華夏英才基金學術文庫

李晓明　闫宏飞　王继民 著

搜索引擎

——原理、技术与系统

科学出版社
www.sciencep.com

華夏英才基金學術文庫

搜索引擎

——原理、技术与系统

李晓明　闫宏飞　王继民 著

科学出版社

北京

内 容 简 介

本书系统地介绍了互联网搜索引擎的工作原理、实现技术及其系统构建方案。全书分三篇共13章内容，从基本工作原理概述，到一个小型简单搜索引擎具体细节的实现，进而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；最后介绍了面向主题和个性化的Web信息服务，阐述了中文网页自动分类等技术及其应用。本书层次分明，由浅入深；既有深入的理论分析，也有大量的实验数据，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、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站点的管理、数字图书馆、Web挖掘等研究和应用开发的科技人员也有很高的参考价值。

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. —北京：科学出版社，2005
（华夏英才基金学术文库）
ISBN 7-03-014633-6

Ⅰ. 搜… Ⅱ. ①李…②闫…③王… Ⅲ. 因特网-情报检索 Ⅳ. G252.7

中国版本图书馆CIP数据核字(2004)第121546号

责任编辑：巴建芬 姚庆爽/责任校对：陈玉凤
责任印制：钱玉芬/封面设计：陈 敬

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
源海印刷有限责任公司印刷

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)
2005年4月第一次印刷 印张：16 1/2
印数：1—3 000 字数：312 000

定价：33.00元

（如有印装质量问题，我社负责调换〈环伟〉）

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，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，Cache 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 月于北大燕园

目　录

中篇　对质量和性能的追求

下篇　面向主题和个性化的 Web 信息服务

图表目录

第一章 引 论

信息的生产、传播、搜集与查询是人类最基本的活动之一。考虑以文字为载体的信息，传统上有图书馆、相应的编目体系和专业人员帮助我们很快找到所需的信息，其粒度通常是“书”或者“文章”。随着计算机与信息技术的发展，有了信息检索(information retrieval，IR)学科领域，有了关于图书或者文献的全文检索系统，使我们能很方便地在“关键词”的粒度上得到相关的信息。

我们注意到，上述全文检索系统一般工作在一个规模相对有限、内容相对稳定的馆藏(collection)上，被检索的对象通常是经过认真筛选和预先处理的(如人工提取出了“作者”、“标题”等元数据，形成了很好的“摘要”等)，并且系统需要同时响应的查询数量通常都不会太大(如每秒钟 10 个左右)。

1994 年左右，万维网(World Wide Web，简记为 WWW 或 Web)出现。它的开放性(openness)和其上信息广泛的可访问性(accessibility)极大地鼓励了人们创作的积极性。作为一个信息源，Web 和上述全文检索系统的工作对象相比，具有许多不同的特征，它们给信息检索领域带来了新的发展机遇和技术挑战。

规模大。在短短的 10 年左右时间，人类至少生产了 40 亿网页(Google 2004)，而人类有文字以来上万年里产生了大约 1 亿本书；中国网上到 2004 年初大致有了约 3 亿网页(天网 2004)，而中华民族有史以来出版的书籍大约不过 275 万种。尽管书籍的容量和质量是一般网页不可比的，但在对应的时间背景上考察其文字的总体数量，我们不能不为人类在 Web 上创造文字的激情惊叹！

内容不稳定。除了不断有新的网页出现外，旧的网页也可能会因为各种原因被删除(有研究指出：50%网页的平均生命周期大约为 50 天(Cho et al. 2000，Cho 2002))。

从原则上讲，读者数和作者数在同一个量级，形式和内容的随意性很强，权威性相对也不高，也不太可能进行人工筛选和预处理。

与生俱来的数字化、网络化。传统载体上的信息，人们目前正忙于将它们数字化、上网(花费极高)，而网络信息天生如此。这个特性是一把双刃剑：一方面便于我们搜集和处理，另一方面也会使我们感到太多，蜂拥而至、鱼目混珠。

而作为要在 Web 上提供服务的信息查询系统，如搜索引擎和数字图书馆，通常要具备同时对付大量访问的能力(如每秒钟 1000 个查询)，而且响应时间还要足够的快(如 1 秒钟)。

本书旨在介绍构建这类搜索引擎的有关技术。传统的 IR 是其基础，同时本书

也充分讨论了由上述 Web 信息的特征所带来的新问题及其解决方案。

第一节 搜索引擎的概念

如上所述，本书的主要内容是介绍搜索引擎的工作原理和实现技术。搜索引擎，在本书指的是一种在 Web 上应用的软件系统，它以一定的策略在 Web 上搜集和发现信息，在对信息进行处理和组织后，为用户提供 Web 信息查询服务。从使用者的角度看，这种软件系统提供一个网页界面，让他通过浏览器提交一个词语或者短语，然后很快返回一个可能和用户输入内容相关的信息列表(常常会是很长一个列表，如包含 1 万个条目)。这个列表中的每一条目代表一篇网页，每个条目至少有三个元素：

1) 标题：以某种方式得到的网页内容的标题。最简单的方式就是从网页的<TITLE></TITLE>标签中提取的内容(尽管在一些情况下并不真正反映网页的内容)。本书第七章会介绍其他形成“标题”的方法。

2) URL：该网页对应的“访问地址”。有经验的 Web 用户常常可以通过这个元素对网页内容的权威性进行判断，例如，http://www.people.com 上面的内容通常就比 http://notresponsible.net(某个假想的个人网站)上的要更权威些(不排除后者上的内容更有趣些)。

3) 摘要：以某种方式得到的网页内容的摘要。最简单的一种方式就是将网页内容的头若干字节(如前 512 字节)截取下来作为摘要。本书第七章会介绍形成“摘要”的其他方法。

通过浏览这些元素，用户对相应的网页是否真正包含他所需的信息进行判断。比较肯定的话则可以点击上述 URL，从而得到该网页的全文。图 1-1 是 2003 年 8 月 20 日在天网搜索引擎(http://e.pku.edu.cn)上的一个例子，用户提交了查询词“伊拉克战争”，系统返回一个相关信息列表。列表的每一条目所含内容比上述要丰富些，但核心还是那三个元素。如果用户主要是想从军事角度关心伊拉克战争，第一条目可能就是很好的选择，不仅摘要看起来军事味道要浓一些，而且从 URL (http://mil.eastday.com)上能看到提供信息的大概是一个专门的军事题材网站。如果用户主要是想关心伊拉克战争对全球经济的影响，则后面的条目可能会更相关些。

这个例子提示了我们一个重要的情况，即搜索引擎提供信息查询服务的时候，它面对的只是查询词。而有不同背景的人可能提交相同的查询词，关心的是和这个查询词相关的不同方面的信息，但搜索引擎通常是不知道用户背景的，因此搜索引擎既要争取不漏掉任何相关的信息，还要争取将那些“最可能被关心”的信息排在列表的前面。这也就是对搜索引擎的根本要求。除此以外，考虑到搜索引擎的应用

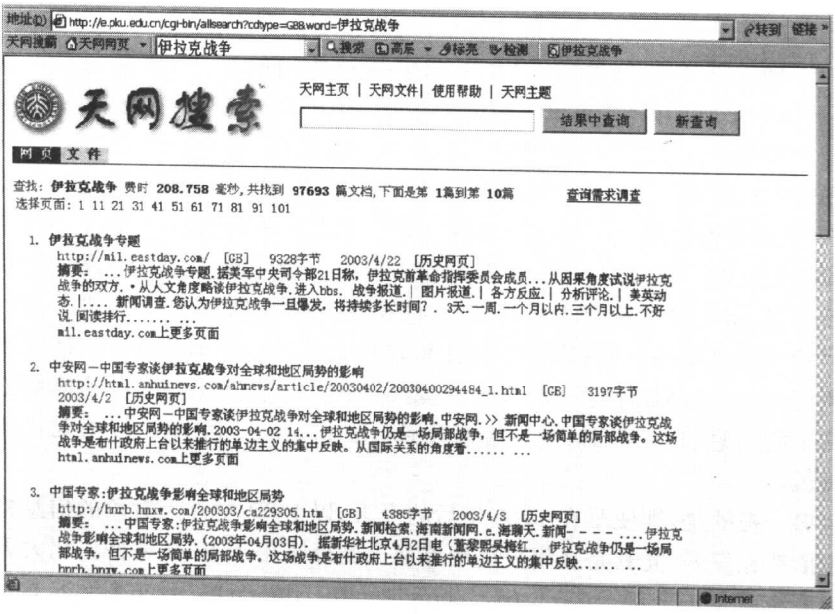

图 1-1 2003 年 8 月 20 日在天网上检索“伊拉克战争”的结果

环境是 Web，因此对大量并发用户查询的响应性能也是一个不能忽略的方面。

作为对搜索引擎工作原理的基本了解，这里有两个问题需要首先澄清。第一，当用户提交查询的时候，搜索引擎并不是即刻在 Web 上“搜索”一通，发现那些相关的网页，形成列表呈现给用户；而是事先已“搜集”了一批网页，以某种方式存放在系统中，此时的搜索只是在系统内部进行而已。第二，当用户感到返回结果列表中的某一项很可能是他需要的，从而点击 URL，获得网页全文的时候，他此时访问的则是网页的原始出处。于是，从理论上讲搜索引擎并不保证用户在返回结果列表上看到的标题和摘要内容与他点击 URL 所看到的内容一致（上面那个“伊拉克战争”的例子就是如此），甚至不保证那个网页还存在。这也是搜索引擎和传统信息检索系统的一个重要区别。这种区别源于前述 Web 信息的基本特征。为了弥补这个差别，现代搜索引擎都保存网页搜集过程中得到的网页全文，并在返回结果列表中提供“网页快照”或“历史网页”链接，保证让用户能看到和摘要信息一致的内容。

第二节 搜索引擎的发展历史

早在 Web 出现之前，互联网上就已经存在许多旨在让人们共享的信息资源了。那些资源当时主要存在于各种允许匿名访问的 FTP 站点（anonymous FTP），