普通高等院校数据科学与大数据技术专业“十三五”规划教材

智能搜索引擎技术

ZHINENG SOUSUO YINQING JISHU

高琰 ⊙ 编著

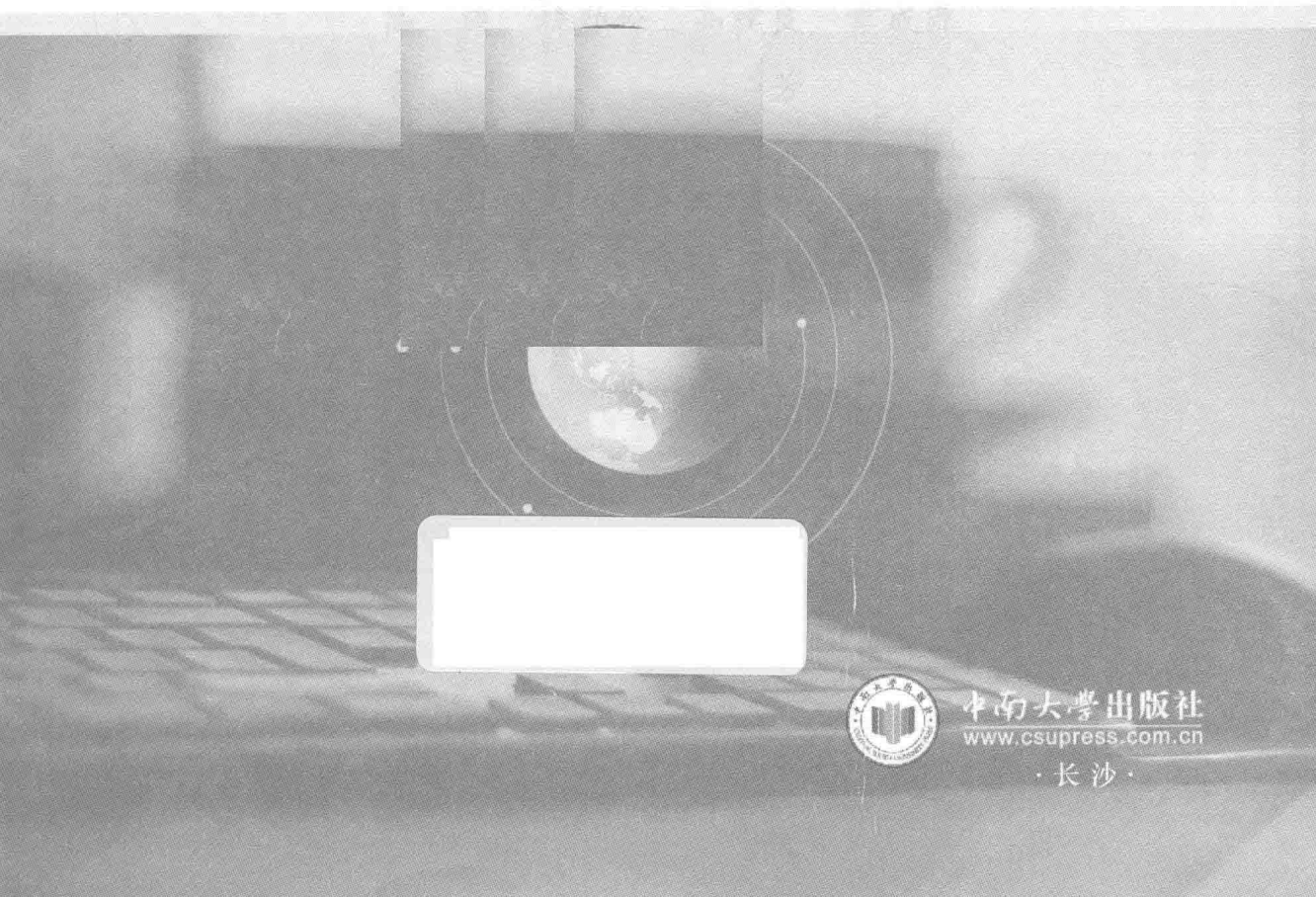
中南大学出版社
www.csupress.com.cn

普通高等院校数据科学与大数据技术专业“十三五”规划教材

智能搜索引擎技术

ZHINENG SOUSUO YINQING JISHU

高琰 ⊙ 编著



中南大学出版社
www.csupress.com.cn
·长沙·

图书在版编目（CIP）数据

智能搜索引擎技术 / 高琰编著. --长沙：中南大学出版社，2018.12

ISBN 978-7-5487-3412-3

Ⅰ.①智… Ⅱ.①高… Ⅲ.①搜索引擎—程序设计 Ⅳ.①TP391.3

中国版本图书馆 CIP 数据核字(2018)第 213377 号

智能搜索引擎技术

高 琰 编著

□**责任编辑** 韩 雪
□**责任印制** 易建国
□**出版发行** 中南大学出版社
社址：长沙市麓山南路 邮编：410083
发行科电话：0731-88876770 传真：0731-88710482
□**印 装** 长沙印通印刷有限公司

□**开 本** 787×1092 1/16 □**印张** 12.5 □**字数** 314 千字
□**版 次** 2018 年 12 月第 1 版 □**印次** 2018 年 12 月第 1 次印刷
□**书 号** ISBN 978-7-5487-3412-3
□**定 价** 35.00 元

普通高等院校数据科学与大数据技术专业“十三五”规划教材

编委会

总序

Preface

随着移动互联网的兴起，全球数据呈爆炸性增长，目前90%以上的数据是近年产生的，数据规模大约每两年翻一番；而随着人工智能下物联网生态圈的形成，数据的采集、存储及分析处理、融合共享等技术需求都能得到响应，各行各业都在体验大数据带来的革命，“大数据时代”真正来临。这是一个产生大数据的时代，更是需要大数据力量的时代。

大数据具有体量巨大、速度极快、类型众多、价值巨大的特点，对数据从产生、分析到利用提出了前所未有的新要求。高等教育只有转变观念，更新方法与手段，寻求变革与突破，才能在大数据与人工智能的信息大潮面前立于不败之地。据预测，中国近年来大数据相关人才缺口达200万人，全世界相关人才缺口更超过1000万人之多。我国教育部门为了响应社会发展需要，率先于2016年开始正式开设“数据科学与大数据技术”本科专业及“大数据技术与应用”专科专业，近几年，全国形成了申报与建设大数据相关专业的热潮。随着专业建设的深入，大家发现一个共同的难题：没有成系列的大数据相关教材。

中南大学作为首批申报大数据专业的学校，2015年在我校计算机科学与技术专业设立大数据方向时，信息科学与工程学院院领导便意识到系列教材缺失的严重问题，因此院领导规划由课程团队在教学的同时积累素材，形成面向大数据专业知识体系与能力体系、老师自己愿意用、同学觉得买得值、关联性强的系列教材。经过两年的准备，针对2017年《教育部办公厅关于推荐新工科研究与实践项目的通知》的精神，中南大学出版社组织对系列教材文稿进行相应的打磨，最终于2018年底出版“高等院校数据科学与大数据技术专业‘十三五’规划教材”。

该套系列教材具有如下特点：

1. 本套教材主要参照“数据科学与大数据技术”本科专业的培养方案，综合考虑专业的来源，如从计算机类专业、数学统计类专业以及经济类专业发展而来；同时适当兼顾了专科类偏向实际应用的特点。

2. 注重理论联系实际，注重能力培养。该系列教材中既有理论教材也有配套的实践教程。力图通过理论或原理教学、案例教学、课堂讨论、课程实验与实训实习等多个环节，训练学生掌握知识、运用知识分析并解决实际问题的能力，以满足学生今后就业或科研的需求；同时兼顾“全国工程教育专业认证”对学生基本能力的培养要求与复杂问题求解能力的

要求。

3. 在规范教材编写体例的同时，注重写作风格的灵活性。本套系列教材中每本书的内容都由教学目的、本章小结、思考题或练习题、实验要求等组成。每本教材都配有 PPT 电子教案及相关的电子资源，如实验要求及 DEMO、配套的实验资源管理与服务平台等。本套系列教材的文本层次分明、逻辑性强、概念清晰、图文并茂、表达准确、可读性强，同时相关配套电子资源与教材的相关性强，形成了新媒体式的立体型系列教材。

4. 响应了教育部“新工科”研究与实践项目的要求。本套教材从专业导论课开始设立相关的实验环节，作为知识主线与技术主线把相关课程串接起来，力争让学生尽早具有培养自己动手能力的意识、综合利用各种技术与平台的能力。同时为了避免新技术发展太快、教材纸质文字内容容易过时的问题，在相关技术及平台的叙述与实践中，融合了网络电子资源容易更新的特点，使新技术保持时效性。

5. 本套丛书配有丰富的多媒体教学资源，将扩展知识、习题解析思路等内容做成二维码放在书中，丰富了教材内容，增强了教学互动，增加了学生的学习积极性与主动性。

本套丛书吸纳了数据科学与大数据技术教育工作者多年的教学与科研成果，凝聚了作者们的辛勤劳动，同时也得到了中南大学等院校领导和专家的大力支持。我相信本套教材的出版，对我国数据科学与大数据技术专业本科、专科教学质量的提高将有很好的促进作用。

桂卫华

2018 年 11 月

前言

Foreword

随着信息技术的快速发展和互联网的广泛应用，Web 已经成为了一个巨大的、分布广泛的全球信息服务中心，发布着新闻、财经、文化、教育等各种海量信息。如何在互联网的海量信息资源中快速准确地定位所需的信息，已经成为人们的迫切需求。搜索引擎是大数据时代下对互联网中的海量信息进行检索的关键技术。并且随着互联网中信息资源的日益快速增长，传统的搜索引擎技术开始向智能化方向发展，为人们提供更精准、更个性化的服务。

本书以当前搜索引擎主流技术为基础，密切关注前沿技术发展趋势，结合当前人工智能和自然语言技术的发展，以深入浅出的形式介绍一套完整的大数据时代背景下的智能搜索引擎的关键技术。本书在吸取国内外经典教材优点的基础上，广泛搜集合适的实例，通过实例从多个视角对智能搜索引擎的核心技术进行全面介绍，加深读者对关键概念和核心技术的理解。本书还对开源软件进行了介绍，将技术理论与应用范例结合。

本书共分为 10 章，通过采用循序渐进的组织方式对搜索引擎的各个组成部分和核心技术进行了介绍。第 1 章引言，对搜索引擎进行了简要概述，介绍了搜索引擎与信息检索的关系，搜索引擎的历史、分类及基本架构。第 2 章信息采集，主要围绕搜索系统的核心——网络爬虫进行介绍。第 3 章文本处理，对搜索引擎的文本处理功能进行了介绍，包括文本信息的提取、自然语言中的统计语言模型、中英文分词技术、网页去重算法等。第 4 章搜索引擎索引构建，主要介绍搜索引擎的索引系统，包括倒排索引、建立索引的方式、索引的更新策略、分布式索引及索引压缩算法。第 5 章基于文本内容的检索模型，对搜索引擎的检索模型进行了介绍，包括传统的检索模型，如布尔模型、向量空间模型、概率检索模型和基于统计语言建模的检索模型，以及基于机器学习的排序模型。第 6 章基于链接的检索模型，主要对基于链接的检索模型和针对链接作弊的反作弊模型进行了介绍。第 7 章查询处理与结果展示，主要对查询条件的纠正与过滤、查询处理与展示的技术进行了介绍。第 8 章相关反馈与查询扩展，主要对围绕着相关反馈和查询扩展的各项技术进行了介绍，通过采用相关反馈和查询扩展的技术理解用户的查询意图。第 9 章分类与聚类，主要介绍了在智能搜索引擎中用到的各种机器学习算法。第 10 章基于知识图谱的搜索引擎，对未来搜索引擎的发展方向——基于知识图谱的智能搜索引擎进行了介绍，包括知识图谱的构建流程、构建中的信息

抽取、知识融合、知识表示与推理等关键技术及其在搜索引擎中的应用。第2至第7章主要介绍了核心的搜索引擎功能与技术，形成搜索引擎技术的基本框架。第8至第10章是扩展内容部分。在教学学时有限的情况下，可以只对前七章内容进行介绍。

本书适用于数据科学与大数据技术专业及其计算机相关专业的本科生或研究生以及从事该领域研究的人员。通过对本书的阅读，可以使读者对智能搜索引擎的相关知识有一个基本的了解，并为将来开展研究工作打下坚实的基础。

本书在编写过程中得到了广泛的支持与帮助。中南大学为数据科学与大数据专业设立了教材出版专项；中南大学出版社与中南大学信息科学与工程学院的相关领导也高度重视，成立了系列教材编写委员会，多次组织专题讨论会，带领编委会成员多次外出学习访问；邀请了厦门大学林子雨老师参加编委会教材专题讨论。在此，对支持、帮助及关注本书的各位同仁表示感谢。

由于作者水平有限，书中难免会存在疏漏，敬请读者批评指正。

编 者

2018年10月

目录

Contents

第1章 引 言

1.1 信息检索与搜索引擎

在信息飞速增长的时代，成千上万的人每天都用信息检索工具对大规模的电子文本进行信息的搜索和处理。信息检索是一种文本处理工具，主要是对文本信息的检索，其核心是文本信息的索引和查询。Gerard Salton 是 20 世纪 60—90 年代信息检索领域的领袖人物之一。在他的经典教科书中，对信息检索给出了以下定义：信息检索是关于信息的结构、分析、组织存储、搜索和检索的领域。从历史上看，信息检索经历了手工检索和计算机检索两个主要阶段。手工检索阶段是信息检索的早期阶段，主要通过人工建立检索目录，应用于图书情报的索引和查询。计算机检索阶段是利用计算机实现自动化处理的阶段，它在 20 世纪 60—80 年代形成并发展，在信息检索领域逐步扩大。

计算机检索阶段发展到更高阶段，信息的存储量越来越大。特别是互联网的发展所带来的知识爆炸，导致了人们快速准确地在信息海洋里发现自己所需要的信息越来越难。搜索引擎是信息检索技术在大规模文本集合上的实际应用，是解决互联网时代信息过载的相对有效的方式。搜索引擎，英文又叫 Search Engine，是指一组特定的软件系统，根据一定的策略从互联网上采集信息，并对信息进行处理与存储，将存储的信息与用户的信息需求(information need)相匹配，将匹配的信息展示给用户的系统。用户首先提交查询关键词给搜索引擎，搜索引擎会返回给用户与查询关键词相匹配的文档。在互联网上，假设你将关键词“搜索引擎”输入到网页的搜索文本框，点击“搜索”按钮，搜索引擎会快速返回结果。其结果是与“搜索引擎”紧密相关的页面 URL，页面标题和一段从网页中提取的简短文字。相关性强的页面会排在前面，大大地方便了用户的信息查找。因此，搜索引擎不仅是查询系统，而且也是用户自定义的信息聚合系统。搜索引擎已经成为互联网应用层上最为重要的应用。目前搜索引擎已经成为大多数人上网查找信息的必要的工具。百度、Google 和 Bing 等著名 Web 搜索引擎已经成为目前人们使用最为普遍、最广泛的搜索引擎。通过这些搜索引擎，人们可以获取最新的技术信息，搜索人和组织、新闻事件等各类资讯。

传统的搜索引擎主要是对文本数据进行索引与查询。这种对文本数据的搜索通常是在文本框中输入查询关键词，点击“查询”按钮，搜索引擎就返回包含这些关键词的相关网页。但随着信息存储与信息处理技术的发展，搜索引擎的检索对象也发生了变化，扩充到了对图像、音乐、视频等各种多媒体信息进行搜索与查询，各大搜索引擎厂商也加强了这方面的工作。百度、Google 和 Bing 等著名 Web 搜索引擎都提供了对图片、音频、视频等的检索。

1.2 搜索引擎的历史

搜索引擎发展至今已有 20 多年的历史。Archie 是人们公认的搜索引擎鼻祖，它由加拿大麦吉尔大学计算机学院的师生于 1990 年共同开发。当时 Web 还没有进入应用阶段，互联网中的资源主要还是以 FTP 协议传输。Archie 是一个用于 FTP 服务器的搜索引擎，它能定期搜集并分析 FTP 服务器上的文件名信息，用户通过输入准确的文件名来检索可以获取该文件的 FTP 服务器地址。Archie 和搜索引擎的基本工作方式是一样的：自动搜集信息资源、建立索引、提供检索服务。

1991 年 Web 技术标准出台，人类对于互联网的使用进入了一个新的阶段，Web 使互联网的使用变得更加丰富、快捷、简单。1993 年 Web 免费开放，同年世界上第一个网络爬虫 World wide Web Wanderer 诞生了，用于追踪 Web 发展规模。起初用于统计 Web 服务器数量，后来增加获取 URL 的功能。同年 10 月，第一个用于 Web 的搜索引擎 ALIWEB(archie - like indexing of the Web)诞生了，命名含义是类似于 Archie 的 Web 索引。按照技术分类的话，它属于分类目录搜索引擎，不使用爬虫获取网页信息，而是通过用户提交自己网页的简介信息来收录网络数据。1994 年 Infoseek 创立，稍后即正式推出搜索服务，并允许站长向 Infoseek 提交网址，这是第一代的搜索引擎。Infoseek 网站如图 1 - 1 所示。

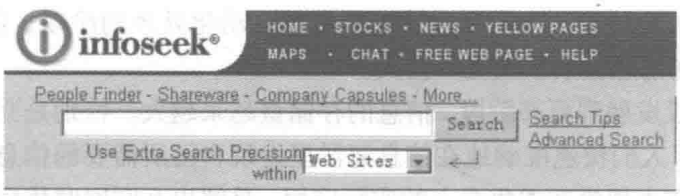


图 1 - 1 Infoseek 网站

1993 年末，出现了大量的基于爬虫的搜索引擎，最有代表的三个是 JumpStation、The World Wide Web Worm 和 Repository - Based Software Engineering(RBSE) spider。前两者没有信息关联度，只是在数据库中检索相关信息再按顺序将结果返回。RBSE 则是首个提供了索引网页正文以及首个按关键字相关性排序返回结果的搜索引擎。

1994 年初，第一个可以索引全文内容的搜索引擎 WebCrawler 诞生。同年，第一个具有现代意义的真正搜索引擎 Lycos 诞生了，它属于第二代搜索引擎，将爬虫程序与索引程序结合，具有前缀匹配、网页相关性排序以及网页自动摘要等功能。

1995 年，元搜索引擎的概念被提出。它将用户的检索请求提交给其他多个独立搜索引擎进行检索，再集中各个引擎的返回结果进行分析排序。这种搜索引擎也不是现在意义上的搜索引擎。年底，AltaVista 的出现更新了搜索引擎的定义。它支持自然语言处理，并且具有高级的搜索语法。

1998 年，Google 出现，搜索引擎又添加了更多方面的功能，增加了对链接的分析。

国内搜索引擎起步较晚，但发展很快。1998 年出现的 Openfind 为当时的新浪、Yahoo 等门户网站提供中文搜索服务。1999 年，李兴平创建了类似于 Yahoo 的导航网站 hao123。2000

年，李彦宏创立百度，专注于中文搜索引擎领域，从为搜狐等公司提供搜索服务开始，很快就成为中文网络世界最大的搜索引擎。同年，Google 推出中文简体和繁体服务。2003 年，中文互联网的四大门户(新浪、搜狐、腾讯、网易)分别涉足搜索领域，并先后推出了自己的搜索引擎服务："爱问""搜狗""搜搜""有道"。

2003 年开始，随着计算智能、数据挖掘领域的快速发展和广泛应用，搜索引擎领域提出了第三代搜索引擎的概念：对万维网中的网页进行更加全面的分析和更深度的数据挖掘，使得其不仅可以产生多个结果，而且使结果更加人性化、智能化、精确化。这一代搜索引擎的目的是让搜索引擎可以更深入地理解用户的需求，并产生更符合用户期望的结果。同时，在第三阶段，搜索引擎的开发也进入了开源搜索引擎的时代。开源搜索引擎为人们提供了透明的搜索引擎，可以根据各种需求对其扩展，对第三代搜索引擎的发展有着重要意义。

第三代搜索引擎之后，又出现了以互动搜索、多模搜索、移动搜索等为中心的新的发展高潮。多模交互搜索是指搜索引擎应用到更加广泛的领域，如图片、视频等多媒体的搜索以及返回结果格式的多样化，结果不仅仅是相关的链接，也包括图片、视频等格式。其中，移动搜索随着移动客户端对于搜索引擎的需求应运而生，移动搜索给用户提供了更好的体验，使得用户可以更加便捷地进行信息检索。

1.3 搜索引擎的分类

搜索引擎经过长时间的发展，目前主流的搜索引擎分为四大类：全文搜索引擎、目录搜索引擎、元搜索引擎、垂直搜索引擎。

1. 全文搜索引擎

全文搜索引擎是当前的主流的搜索引擎，其代表是 Google 及百度等第二代商用搜索引擎。全文搜索引擎是针对万维网所有网页进行全文检索的搜索引擎。由信息采集系统以某种策略自动地在万维网上搜集网页，并且全文搜索引擎的索引系统为采集到的网页的每个词都建立索引，索引的内容包括在文档中出现的位置和次数。它允许用户提交查询关键词，搜索引擎在所有的网页上进行全文查询匹配，返回给用户与查询关键词相关的网页。该类搜索引擎的优点是信息量大、更新及时、不需要人工干预；缺点是返回信息过多、有很多无关信息、用户需要对结果进行筛选。

2. 目录搜索引擎

目录搜索引擎又叫作分类目录搜索引擎。它通过人工操作，将收录到的网站分门别类地进行整理，形成树型的目录结构。用户通过树型的目录，在各级目录下进行信息的查找。因此，它是通过目录导航的方式为用户提供搜索服务。其最具代表的是 Yahoo 和国内的 hao123。该类搜索引擎收录的网站通常质量都比较高，但收录的范围有限，这种方式的可扩展性不强。