前 言

进入21世纪，互联网技术的发展及其所催生的“新经济”在世界经济格局中所占的分量呈现大幅上升的趋势。互联网的发展和应用趋势也为越来越多的人所关注。近几年来搜索引擎与信息获取技术随着万维网（WWW）这种方便易用的媒体的普及而得到了迅速的发展。由于互联网技术开放性的特点，使得网上的信息越来越丰富，这一方面为用户获取信息提供了很大的便利，另一方面由于信息量的飞速增长，使得用户查找所需信息时犹如大海捞针。新的信息获取技术——搜索引擎技术应运而生，并得到了飞速发展。搜索引擎已经成为在互联网上仅次于电子邮件的第二大网络应用。

信息获取技术从出现到现在已经有20多年的历史。在WWW出现之前，信息获取技术的发展比较缓慢，主要应用在图书馆、科技信息研究等专业部门，涉及的用户相对较少。在WWW出现之后，基于WWW的信息获取技术——搜索引擎技术出现并得到迅速发展。由于互联网

的开放性，使得搜索引擎可以面向普通用户，用户的需求推动其进一步发展。我国在信息获取领域起步较晚，在中文信息获取需求的驱动下，出现一些中文搜索引擎网站。但是从国内搜索引擎的应用效果和对搜索引擎技术的掌握上与发达国家比较，则仍然存在较大的差距，特别是在智能搜索引擎的开发、建设和应用水平方面差距更大。这种差距主要表现在两个方面：第一是搜索引擎的性能和搜索结果的质量与国外的搜索引擎相比存在很大的差距。这就是为什么国人在选择搜索引擎的时候普遍选择国外著名搜索引擎的缘故。第二是对搜索引擎与信息获取技术的掌握和应用上有待于提高和加强，应用人才急需培养，经验有待积累和总结。前者的改善应依靠于国内网络性能的整体改善和先进信息获取技术的应用；而后者的改善则更多地依赖于人们对搜索引擎和信息获取技术的掌握。

我的博士生张卫丰同志从本科学习阶段开始就跟随我进行搜索引擎的研究，后又跟我攻读硕士学位（因成绩优异被批准改为硕博连读），经过近五年的刻苦努力，在国内外发表了大量学术论文，取得了一系列研究成果。本书即是我们在对搜索引擎的有关问题进行深入研究的基础上，充分吸收现有研究成果，一起编写而成的。

本书涵盖了搜索引擎与信息获取技术的主要内容，力争系统全面地介绍信息获取的关键技术，并通过实例来说明，使得复杂的概念容易理解。本书主要包含三部分：第一部分（第 1 章到第 9 章）介绍信息获取的基本原理与技术；第二部分（第 10 章到第 11 章）主要介绍我们在搜索引擎与信息获取领域所做的工作和取得的最新科研成果，它主要是实现用户个性化搜索的相关技术；第三部分（第 12 章到第 13 章）面向普通用户的应用需求，分别从 Web 站点维护者的角度和普通用户使用搜索引擎的角度讨论了如何提高自己的网页在搜索引擎的排名和如何提高查询的搜索

精度。

我们希望本书的出版能够对搜索引擎的设计者、Web 站点的管理员以及广大用户有所裨益，也希望它能成为有关领域学生的学习参考书。

十分感谢清华大学出版社的同志为本书的出版所做的工作。

作　者

2002 年 2 月于南京

目 录

图 目 录

表格目录

第1章 概 述

1.1 引 言

随着计算机技术和互联网技术的飞速发展，信息获取已经从手工获取，到计算机信息获取，以及到现在的通过网络进行信息获取。网络的最大优点就是将大量的信息相互共享，而且只要通过一台接入互联网的计算机就可以方便地获取信息。利用互联网，用户一方面可以快速、方便地接触到各种信息，但是另一方面通过普通浏览的方式很难在信息的海洋中找到真正需要的信息。网络时代的信息量每8个月就翻一倍，如今的网页以10亿来计算，要在浩如烟海的网络世界寻找需要的信息，作为现代信息获取技术的主要应用——搜索引擎是必不可少的。搜索引擎正在不断地改变人们获取信息的方式。利用搜索引擎可以快速找到需要的信息。信息获取技术现在广泛应用于搜索引擎、数字图书馆等。

搜索引擎是仅次于门户的互联网第二大核心技术，伴随互联网的普及和网上信息的爆炸式增长，它越来越引起人们的重视。

搜索引擎技术的市场不仅限于门户网站、专业网站同样需要快速有效的搜索。此外，各个企业、机构自己的网站也是一个极其广阔的市场领域。

目前，国内不少企业花了很多钱构建了一个内容丰富的网站，