信息检索
理论方法及问题分析

XINXI JIANSUO LILUN FANGFA JI WENTI FENXI

王　彪　高光来◎编　著

中国工信出版集团

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

信息检索理论方法及问题分析

王 彪 高光来 编著

電子工業出版社
Publishing House of Electronics Industry
北京 • BEIJING

内 容 简 介

本书围绕信息检索的基本内容，结合当前的研究进展和取得的成果，就信息检索领域的研究内容、理论方法及存在的问题进行阐述和分析，主要包括信息检索的基本内容、信息需求表达、检索模型、文档索引及检索性能评价等方面。

本书适合于对信息检索学习和研究感兴趣的读者阅读参考。

未经许可，不得以任何方式复制或抄袭本书之部分或全部内容。
版权所有，侵权必究。

图书在版编目（CIP）数据

信息检索理论方法及问题分析/王彪，高光来编著. —北京：电子工业出版社，2015.11
ISBN 978-7-121-27437-4

Ⅰ. ①信… Ⅱ. ①王… ②高… Ⅲ. ①情报检索－研究 Ⅳ. ①G252.7

中国版本图书馆 CIP 数据核字（2015）第 249428 号

策划编辑：赵 娜
责任编辑：张 慧
印 刷：**北京季蜂印刷有限公司**
装 订：**北京季蜂印刷有限公司**
出版发行：电子工业出版社
北京市海淀区万寿路 173 信箱 邮编 100036
开 本：720×1 000 1/16印张：10.25 字数：113.22千字
版 次：2015 年 11 月第 1 版
印 次：2015 年 11 月第 1 次印刷
定 价：36.00元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888。

质量投诉请发邮件至 zlts@phei.com.cn 盗版侵权举报请发邮件至 dbqq@phei.com.cn。
服务热线：（010）88258888。

前　言

随着信息时代的不断深入发展，人类对信息有了新的要求，不仅在信息种类和数量上要求越来越多，而且在信息质量上要求越来越高。人类在对衣食住行等基本需求的追求过程中常常伴随着相应的信息需求。在对物质需求逐步满足的基础上，人类对信息的需求往往超过了对其他物质的需求。同样，人类自身的发展越来越依赖于对信息的获取和掌握程度。

信息时代的特点是谁能以最短的时间获取最新的、最有价值的信息，谁就能在激烈的竞争中处于有利地位。而现实情况是，随着信息技术、大数据的不断发展，一方面是日积月累的海量信息，而另一方面是信息获取的困难。

在这种情况下，信息检索理论和技术变得越来越重要了。在大数据时代，信息检索理论与技术面临着新的机遇和挑战。

本书是作者在对信息检索相关理论和应用学习及研究分析的基础上，将一些结果和应用加以汇总、总结和整理而成的。

全书共 7 章，主要内容如下。

第 1 章，信息检索及其主要研究内容。该章主要介绍信息检索的基本概念、主要研究内容，并对信息检索的研究现状和发展趋势，以及大

数据背景下的信息检索进行分析。

第 2 章，信息检索的需求表达。该章介绍需求表达的含义，分析需求表达的难点及建立信息需求域的方法。

第 3 章，信息检索的检索模型。该章主要介绍已有的检索模型、查询扩展及相关反馈的发展情况，讨论需求域基础上的信息检索。

第 4 章，文档索引的建立。该章介绍倒排索引的基本思路和方法。

第 5 章，信息检索系统的评价方法。该章介绍几种常用的评价模型，包括正确率、召回率、F 值指标和平均正确率均值等。

第 6 章，伪相关文档反馈需求域模型信息检索。该章讨论并分析伪相关文档反馈机制下的需求域模型信息检索，分析伪相关文档反馈机制下需求域的特点，介绍相关模型，设计实验，对实验结果进行分析，并评价模型的性能。

第 7 章，用户相关文档反馈需求域模型信息检索。该章介绍并分析用户相关文档反馈机制下的需求域及其检索模型，设计实验，并进行模型训练和实验分析。

需要说明的是，信息检索理论方法极其博深，且在不断丰富发展，本书仅是一些初探。

鉴于作者对该领域的浅薄认识及自身知识的局限性，错误和不当之处在所难免，敬请广大同仁不吝批评、指正。

编著者

2015 年 10 月

此为试读，需要完整PDF请访问：www.ertongbook.com

目　录

第1章

信息检索及其主要研究内容 1

1.1 信息检索 ……3

1.1.1 信息检索的基本概念 ……3

1.1.2 信息检索的研究内容 ……3

1.1.3 研究现状和发展趋势 ……4

1.1.4 结构化、半结构化和非结构化信息 ……5

1.2 大数据背景下的信息检索……6

参考文献 7

第2章

信息检索的需求表达 11

2.1 需求表达 ……13

2.2 需求表达的主要理论方法……13

2.3 需求表达存在的主要问题分析……14

2.4 信息需求域 ……15

2.4.1 机器信息检索：用关键词匹配近似语义匹配 ……15

2.4.2 文档、句子及词语之间的语义关系 ……15
2.4.3 信息需求域 ……18
2.4.4 信息需求域的理论推导 ……22
2.4.5 信息需求域的子域、近似域 ……24
2.4.6 查询请求与信息需求的关系 ……26
2.4.7 信息需求域的理论意义 ……29
2.4.8 信息需求域的一种粗糙集解释 ……29
2.5 小结与讨论 ……33
参考文献 34

第3章
信息检索的检索模型 37
3.1 信息检索的主要检索模型……39
3.2 查询扩展、相关反馈研究现状……42
3.3 检索存在的主要问题分析……43
3.4 信息需求域基础上的信息检索……45
3.4.1 信息需求域的结构 ……45
3.4.2 文档相似度的定义 ……50
3.5 检索模型的发展方向分析……59
参考文献 60

第4章

文档索引的建立 67

4.1 附加统计信息的倒排索引……69

4.2 停用词……71

4.3 词干提取……71

4.4 词形归并……72

4.5 小结与讨论……73

参考文献 73

第5章

信息检索系统的评价方法 75

5.1 测试集……77

5.2 无序检索结果的评价……79

5.3 排序检索结果的评价……80

5.4 小结与讨论……82

参考文献 82

第6章

伪相关文档反馈需求域模型信息检索 85

6.1 伪相关文档反馈机制……87

6.2 需求域去噪 ……87
6.3 伪相关文档反馈机制的模型分析……89
6.3.1 去噪性能分析与实验 ……91
6.3.2 去噪参数 β 的取值分析与实验 ……95
6.3.3 参数 α 的取值分析与实验 ……99
6.3.4 伪相关反馈文档数目及稳定性分析与实验 ……101
6.4 伪相关文档反馈机制下的需求域模型结论……103
6.4.1 需求域模型结论 ……104
6.4.2 检索性能对比实验分析 ……106
6.5 小结与讨论 ……111
参考文献 112
本章附录 112

第 7 章
用户相关文档反馈需求域模型信息检索 117
7.1 用户相关文档反馈机制……119
7.2 用户相关文档反馈机制下的模型分析……120
7.2.1 用户相关文档反馈下的上界优化分析与实验 ……121
7.2.2 优化参数 β 的取值分析与实验 ……124
7.2.3 参数 α 的取值分析与实验 ……127

7.2.4 相关反馈文档数目及稳定性的分析与实验 ……130
7.3 用户相关文档反馈机制下的需求域模型结论……133
7.3.1 需求域模型结论 ……133
7.3.2 检索性能对比实验分析 ……135
7.4 需求域模型计算性能分析……139
7.5 小结与讨论 ……140

全书参考文献 143

第 1 章

信息检索及其主要研究内容

1.1 信息检索

1.2 大数据背景下的信息检索

此为试读，需要完整PDF请访问：www.ertongbook.com

1.1　信息检索

1.1.1　信息检索的基本概念

信息检索的含义及内容非常广泛。例如，图书馆管理员帮助读者从图书馆的书架上找到一本书，这就是一种信息检索，是人工形式的信息检索；计算机从银行数据库中找到某个客户账户的信息，这也是一种信息检索，是机器形式的信息检索。现在人们所研究的信息检索（Information Retrieval, IR）主要是指利用计算机，根据用户提出的查询请求（query），从存储在计算机中的大规模非结构化数据集中，如文本文档集（Collection D），查找到用户所需要的信息资料（若干个文档），并自动将查找结果（Result）反馈给用户的过程。

1.1.2　信息检索的研究内容

信息检索主要完成三个方面的任务：信息需求的表达方法、信息存储方法和检索方法。相应地，信息检索研究主要有三个方面：查询表达、信息表达和检索理论与方法。其中，检索主要指检索模型（Retrieval Model）。信息检索的基本过程如图 1.1 所示。

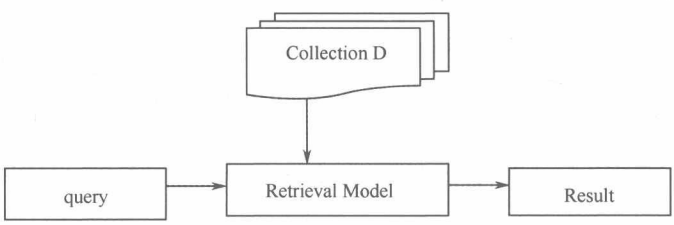

图 1.1　信息检索的基本过程

按照检索的不同内容，信息检索分为文本检索、图形图像检索、声音检索、视频检索等。它们的检索理论与方法既有相同之处又有区别之处。

随着信息检索的不断发展和应用，检索结果的呈现也显得越来越重要。通常，用户希望将检索到的内容以可视化、直观化、美观化的形式展现出来。因此，检索结果的呈现也日益成为信息检索的研究内容之一。

1.1.3 研究现状和发展趋势

需求推动研究、创新和发展。可以说，自从人类有了信息开始，就有了信息检索的需求。至今，信息检索经历了人工检索、机械检索、计算机检索三个发展阶段。

计算机信息检索始于 20 世纪 40 年代[1-2]。1950 年，信息检索先驱，美国人 Calvin N. Mooers 首次提出了信息检索的概念[3]。1959 年，Calvin N. Mooers 提出了穆尔斯定律[4]：当拥有信息比不拥有信息会让用户付出更大的努力或给用户造成更大的麻烦时，用户会倾向于不使用信息检索系统。该定律既表达了计算机信息检索系统效率的重要性，也从侧面反映了机器信息检索系统实现的难度。

当今，人类社会已经发展并进入到信息化、网络化阶段。人类的生产、生活日益高度依赖于信息。诸如 Web、博客、微信、数字图书馆、电子商务、企业网站、网上股票、网上银行等，都是信息的来源。信息的种类和数量以惊人的速度不断地增长，与此形成鲜明对比的是信息获取的手段和效率日益相对滞后。信息处理技术迫切需要更有效的理论和

方法来处理如此海量的信息，特别是如何从如此海量的信息中获取用户所需的信息。随着人类社会的日益进步，信息获取已经关系到人类生产、生活、学习等质量的提高。

顺应这样的需求，信息检索成为当前信息处理研究领域中的研究热点，布尔模型、向量空间模型、概率模型、统计语言模型、基于机器学习的检索模型等模型被先后提出并取得了一定的应用效果。百度、Google 等一些成功案例已经出现。但是，总的来讲，当前已有的信息检索理论与方法远未满足人们的需要。因此，信息检索是当前以及未来一定时期内信息处理研究领域中的研究热点，各种新的检索理论方法将不断涌现。

1.1.4　结构化、半结构化和非结构化信息

结构化信息指的是这类信息的各个组成部分的语义都是明确的，各个组成部分之间的关系也是明确的。结构化信息处理的主要方法是使用数据库技术，结构化信息的检索理论与方法主要也是基于数据库的。基于数据库的结构化信息检索理论与技术相对已经成熟，主要是 SQL 技术。参考文献[5]从数据库的角度出发介绍了结构化文本检索。参考文献[6]详述了 SQL 技术。

半结构化信息指的是这类信息的一部分组成内容的语义是明确的，而另一部分组成内容的语义是不明确的。半结构化信息的典型代表是 HTML 网页。较早的半结构化信息检索见参考文献[7]。XML 是半结构化信息检索的基础，参考文献[8]、[9]是关于 XML 的综述。向量空间的

XML 检索见参考文献[10]、[11]，语言模型见参考文献[12]～[14]。参考文献[15]介绍了基于概率权重的计算机制。

非结构化信息指的是这类信息的内容在结构上一般没有进行语义上的划分，没有清楚的语义结构。非结构化信息分为图形图像信息、语音信息及文本信息等类型。

随着网络技术的不断发展，网络用户越来越多，网络应用越来越广泛，特别是 Internet 和 Intranet 技术，使得非结构化信息占全部信息的比例越来越大，绝对数量也日益增加，对于非结构化信息检索的需求越来越迫切。同时，非结构化信息检索也是当前整个信息检索研究中的难点和热点。

1.2 大数据背景下的信息检索

必须注意的是，随着大数据时代的到来，信息检索面临着新的挑战和机遇见参考文献[16]～[23]。大数据下的信息检索不仅只是从数据集中找到与用户需求相关的信息资料，更重要的是要找到经过分析和加工整理后的信息。例如，一位初学信息检索的用户想查找信息检索的概念的相关资料，基于不同的检索环境将出现不同的检索结果，如下所示。

百度检索：

查询请求：信息检索的概念。检索结果：13800000 个。

查询请求: 什么是信息检索。检索结果：58900000 个。

Google 检索：

查询请求: concept of Information Retrieval。检索结果：12900000 个。

查询请求: what is Information Retrieval。检索结果：14700000 个。

百度学术：

查询请求:信息检索的概念。检索结果：40700 个。

查询请求: 什么是信息检索。检索结果：343000 个。

Google 学术。

查询请求: concept of Information Retrieval。检索结果：3430000 个。

查询请求: what is Information Retrieval。检索结果：3070000 个。

上述检索结果往往出乎用户意料：（1）不需要如此多的资料；（2）在如此多的资料中，哪些是所需要的资料。

面对大数据，信息检索面临的机遇和挑战：（1）能否找出有价值的若干资料；（2）能否经过分析整理后仅生成一份关于问题的最终资料。

参 考 文 献

[1] R. Baeza-Yates，B. Ribeiro-Neto. Modern Information Retrieval: The Concepts and Technology behind Search. 2nd ed. Addison Wesley，2011.

[2] Liddy Elizabeth D. Automatic document retrieval. In Encyclopedia of Language and Linguistics. 2nd ed. Elsevier，2005.

[3] Mooers Calvin E. Coding，information retrieval and the rapid selector. American Documentation，1950，1（4）:225-229.

此为试读,需要完整PDF请访问: www.ertongbook.com