普通高等教育“十一五”国家级规划教材
计算机科学与技术系列教材 信息技术方向

搜索引擎技术基础

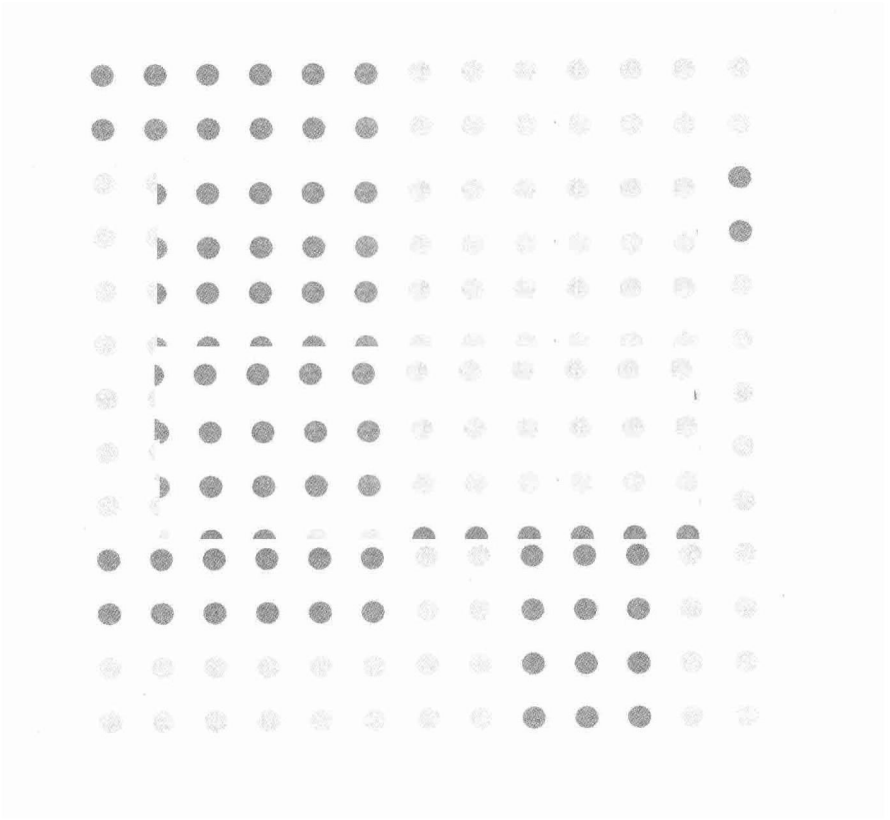
刘奕群 马少平 洪涛 刘子正 编著

清华大学出版社
http://www.tup.com.cn

普通高等教育“十一五”国家级规划教材
计算机科学与技术系列教材 信息技术方向

搜索引擎技术基础

刘奕群 马少平 洪涛 刘子正 编著



清华大学出版社
北京

内 容 简 介

这是一本关于搜索引擎的教科书，它从研究实践者的角度介绍了搜索引擎的相关技术及其产业，并试图协助读者成为搜索引擎领域的局内人。与传统的将搜索引擎作为信息检索系统实现的一个特殊实例的做法不同，作者试图把搜索引擎作为一个独立的研究课题，从纷繁复杂的互联网数据现象和搜索引擎工作案例中提炼知识点，对现代商业搜索引擎的体系结构、运行原理、运营机制和核心算法进行总结和讲解。

本书是清华大学计算机系与百度公司合作在清华大学开设的“搜索引擎技术基础”课程的教材，适合作为高等院校信息科学技术、图书馆学等相关专业本科生与研究生相关课程的教材，也可作为相关领域技术人员与搜索引擎技术爱好者的参考资料。

图书在版编目（CIP）数据

搜索引擎技术基础/刘奕群等编著．—北京：清华大学出版社，2010.7
（计算机科学与技术系列教材　信息技术方向）
ISBN 978-7-302-22796-0

Ⅰ．①搜…　Ⅱ．①刘…　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(2010)第 095896 号

责任编辑：张瑞庆　薛　阳
责任校对：白　蕾
责任印制：李红英

出版发行：清华大学出版社　　**地　　址**：北京清华大学学研大厦 A 座
http://www.tup.com.cn　　**邮　　编**：100084
社　总　机：010-62770175　　**邮　购**：010-62786544
投稿与读者服务：010-62795954，jsjjc@tup.tsinghua.edu.cn
质　量　反　馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 刷 者：清华大学印刷厂
装 订 者：三河市新茂装订有限公司
经　　销：全国新华书店
开　　本：185×260　　**印　　张**：17.25　　**字　　数**：413 千字
版　　次：2010 年 7 月第 1 版　　**印　　次**：2010 年 7 月第 1 次印刷
印　　数：1～4000
定　　价：26.00 元

产品编号：032454-01

计算机科学与技术系列教材 信息技术方向

编 委 会

PREFACE

序 言 1

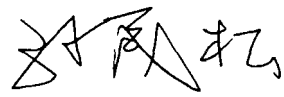
面对浩瀚的万维网信息海洋，人类并没有如《庄子·秋水》中的河伯那样望洋兴叹、徒唤奈何，这实在是拜搜索引擎之功。搜索引擎是人们从无远弗届、无深不入的万维网中获取信息不可或缺的手段，是人们遨游于这个海洋里孜孜以求的“探海金针”。搜索技术也因此成为当今最热门的研究热点之一，为信息检索、数据挖掘、自然语言处理等众多领域所共同关注。

与世界上其他国家的发展路径有所区别，中国搜索引擎的发展一直坚持本土化的道路，一方面，确保了数以亿计的中文网民获取互联网信息过程的便利；另一方面，也确立了中文信息处理技术在世界范围的影响，这是与以百度、搜狗、搜搜等为代表的一系列“国产”搜索引擎的技术创新和产业发展分不开的。

技术创新和产业发展都需要优秀人才的支撑。培养对搜索技术具有比较深刻理解的计算机专业高端人才是中文搜索引擎乃至信息处理产业发展的迫切需求。然而，搜索引擎属于比较新的研究方向，其核心技术研发与知识体系演化的速度很快，如何从纷繁复杂的产品及其功能中凝炼出搜索引擎人才真正需要的知识与技能，是相关教学工作开展中面临的重要问题。鉴于搜索引擎发展过程中融合了学术界与产业界两方面的创新成果，我们认为，解决这一问题也需要大学与搜索引擎企业的共同努力。

作为这方面的一个积极探索，清华大学计算机系和百度公司从2009年春季起开始合作开设“搜索引擎技术基础”课程，希望为相关人才培养贡献绵薄之力。课程受到了清华大学同学的欢迎与好评，也激励了不少同学尝试开展搜索引擎方面的研究与创新。清华大学的刘奕群博士、马少平教授与百度公司的洪涛先生、刘子正先生合作完成的这本书就是该课程的教材。作为为数不多的搜索引擎技术中文教科书之一，该教材系统评价了搜索引擎技术与产业发展的概况，对搜索引擎领域得到广泛应用的各种核心算法和应用模式进行了阐述与探讨。“鸳鸯绣出从君看，更把金针度与人”。相信每一位对搜索引擎感兴趣的学生和学者都能通过学习或参考此书而有所收获。

最后，希望读者通过本书尽早成为搜索引擎领域的“行内人”。万维网上的信息构成了清风与明月交织的虚拟世界，“耳得之而为声，目遇之而成色，取之无禁，用之不竭”。掌握了搜索引擎技术，会使我们从一个更高的境界去品味、享受万维网这个“造物者之无尽藏也”所带来的无尽乐趣。



清华大学计算机科学与技术系主任

2010 年 3 月

序 言 2

18 世纪的著名英国作家塞缪尔·约翰逊曾经说过："知识有两种，其一是我们自己精通的问题，其二是我们知道在哪里找到关于某问题的知识。"但显然，在互联网和搜索引擎出现之前，获得第二种知识的渠道艰难而匮乏。

互联网堪称是 20 世纪以来影响整个世界的最伟大发明。在互联网的带动下，人们一下子进入了一个崭新的信息爆炸时代，各种各样的知识和信息层出不穷，令人眼花缭乱。面对浩瀚的信息海洋，人类所面临的最大困扰是：如何在尽可能短的时间里，找到最想要的东西？

搜索引擎技术的出现和发展，让这一切变得简单。

借助搜索引擎，塞缪尔·约翰逊所说的第二种知识的获取过程变得容易起来，每个人都可以轻而易举地在互联网上找到所求。可以说，搜索引擎的问世，拉近了各种地域、阶层和职业的人们与信息之间的距离，在消除信息鸿沟和加速知识进化过程中发挥着越来越重要的作用；而同时，由于具备更加精准、高效、低成本、高覆盖等特点，搜索引擎也彻底颠覆了传统的营销观念和模式，成为众多企业首选的营销推广方式，在推进传统经济向数字化经济迈进的过程中贡献出了巨大力量。

更为重要的是，搜索引擎技术的发展，不仅关系到用户和企业的利益，也关系到一个国家的信息安全和更长远、更深层次的网民的知情权。虽然互联网信息技术发展一日千里，但至今拥有市场认可的独立搜索引擎技术的国家也只有四个：美国、韩国、俄罗斯和中国。

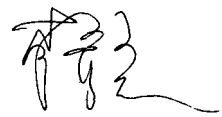
目前，拥有近 4 亿网民的中国，已经成为全球最大的互联网国家，互联网相关应用与创新层出不穷，搜索引擎也已经日益渗透到社会生活和国家发展的各个层面，成为衡量我国信息技术发展水平的重要标志。在这一局势下，建立中国自己的搜索引擎技术体系并培养更多的搜索引擎技术人才，增强我国搜索引擎技术的持续创新能力，无疑是至关重要的。

然而，在我国当前现有的计算机教学和培训体系中，涉及搜索专业技术知识的内容却少之又少，针对搜索引擎技术的系统讲授和相关应用案例更是远远不够。这种相关知识的匮乏显然不利于我国未来搜索引擎技术的发展，甚至可能导致我国未来搜索引擎技术的发展瓶颈。

PREFACE

这也是百度与清华大学联合开设“搜索引擎技术基础”课程，并编撰本教材的初衷。对于百度而言，“技术改变世界”一直是我们的梦想和努力的方向，我们希望通过将公司十年来在搜索技术研发和应用方面的经验和已初步形成体系的技术培训课程引入高校，与更多的人一起分享搜索引擎领域最新的技术进展，为我国计算机专业人才培养和搜索引擎技术进步带来一些帮助。

这只是一个开始。我们相信，随着未来越来越多的科研机构和技术企业投入其中，中国的搜索技术和互联网发展必将迎来一个更加迅猛的发展阶段，而我们的大国崛起之梦，也将随着技术自主创新能力的日益增强，变得越来越真实。



百度董事长兼首席执行官

2010 年 3 月

FOREWORD

前　言

2008年底，笔者开始筹备“搜索引擎技术基础”课程的教学。尽管搜索引擎作为互联网时代最重要的基础设施之一的地位已经得到了广泛的认可，而百度等优秀企业的崛起也使得搜索引擎行业成为近几年高校信息技术专业毕业生最热门的就业去向之一，但我们却发现国内外在这方面的教学资源都处于相对匮乏的状态。这一方面是因为搜索引擎仍是一个相对年轻的技术领域；另一方面也是与搜索引擎的数据对象——互联网日新月异的变化分不开的。

面对这一问题，当时一个较为简便的解决方式是尝试从信息检索原理与系统的角度入手，把搜索引擎作为一个以互联网用户为服务对象的网络信息检索系统来进行分析和讲述。由于信息检索领域已经有半个多世纪的研究积累，知识体系相对完整，包括经典教材在内的教学资源更是数量繁多，开展教学工作会有很多有利的条件。相比较而言，把搜索引擎作为一个独立的研究课题，从纷繁复杂的互联网数据现象和搜索引擎工作案例中提炼出知识点、进而形成体系是一件繁重得多的工作。

然而，我们最终还是弃简就繁，选择抛开研究信息检索领域的传统思路，而把搜索引擎作为一个融合了信息检索、数据挖掘、机器学习、自然语言处理、网络经济、用户行为分析等多方面内容在内的综合性课程来组织教学、撰写教材。这是由于我们深切地体会到，尽管搜索引擎与传统信息检索系统都是协助用户从海量数据中获取有用信息的工具，但数据对象和服务对象的不同造成了两者在体系结构、核心算法、运营机制方面本质性的差异。这也正是在搜索引擎企业招聘人才时，“懂得互联网”往往比“懂得信息检索”更为“吃香”的原因。把搜索引擎作为一个独立的研究领域，而不是一种传统信息检索技术的应用系统，这是我们在课程设计和教材撰写中的最大特色，也是我们希望读者在进行搜索引擎研究与学习时需要时刻铭记的一点。

搜索引擎是计算机科学研究中学术界与产业界最紧密合作的领域之一，因此在课程教学方面得到产业界的支持也是十分重要的。我们有幸能够与中文搜索引擎领域的领先者、也是全球最大的中文搜索引擎公司——百度公司开展了合作教学项目。合作项目从课程设计之初就开始进行，百度公司的搜索引擎技术研究与培训机构——百度技术学院（BIT）对合作项目给予了大量的支持。在与包括洪涛先生、刘子正先生在内的多位公司高级研发人员的反复讨论与雕琢

FOREWORD

之后，课程内容和知识体系逐渐成型，合作内容讲授与多层次动手实践相结合的教学方法也确定了下来。2009 年春，清华大学计算机系的同学有机会第一次进行了这门内容和形式都比较新颖的“搜索引擎技术基础”课程的学习。而我们也在清华大学出版社的支持下，把课程相关的讲义加以整理，编写成了读者手中的这本教材。

本书首先从搜索引擎的基本概念和发展历史出发，使读者对搜索引擎产业和搜索引擎技术有一个概括的认识；随后介绍搜索引擎性能评价的理论与方法，吸引读者站在搜索用户的角度审视搜索引擎性能的优劣，并思考其内在原因；接下来通过对搜索引擎体系结构的讲解，并结合具体应用案例分别讲述数据抓取、内容索引、内容检索、链接结构分析 4 个搜索引擎组成模块的系统设计与核心算法，以使读者由浅入深地掌握搜索引擎运行的基本机理；紧接着通过对数据质量评估和垃圾网页识别两个搜索引擎运行过程中对性能有关键性影响的问题的阐述，协助读者理解万维网数据环境的特点及其对搜索引擎造成的挑战；而后通过对搜索引擎广告技术的剖析说明搜索引擎的运营和盈利模式；最后进行搜索引擎技术未来发展的一点展望。

林语堂在爱斯嘉拉《中国之过去与现在》的英文版序言中提到“有时一个人需要花一生的时间，学习以权威而又简明的语言向其他人介绍一件事”。我们远未达到可以熟练地使用“权威而又简明的语言”的地步，但在本书的整个撰写过程中，我们也努力避免晦涩空洞的介绍以及“为了形式化而形式化”的讲述。搜索引擎是一门鲜活的年轻学科，我们希望通过大量的示例，引导读者站在搜索引擎设计者的角度体会其设计思路与核心算法，唤起读者使用和研究搜索引擎的热情。我们也希望通过对本书的阅读，即使是非信息技术领域的研究人员和学生、甚至普通的搜索引擎用户都能够更加深入了解搜索引擎、有效利用搜索引擎，投身到这个方兴未艾的“有趣”的行业中去。

本书是在清华大学计算机系以及百度公司两方面的人员通力合作之下完成的。其中，第 1～3 章由刘奕群、马少平、陈磊执笔，第 4 章由刘奕群、马少平执笔，第 5 章由赵柏敏、刘子正、马飞、刘奕群执笔，第 6 章由李智超、刘奕群执笔，第 7 章由刘知远、李智超、刘奕群执笔，第 8～10 章由刘奕群执笔，第 11、12 章由洪涛、刘奕群执笔。最后由刘奕群统稿并整理。

笔者在本书的成稿过程中，得到了来自各方面同行、同事的指导与帮助，下面逐一表示感谢之意。

感谢清华大学计算机系主任孙茂松教授与百度公司董事长兼首席执行官李彦宏先生为本书撰写序言，他们分别站在学术界与产业界的制高点对搜索引擎的未来发展进行了展望，他们对本书的推荐相信能够使本书为更多读者所知。

感谢百度公司校园关系部刘湘雯女士与杨斌先生，他们的积极协调与努力使得清华与百度的课程合作项目成为现实。

FOREWORD

感谢百度技术学院参加合作授课的诸位老师：孙云丰先生、马飞先生、吴凯先生、侯震宇先生和秦首科先生，他们使得清华大学计算机系的同学得以聆听来自搜索引擎技术与产品一线人员的声音，他们的授课也为本书的撰写提供了宝贵的参考资料。

感谢清华大学计算机系张敏博士、金奕江博士以及岑荣伟、周博、方奇、花贵春、余慧佳、薛宇飞、朱彤、邢千里、缪钧伟等同学，笔者与他们共同进行的搜索引擎领域研究是课程教学和教材撰写能够顺利完成的最重要保证。

感谢清华—搜狐搜索技术联合实验室提供的宝贵数据资源，这些资源使得本书各方面的资料更加翔实，论证更加全面和有力。

感谢哈尔滨工业大学李生教授、刘挺教授，北京大学孙斌博士等提供的宝贵意见、建议与鼓励。

最后需要着重指出的是，正如书中反复提到的，搜索引擎是一门年轻的学科，尽管我们在产业界同行的协助下，尽量尝试对其中相对成熟和稳定的技术原理、体系结构等内容进行提炼，但知识结构体系中的缺漏以及内容中的错误一定不少。我们欢迎并期望广大同行给予批评指正。

刘奕群、马少平、洪涛、刘子正

2010 年 3 月

CONTENTS

目 录

CONTENTS

CONTENTS

第 1 章　为什么要关注搜索引擎

The highest purpose of science is the search for knowledge, truth and a greater understanding of the world around us

——Barack Obama, U. S. president

这是一本关于搜索引擎的教科书，它从研究实践者的角度介绍了搜索引擎的相关技术及其产业，并试图协助你成为搜索引擎领域的局内人。在开始本书的内容之前，让我们先花一点篇幅思考这个问题：我们为什么要关注搜索引擎？它到底对我们的生活、工作、学习产生了什么样的影响，以至于我们需要对一个互联网应用系统给予如此多的关注？解答这个问题有助于我们更好地揭开搜索引擎的神秘面纱。

1.1　互联网上最重要的应用系统

图 1.1 是来自于互联网著名网站访问量统计公司 Alexa. com 的访问量统计信息。从图 1.1 中可以看到，全球用户访问量排名前 5 位的网站分别为：Yahoo!，Google，YouTube，Windows Live 和 Facebook。其中 YouTube 为视频分享网站，Facebook 为社交网站，而其余的 3 个网站均为搜索引擎功能为主的综合性网站。

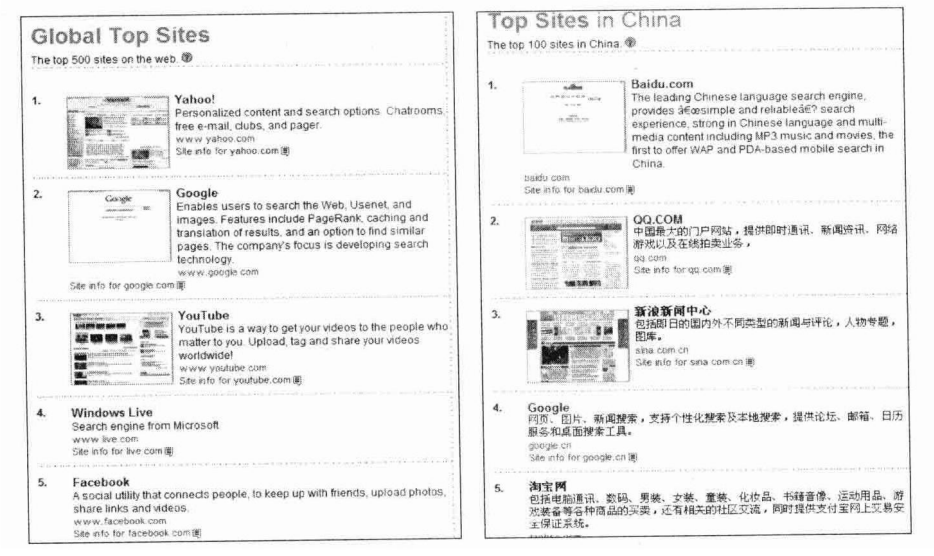


图 1.1　Alexa.com 统计的全球和中国用户访问量最大的网站情况

（数据采集于 2009 年初）

在中国，用户访问量最大的 5 个网站分别为百度、腾讯网、新浪网、谷歌和淘宝网。其中，淘宝网是在线购物网站，百度和谷歌均为搜索引擎网站，而腾讯网和新浪网则为包含搜