章第 5 节。

4. 转述与引用

为了保持内容完整性,除了笔者研究实践的内容以外,还增加了一些内容,这些内容是对他人成果的转述。具体包括信息分析方法与自然语言处理概论,中文分词的未登录词处理、分词消歧、词性标注,句法模型与分析过程,全文数据库与全文索引平台的介绍,面向情报检索、信息管理及知识工程的文本信息分析,分布在第 1～3、第 6～10 章。对于这些内容,笔者在理解内容的基础上,用新的框架重新组织与梳理,在转述他人成果的同时,也试图加入一些自己的认知与看法,并把原有的例子替换成适合本书的例子。具体内容如表 F-2 所示。

表 F-2 本书内容中转述他人成果之创新点

内容	章节	创新点
信息分析方法	第 1 章第 4 节	发展轨迹与实例分析
自然语言处理概论	第 1 章第 5 节	无
中文分词方法	第 2 章第 1 节	按过程分类
分词后处理技术	第 2 章第 4 节	一些举例
词性标注	第 2 章第 5 节	无
语法分析基础理论	第 3 章第 1 节	规则的归一
句法分析过程	第 3 章第 2 节	无
搜索引擎工作原理	第 6 章第 1 节	无
全文数据库与全文索引平台	第 7 章第 1 节	通过实例探析技术内幕
分析标引过程	第 7 章第 2 节	使用专业内容的例子
信息抽取技术	第 8 章第 1 节	一些举例
自动问答系统	第 8 章第 3 节	一些举例
自动分类	第 9 章第 1 节	分类思想
聚类分析	第 9 章第 2 节	联想聚类
自动文摘	第 9 章第 3 节	无
知识的定义与分类	第 10 章第 1 节	一些举例
知识表示	第 10 章第 2 节	无
知识利用	第 10 章第 4 节	知识推理案例及分析

本书除了转述他人成果外,大部分内容皆为个人研究与实践的成果,有些成果发表在一些情报学期刊上,这些期刊包括《情报学报》、《现代图书情报技术》、《情报理论与实践》、《图书情

报工作》、《情报科学》、《情报杂志》等。

特别感谢中国科学技术信息研究所(以下简称中信所)总工程师武夷山研究员对本书就逻辑框架、读者定位、术语规范等问题提出的一系列宝贵意见与建议。特别感谢中信所情报方法中心主任郑彦宁研究员为本书出版提供了大量的帮助与支持。特别感谢中信所情报方法中心副主任张新民副研究员对本书进行全文审稿与修改。感谢北京万方数据股份有限公司王胜海副总工程师、于晓松工程师、宋丽哲博士。

此外,许多情报学专业的博士生、硕士生为本书书稿作了大量工作,包括中国科学院国家科学图书馆博士生王立学、李勇、赵凡、范炜、徐树维先生,吴霞、侯丽、刘兰女士,中国人民大学信息资源管理学院博士生王克平先生,北京大学信息管理系博士生贾佳女士,中信所硕士生甘大广、史豪杰、杨阳先生,还有山东理工大学 2004—2006 级情报学硕士生、中信所 2006—2007 级全日制及在职研究生,对于他们的建议与辛勤工作一并表示感谢! 感谢科学技术文献出版社周国臻编辑,认真高效的编辑为我们弥补了许多漏洞,感谢科学技术文献出版社对本书的出版。

限于作者水平和时间有限,错误疏漏不足之处在所难免,肯请各位专家读者批评指正。

化柏林

2008 年 8 月于北京

目录