搜索引擎
——原理、技术与系统
（第二版）

李晓明　闫宏飞　王继民　著

科学出版社

搜索引擎

——原理、技术与系统

（第二版）

李晓明　闫宏飞　王继民　著

科学出版社

北京

内 容 简 介

本书系统介绍了互联网搜索引擎的工作原理、实现技术及系统构建方案。全书分三篇共13章。上篇介绍搜索引擎的基本原理和技术，讲述一个小型简单搜索引擎实现的具体细节；中篇详细讨论了大规模分布式搜索引擎系统的设计要点及其关键技术；下篇结合“中国Web信息博物馆”和“中国互联网数字资源财富库藏”的实践经验，介绍了构建大规模Web历史网页和非网页仓储系统的技术和方法，以及中文网页的自动分类与聚类、开放域问题系统的构建等。

本书层次分明，由浅入深，上篇和中篇涉及内容提供了源代码下载地址；既有深入的理论分析，也有大量的实验数据和程序，具有学习和实用双重意义。

本书可作为高等院校计算机科学与技术、软件工程、信息管理与信息系统、电子商务等专业的研究生或高年级本科生的教学参考书和技术资料；对广大从事网络技术、Web站点管理、数字图书馆、Web挖掘等研究和应用开发的科技人员有很高的参考价值；书中提供了大量源代码，除了用于构建搜索引擎之外，对于学习编程，提高编程技巧，以及实现一个大规模应用开发也有一定的参考价值。

图书在版编目(CIP)数据

搜索引擎：原理、技术与系统/李晓明，闫宏飞，王继民著. —2版. —北京：科学出版社，2012

ISBN 978-7-03-034258-4

Ⅰ.①搜… Ⅱ.①李…②闫…③王… Ⅲ.①互联网络-情报检索-高等学校-教材 Ⅳ.①G354.4

中国版本图书馆CIP数据核字(2012)第090075号

责任编辑：匡 敏 刘兰霞/责任校对：包志虹
责任印制：阎 磊/封面设计：迷底书装

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

北京市文林印务有限公司 印刷
科学出版社发行 各地新华书店经销
*
2005年4月第 一 版 开本：B5(720×1000)
2012年5月第 二 版 印张：21 3/4
2012年5月第八次印刷 字数：451 000

定价：48.00元

（如有印装质量问题，我社负责调换）

第二版前言

2005年4月，在华夏英才基金的支持下，本书的第一版问世。当时，互联网搜索引擎虽然在网民中已经不是一个陌生的概念，但是很少见到针对实际系统的构建，比较系统地介绍搜索引擎原理与实现技术的书。因此本书第一版的出版应该说比较及时，对不少年轻计算机技术人员起到了启蒙的作用。的确，在过去这些年里，我们多次收到读者的反馈，告知那本书对他们职业生涯的作用。最初的读者中，现在有许多都是搜索公司的骨干了。截止到2011年5月，《搜索引擎——原理、技术与系统》已经印刷了七次。一本比较专业的书，能得到那么多读者的喜欢，我们甚感欣慰。

同时我们也看到，随着互联网长盛不衰的发展，搜索的重要性日益突出；而且我们还看到一个现象，那就是相当一批有实力的互联网公司都有进入搜索领域的战略和行动，并不因为有了百度和谷歌就放弃那个市场。这种现象造成了搜索技术人才的很大缺口，尤其是有一定实际开发搜索引擎经验的人才缺口。很多高校看到这种需求，纷纷开设相关课程。许多学者看到这种需求，也纷纷写出各具特色的教材和专著。例如，郭军的《Web搜索》，董守斌和袁华的《网络信息检索》，以及刘奕群、马少平、洪涛和刘子正的《搜索引擎技术基础》。许多出版社也看到了这种需求，引进了大量相关的外版书籍。

那么，既然市场上已经有了这么多关于搜索引擎技术的书，为什么还要出版本书呢？其实这个想法我们4年前就有了，由于客观原因而耽误下来。最根本的原因是，搜索引擎技术的发展和对搜索引擎技术前沿认识的深入，使得我们感到原来书中的有些内容不再重要，而有些新的内容应该包括进去。

2003年秋，在编写本书第一版的时候，主要工作基础是“天网搜索”，它曾经是中国最好的，是我们引以为自豪的搜索引擎。围绕“天网搜索”的开发，北京大学网络实验室培养了一批优秀的学生。本书第一版的内容大都是那些同学实际工作经验的总结，因此一方面总体看的确比较实用，但另一方面某些部分也不够成熟和深入。同时，过去7年来，北京大学网络实验室在搜索引擎技术方面的研究工作也有了深入进展，尤其在搜索评测与高性能索引结构方面取得了一批前沿成果，这些都是在修订中要考虑的内容。

本书保留了第一版上篇的大部分内容，即搜索引擎的基本原理，过去这么些年并没有什么变化；删除了第一版中的第九，第十二和十三章，增加了第十，第十一和十三章，分别介绍基于搜索引擎技术开发并从2002年一直运行至今的“中国Web信息博物馆”、“中国数字财富库藏”及开放域问答系统。同时，较大幅度修订了第一版中的部分小节内容。总的来看，第二版中约45%的内容是新的，且总篇幅比第一版增加

约30%。

鉴于我们在第一版中的一个特色——详细介绍了一个小型搜索引擎(TSE)并提供源代码,引起了许多读者的兴趣,纷纷下载和来邮件咨询,在此高兴地告诉读者:北京大学网络实验室将开放天网搜索系统的所有源代码。用该源代码构建的系统能搜集和处理上亿量级的网页,其体现的技术与本书中的几章内容相对应。

还有两个原因促使我们完成第二版的修订。一是2011年我们与百度合作承担一个国家项目"基于框计算的新一代搜索引擎与浏览器",尽管规定的任务中没有要求,但我们认为能有这么一本书在项目完成之际面世,也是一件令人高兴的事情。二是2003年我们发起了"全国搜索引擎与网上信息挖掘学术研讨会",首届在北京大学举办后,每年在全国各地由不同的高校轮流举办,很多人在会议组织工作中付出了巨大努力,他们是华南理工大学的董守斌、清华大学的李星、山东大学的马军、海南大学的雷景生、江西师范大学的王明文、大连理工大学的林鸿飞、西华大学的杜亚军,河北大学的袁方。今年已是第十届了,又回到北京大学来举办,本书的出版既是对本届会议的一个献礼,也是对过去10年来曾经为"全国搜索引擎与网上信息挖掘学术研讨会"做过贡献的所有朋友,以及参加会议的所有人员的感谢。

只要互联网还在发展,Web还在发展,搜索引擎技术就会不断发展。我们相信《搜索引擎——原理、技术与系统(第二版)》能向读者更深入地展示搜索技术视野,同时也有助于启迪创新的搜索技术。同时我们愿意与读者交流,并欢迎读者指出书中难免的疏漏(lxm@pku.edu.cn;yhf@net.pku.edu.cn;wjm@pku.edu.cn;http://sewm.pku.edu.cn/book/)。

作 者

2012年3月于北京大学燕园

第一版前言

随着互联网的不断发展和日益普及，网上的信息量在爆炸性增长，全球 Web 页面的数目已经超过 40 亿，中国的网页数目估计也超过了 3 亿。目前人们从网上获得信息的主要工具是浏览器，而通过浏览器得到信息通常有三种方式：第一，直接向浏览器输入一个关心的网址（URL），如 http://net.pku.edu.cn，浏览器返回所请求的网页，根据该网页内容及其包含的超链接文字（anchor text）的引导，获得自己需要的内容；第二，登录到某个知名门户网站，如 http://www.yahoo.com，根据该网站提供的分类目录和相关链接，逐步“冲浪”浏览，寻找自己感兴趣的东西；第三，登录到某个搜索引擎网站，如 http://e.pku.edu.cn，输入代表自己所关心信息的关键词或者短语，依据返回的相关信息列表、摘要和超链接引导，试探寻找自己需要的内容。

这三种方式各有特点，各有自己最适合的应用场合。第一种方式的应用是最有针对性的，例如，要了解北京大学计算机系网络与分布式系统实验室在做些什么工作，从某个渠道得知该实验室的网址为 http://net.pku.edu.cn，于是直接用它驱动浏览器就是最有效的方式。第二种方式的应用类似于读报，用户不一定有明确的目的，只是想看看网上有什么有意思的消息；当然这其中也可能是关心某种主题，如体育比赛、家庭生活等。第三种方式适用于用户大致知道自己要关心的内容，如“国有股减持”，但不清楚哪里能够找到相关信息（即不知道哪些 URL 能给出这样的信息）；在这种场合，搜索引擎能够为用户提供一个相关内容的网址及其摘要的列表，由用户一个个试探看是否为自己需要的。现在的搜索引擎技术已经能做到在多数情况下满足用户的这种需要。CNNIC 的信息统计指出，目前搜索引擎已经成为继电子邮件之后人们用得最多的网上信息服务系统。

同时，随着网上信息资源规模的增长，尤其是其内容总体和我们社会的演化发生着越来越密切的联系，研究网上存在的海量信息逐渐成为许多学科关注的一个方向。为此，不少研究人员也有采样搜集特定内容、一定数量网页的需要。

本书以我们设计、实现并维护、运行北大“天网”搜索引擎的实际经验，介绍大规模搜索引擎的工作原理和实现技术。我们要向读者揭示，为什么向搜索引擎输入一个关键词或者短语，就能够在几秒钟内得到那么多相关的文档及其摘要，而点击其中的链接就能够被引导到文档的全文，且其中相当一部分可能正是用户需要的。

我们按照上、中、下三篇展开相关的内容。上篇讲搜索引擎的基本工作原理和技术，要解决的是为什么搜索引擎能提供如此庞大的信息查找服务这一问题，以及它在功能上有什么本质的局限性。这一篇的内容包括网页的搜集过程，网页信息的提取、组织方式和索引结构，查询提交和响应的过程以及结果产生，等等。这其中，虽然我们假定读者熟悉 URL、HTML、HTTP、CGI、MIME 等基本概念，但在上下文中也给

予了必要的介绍，力图保持行文的流畅性。这一部分内容对于需要构建小规模搜索引擎的研究人员会有直接的参考价值。

中篇讨论和大规模实用搜索引擎有关的技术问题。所谓大规模在这里指至少维护超过1000万的网页信息，提供相关的查询服务。所涉及的内容包括并行分布处理技术的应用，数据局部性的开发，缓存技术的应用以及搜集的网页在提供服务之前的预处理问题和高效倒排文件的建立技术等。这一部分的讨论有比较强的计算机系统结构的风格，我们将向读者展示计算机系统结构课程中的那些概念是如何生动地体现在一个实际应用系统中的。这一部分内容对构建大规模数字图书馆的技术人员也应该有帮助。

下篇介绍挑战性更强一些的内容。一般地讲，前面所述可以称为是“通用搜索引擎”，为最广泛的人群提供信息查询服务是它的基本宗旨。这意味着它的应用模式必须尽量简单，即关键词或查询短语的提交和匹配响应。尽管这已经可以解决许多问题了，但对有些重要的信息需求依然显得力不从心。例如，一个人可能会关心最近半年来网上出现了哪些关于他(她)的信息，一个企业可能要关心它做了一次大规模促销活动后一个月内网上有什么反响，一个政府机构可能会关心在一项政策法规颁布后的网上舆论。面向主题和个性化的信息查询服务就是我们试图描述的一种基本途径。这一部分内容更多地和网上中文信息处理技术有关。更准确地讲，我们要介绍网络与并行分布处理技术和中文处理技术的结合，从而实现大规模、高性能、高质量、有针对性的网上信息查询服务。这一部分内容反过来可能对从事中文信息处理的研究人员有启发作用。

本书的内容是集体智慧的结晶，主要概括了北京大学计算机科学技术系网络与分布式系统实验室自1996年以来的研究成果。其中许多段落直接来自同学的博士和硕士论文，他们是雷鸣、赵江华、冯是聪、单松巍、谢正茂、彭波、张志刚、龚笔宏、孟涛、昝红英等。署名作者的主要工作是将这些内容系统化，使其表述的风格统一。我们特别感谢陈葆珏教授，是她在北京大学计算机系开创了搜索引擎这一研究方向，从而使我们能在其后发扬光大，还要感谢刘建国和王建勇，是他们分别带领攻关队伍，实现了天网1.0和天网2.0版本。感谢黄蕊为本书进行的文字校对。最后，我们要感谢国家九五攻关计划、973计划和985计划的支持，是它们的不断支持使我们得以将天网不断推上新的台阶，实现“让天网和中国网上信息资源规模同步成长”的理想。

作　者

2004年5月于北京大学燕园

目　录

中篇 对质量和性能的追求

下篇 Web 信息资源的组织与应用服务

图 表 目 录