网络信息检索技术及搜索引擎系统开发

高　凯　郭立炜　许云峰　编著

科学出版社
www.sciencep.com

网络信息检索技术及搜索引擎系统开发

高　凯　郭立炜　许云峰　编著

科学出版社
北京

内 容 简 介

本书全面、系统地讲述了网络信息检索技术的基本原理，并阐述了其在搜索引擎系统开发及其智能化实现中的应用。在全面介绍了网络信息检索技术、标引与索引、检索结果处理、中英文分词、网络信息获取及预处理之后，本书对信息采集中的网页去重与相似网页聚类、信息的动态采集、基于自然语言理解的检索处理、相关概念反馈、检索纠错、检索结果排序、基于用户浏览历史的网页预取技术等多个方面进行了较深入的研究与分析。

全书体系完整，内容新颖，条理清晰，组织合理，可为高校相关专业学生的学习和科研工作提供帮助，也可为从事搜索引擎技术开发的工程技术人员、希望了解搜索引擎技术的爱好者等提供参考。

图书在版编目(CIP)数据

网络信息检索技术及搜索引擎系统开发/高凯，郭立炜，许云峰编著.
—北京：科学出版社，2010
ISBN 978-7-03-026143-4

I. 网… II. ①高…②郭…③许… III. ①计算机网络–情报检索
②计算机网络–程序设计 IV. G354.4 TP393.09

中国版本图书馆 CIP 数据核字(2009)第 220299 号

责任编辑：赖文华 陈晓萍/责任校对：赵 燕
责任印制：吕春珉/封面设计：耕者设计工作室

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

骏走印刷厂印刷

科学出版社发行 各地新华书店经销
*
2010 年 2 月第 一 版 开本：B5 (720×1000)
2010 年 2 月第一次印刷 印张：16
印数：1—3 000 字数：323 000

定价：32.00 元

(如有印装质量问题，我社负责调换〈环伟〉)
销售部电话 010-62134988 编辑部电话 010-62135120-8003

前　言

在社会信息化的今天，伴随着因特网的迅速普及和应用，网络已成为人们获取信息的重要渠道。为了帮助人们方便、高效地利用网络信息，搜索引擎应运而生。由于网上信息的海量性、冗余性及用户需求的多样性等，迄今搜索引擎在信息采集与处理、检索、个性化等智能化方面尚不能很好地满足用户的需求。如何借助搜索引擎来帮助人们方便、高效地利用网络信息已成为当前 IT 业的研究热点之一。虽然目前搜索引擎在一定程度上满足了人们的检索需求，但随着信息检索的对象从相对封闭、稳定一致、由独立数据库集中管理的内容扩展到开放、海量、冗余、更新快、分布广泛的 Web 内容，同时由于 Web 内容的特殊性（如网页质量参差不齐、镜像网页的存在等），又由于使用者由原来的专业检索人员扩展到包括务工人员、管理人士、教师、学生等在内的普通大众，人们对网络信息检索技术也就提出了新的、更高的要求。目前，网络信息检索在信息采集与处理、检索处理、个性化处理等智能化方面尚不能很好地满足用户的需求。因此，研究网络信息检索技术，并利用包括开源架构等在内的工具构建自己的搜索引擎系统，进而提高搜索引擎的智能化水平，是一项很有意义的工作。

本书较系统地论述了网络信息检索技术的基本原理，并进一步阐述了其在搜索引擎系统开发及其智能化实现中的应用。本书分为三部分。第一部分是基础知识和相关背景介绍部分，包括从第 1 章到第 3 章的内容。其中，第 1 章概要介绍了信息检索的起源和发展、信息检索模型及方法、网络信息检索的过程、网络信息检索性能评价、网络信息智能化处理、网络信息检索技术的未来发展等问题；第 2 章简要介绍了 Web 信息下载、页面分析与信息抽取方法、基于链接分析的网页相关性算法、检索结果排序、自然语言处理等问题；第 3 章就搜索引擎的发展、分类、功能、资源等进行了介绍。第二部分为利用开源资源实现搜索引擎系统的部分，包括从第 4 章到第 8 章的内容，主要介绍如何利用 Lucene 等开源资源来构建自己的搜索引擎。作为开源项目中的一朵奇葩，Lucene 提供了强大的全文索引和检索功能，并在搜索引擎、桌面检索系统、网站站内搜索、企业级内部文档管理与检索、情报分析系统、知识管理系统、数字图书馆检索系统中发挥了很好的作用。本书第 4 章介绍了 Lucene 的索引与检索机制及其应用、开发平台的搭建与配置等；第 5 章介绍 Lucene 中的中英文分词处理及其效果；第 6 章介绍检索结果排序及处理技术；第 7 章介绍如何利用开源资源来获取网络信息；第 8 章介绍如何对常见格式的网络资源进行解析与预处理。第三部分为搜索引擎智能化的研究与实现部分，包括从第 9 章到第 14 章的内容。其中，第 9 章对信息采集中的网页去重与相关网页聚类进行了研究；第 10 章讨论了信息的动态采集

与更新策略，以期搜索引擎能根据网站及其更新速度的不同，动态调整其信息采集与更新的频度；第 11 章则是面向自然语言提问的理解与处理，提供面向大众的支持自然语言提问的智能检索接口不仅能使人机交互更加人性化，还能促进搜索引擎的应用普及；第 12 章则给出一种参照多数用户在检索类似问题时的经验，为用户提供一些关联性和扩展性的相关概念反馈的方法；第 13 章给出一种相近检索与检索结果排序方法；第 14 章阐述了一种基于用户浏览兴趣的网页预取策略。

全书理论联系实际，涉及面广，体系完整，内容新颖，条理清晰，组织合理，图例丰富，说明详细，既可作为高等院校计算机应用技术专业和图书馆等相关专业的教材，也可作为工程技术人员的参考资料。本书由高凯、郭立炜、许云峰合作编著。编著工作分工如下：高凯组稿并提出写作大纲，郭立炜撰写第 1 章、第 2 章，许云峰撰写第 3 章，高凯完成了其他章节的撰写以及最后的审订和统稿工作。

在本书的写作与相关科研课题的研究工作中，得到了多方面的支持与帮助。自动标引和自动文摘模块分别是采用第一作者所在原 OA 实验室标引课题组和自动文摘课题组提供的 DLL，相关科研课题的研究以及项目开发得到王永成教授和李明禄教授的指导，得到国家高科技研究发展计划（863）项目子课题“教育资讯搜索引擎系统”及上海市信息化专项资金项目“智能中文新闻搜索引擎”的支持，得到上海市软件评测中心、上海市远程教育集团等相关单位对项目研发的支持与协作。相关工作中，课题组宋聚平实现了新闻搜索引擎系统的基本体系架构，李刚实现了新闻分类与摘要的动态显示，龙宇巍对国内搜索引擎信息覆盖面进行了调研并初步确定了新闻搜索引擎网页下载种子集，许欢庆实现了股票检索原型系统架构，陈肯完成了对简单逻辑词的初步分析和获取检索项的拼音码、初步的相关概念反馈处理等工作，课题组肖君、丘振华、杨威、钱兵、刘杰等也给予了大力的协作。宗宝琴、张林利协助完成后续的部分文字校对、参考文献整理工作。另外，中国科学院龙星计划、北京大学搜索引擎与互联网信息挖掘组以及闫宏飞副教授均为部分研究工作提供了帮助，国内外众多的信息检索与搜索引擎智能化方面的研究和相关网站亦为本书提供了良好的基础，本书的顺利完成也得益于参阅了大量的相关工作及研究成果，在此谨向这些文献的作者以及为本书提供帮助的老师、同仁和课题组成员致以诚挚的谢意和崇高的敬意。本书亦得到 2009 年度河北省教育厅科学研究计划（编号：2009435）的资助。在本书写作过程中，也得到了科学出版社赖文华、陈晓萍等的大力支持和帮助，在此一并表示衷心感谢。

由于我们的学识、水平均有限，书中不妥之处在所难免，恳请广大读者批评指正。

高 凯 郭立炜 许云峰

2010 年 1 月

目　录

第1章 绪 论

随着现代网络的飞速发展，中国的互联网普及实现再次飞跃，赶上并超过了全球平均水平。据中国网络信息中心CNNIC在2009年1月发布的统计数据显示，截至2008年底，中国网民规模达到2.98亿人，较2007年增长41.9%（见图1-1），因特网普及率达到22.6%，略高于全球平均水平21.9%（注：数据来源于http://www.internetworldstats.com，对比的其他国家和地区因特网普及率为2008年6月底数据）。

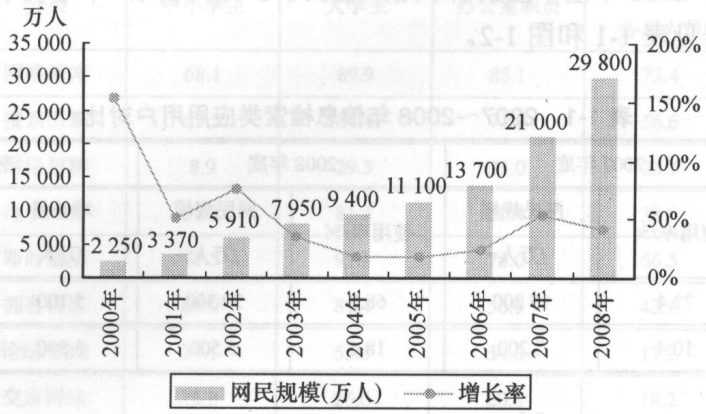

图1-1 2000～2008年中国的网民规模及增长率

伴随着网络应用的普及，网络信息也呈爆炸式增长。英国科学家詹姆斯·马丁认为，人类的知识在19世纪是每50年增加1倍，20世纪中叶是每10年增加1倍，到20世纪70年代就已经缩短为每5年增加1倍[1]。迄今，信息更如爆炸般产生，而且信息的生产能力已超过了人们对其处理和吸收的能力。正如美国作家奈斯比特在其著作《大趋势》一书中曾指出的那样：“我们虽淹没在信息的海洋中，但是却渴求所需的知识”。为什么会出现这种情况呢？主要原因之一是因为缺乏有效的信息检索与知识获取手段。因此，如何帮助人们快速、方便、准确地从信息海洋中寻找到所需信息已成为时代发展的迫切需要。

1.1 引 言

信息检索是伴随着人类社会的进步而发展起来的。从广义上来说，信息检索一般指用户为处理解决各种问题而查找、识别、获取相关的事实、数据、文献的

活动及过程，例如，用户在图书馆查询相关信息资料的行为就属于一种信息检索行为。而狭义的信息检索就是指用户在计算机信息检索系统上进行信息查询的行为[1]。一般认为信息检索是经过了手工检索、脱机批处理检索、联机检索、光盘检索等多个阶段后，逐步发展到今天的网络信息检索。可以说，利用信息检索工具，特别是网络信息检索工具，已经成为现代商业社会和日常生活中不可或缺的一部分。在国外，“Google”已经不再单纯是一个名词，有时它充当的则是动词的角色；在国内，“百度一下”等流行语的出现也说明现在人们越来越依赖于搜索引擎。据 2009 年 1 月我国澳门特别行政区因特网使用现状统计报告公布的数据，在网上娱乐方面，利用因特网搜索引擎寻找信息的网民是最多的，占 78%。而在中国大陆，中国互联网络信息中心 2009 年 1 月发布的统计报告显示，搜索引擎是网民在因特网中获取所需信息的基础应用，其使用率为 68%，在各因特网应用中位列第 4，2008 年全年搜索引擎用户增长了 5 100 万人，年增长率达到 33.6%，具体统计数据见表 1-1 和图 1-2。

表 1-1　2007～2008 年信息检索类应用用户对比

应用	2007 年底		2008 年底		变　化	
	使用率/%	网民规模/万人	使用率/%	网民规模/万人	增长量/万人	增长率/%
搜索引擎	72.4	15 200	68.0	20 300	5 100	33.6
网络求职	10.4	2 200	18.6	5 500	3 300	150.0

图 1-2　搜索引擎用户规模

但搜索引擎的使用存在明显的城乡、年龄、学历、收入差异：城镇网民搜索引擎使用率明显高于农村；20～40 岁网民搜索引擎使用率明显高于其他人群；学历越高，搜索引擎使用率越高；收入越高，搜索引擎使用率越高。各网络应用在重点群体中的普及率见表 1-2（注：数据来源于中国互联网络信息中心）。可见，搜索引擎应用人群的特点决定了它在网络领域的高商业价值。虽然由于网民规模快速增长，新增网民中低学历网民比重增大，而该部分网民的搜索引擎的使用率

较低，导致搜索引擎的整体使用率下降，但这并不妨碍搜索引擎正在成为人们获取信息的重要方式。网络信息检索工具——搜索引擎——已经成为人们获取网络信息的重要手段，是载着人们在信息海洋中遨游的快艇。但由于人们对搜索引擎提出了新的、更高的要求[2~4]，又由于网上信息的海量性、冗余性，及用户需求的多样性等，迄今搜索引擎在信息采集与处理、检索、个性化等智能化方面尚不能很好地满足用户的需求。据中国互联网络信息中心2003年7月至2005年发布的几次统计报告显示，用户对搜索引擎性能感到非常满意的只有23.4%、27.4%、26.9%、28.4%（注：2005年7月后的调查报告中无此项统计数据）。可见，搜索引擎的性能仍有许多需改进之处[5]。

表1-2 各网络应用在重点群体中的普及率/%

应用	分类	中小学生	大学生	办公室职员	农村外出务工人员	总体
网络媒体	网络新闻	68.1	89.9	83.1	73.4	78.5
信息检索	搜索引擎	63.5	84.4	71.9	56.6	68.0
	网络招聘	8.9	29.5	23.0	23.7	18.6
网络通信	电子邮件	52.2	81.4	60.4	38.9	56.8
	即时通信	77.5	91.1	75.0	66.5	75.3
网络社区	拥有博客	64.0	81.4	50.9	43.1	54.3
	论坛/BBS	24.1	55.5	34.6	17.2	30.7
	交友网站	16.8	26.0	20.2	18.2	19.3
网络娱乐	网络音乐	86.9	94.0	83.4	78.2	83.7
	网络视频	67.4	84.4	68.1	57.3	67.7
	网络游戏	69.7	64.2	60.6	55.5	62.8
电子商务	网络购物	16.2	38.8	29.4	11.7	24.8
	网上卖物	2.1	5.2	4.4	0.8	3.7
	网上支付	9.6	30.5	22.4	7.9	17.6
	旅行预订	2.0	6.8	6.8	2.5	5.6
其他	网上银行	7.7	29.9	25.5	7.4	19.3
	网络炒股	4.7	4.7	15.5	4.1	11.4
	网上教育	16.2	25.6	17.3	7.8	16.5

近年来，随着Internet应用的普及，网络信息检索技术吸引了众多学者的目光，许多原先从事人工智能、数据挖掘、自然语言处理等领域的学者也先后加入到信息检索的研究与开发队伍中来，国际上也有专门针对这个研究领域的国际会

议，如 ACM SIGIR 等。纵观网络信息检索技术的发展，未来的搜索引擎将朝着智能化、多功能化、人机交互等方向发展[6]。近年来，研究者投向用户的目光愈来愈多了，而搜索引擎的智能化技术则是当前研发的热点。智能化技术应用了相关领域的多项成果，在信息采集与处理、面向自然语言的理解与检索、个性化服务、信息自动分类等方面能够向用户提供更有价值的服务，因而能较好地满足用户的需要。单就如何提高搜索引擎的智能性来说，文献[7]的观点是："全、快、好、准"应成为搜索引擎系统所追求的主要目标之一。"全"是指搜索引擎应尽量全面地采集所需信息。虽然目前尚无任何一个搜索引擎可以覆盖整个 Internet（据统计，搜索引擎索引网页数据量最大的也不超过整个 Internet 的 16%，所有搜索引擎索引网页数据量的总和大约是整个 Internet 网页量的 42%左右[8,9]），但要力争采集到尽可能全面的信息；同时，由于目前 Internet 上有大量的重复及转载网页[10,11]，因此在尽量全面采集信息的同时，还要考虑到对重复内容的处理。"快"一般指信息采集快、加工处理快、检索快等。而"好"和"准"一般指便于用户方便地与系统进行交流、系统能提供相关的检索结果等，譬如，是否允许用户用自然语言进行提问、能否较为准确地理解用户的检索需求、能否进行相关概念反馈、能否提供信息自动分类等。

本书较全面地介绍了网络信息检索技术的原理技术、进展，并阐述了其在搜索引擎系统中的应用。全书共分三部分。第一部分（第 1 章～第 3 章）对信息检索技术、网络信息智能处理、搜索引擎技术等进行综述，力图使读者对网络信息检索技术和搜索引擎系统有一个全面认识；第二部分（第 4 章～第 8 章）介绍如何借助开源的 Lucene 来搭建自己的搜索引擎应用系统，介绍 Lucene 用于索引和检索 API 的使用、基于 Nutch 的搜索引擎系统的构建等；第三部分（第 9 章～第 14 章）则针对搜索引擎系统实现中涉及的部分智能化处理技术进行研究与实现，通过对信息采集过程中对内容雷同网页的去重与聚类处理、信息动态采集的研究与实现、面向自然语言提问的理解与检索、相关概念反馈技术、相关检索与查询纠错、检索结果排序、基于用户浏览历史的网页预取等的叙述，较全面地阐述了网络信息智能检索技术的研究与应用。

作为全书的绪论，本章首先从信息检索的起源、发展起步，介绍信息检索的基本原理，并给出网络信息检索的评价指标、研究内容等。

1.2　信息检索的起源和发展

1.2.1　手工检索

顾名思义，手工检索就是指人们以手工为主的方式进行信息检索。在长期的社会和生产实践中，手工检索曾经扮演了重要的角色，特别是在计算机出现以前