

# 搜索引擎



[搜索引擎 下载链接1](#)

著者:李晓明

出版者:科学出版社

出版时间:2012-5

装帧:平装

isbn:9787030342584

《搜索引擎:原理技术与系统(第2版)》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。

作者介绍：

李晓明：天网搜索引擎领域负责人

闫宏飞 王继民：天网搜索引擎项目负责人

目录: 第二版前言

第一版前言

第一章引论

第一节搜索引擎的概念

第二节搜索引擎的发展历史

第三节一些著名的搜索引擎

第四节小结

上篇Web搜索引擎基本原理和技术

第二章Web搜索引擎工作原理和体系结构

第一节基本要求

第二节网页搜集

第三节预处理

第四节查询服务

第五节体系结构

第六节小结

第三章Web信息的搜集

第一节概述

一、超文本传输协议

二、一个小型搜索引擎系统

第二节网页搜集

一、定义URL类和Page类

二、与服务器建立连接

三、发送请求和接收数据

四、网页信息存储的天网格式

第三节多道搜集程序并行工作

一、多线程并发工作

二、控制对一个站点并发搜集线程的数目

第四节如何避免网页的重复搜集

一、记录未访问、已访问URL和网页内容摘要信息

二、域名与IP的对应问题

第五节搜集信息的类型

第六节小结

第四章对搜集信息的预处理

第一节索引网页库

第二节网页编码识别

一、基本而重要的概念

二、常用字符编码

三、常用字符编码算法

四、字符的输入和显示

五、编码识别

第三节中文自动分词  
第四节分析网页和建立倒排文件

第五节小结

第五章信息查询服务

第一节检索的定义  
第二节查询服务的实现

一、结果集合的形成  
二、查询结果显示

第三节小结

中篇对质量和性能的追求

第六章可扩展搜集子系统

第一节天网系统概述和集中式搜集系统结构

一、天网系统结构  
二、集中式搜集系统

第二节利用并行处理技术高效搜集网页的一种方案

一、节点间URL的划分策略

二、关于性能的讨论

三、性能测试和评价

四、系统的动态可配置性设计

第三节天网分布式搜集系统

第四节对DeepWeb的认识

一、DeepWeb的成因

二、搜索DeepWeb的方法

第五节小结

第七章网页净化与消重

第一节网页净化与元数据提取

一、DocView模型

二、网页的表示

三、提取DocView模型要素的方法

四、模型应用及实验研究

第二节网页消重算法

一、消重算法

二、算法评测

第三节小结

第八章高性能检索子系统

第一节检索系统基本技术

一、系统设计与结构

二、索引创建

三、检索过程

第二节适于查询的网页索引结构

一、倒排索引结构

二、平面位置索引

第三节倒排索引压缩

一、倒排索引压缩技术

二、词典与倒排表的压缩

第四节索引剪枝

一、静态索引剪枝方法

二、动态索引剪枝方法

第五节混合索引技术

一、混合索引的原理

二、混合索引的实现

第六节倒排文件缓存机制

一、倒排文件缓存

二、负载特性

### 三、缓存策略的选择

#### 第七节小结

##### 第九章相关排序与系统质量评估

###### 第一节传统IR的相关排序技术

###### 第二节链接分析与相关排序

###### 一、链接分析

###### 二、Web查询模式下的新信息

###### 第三节相关排序的一种实现方案

###### 一、形成网页中词项的基本权重

###### 二、利用链接的结构

###### 三、收集用户反馈信息

###### 四、计算最终的权重

##### 第四节信息检索技术评估

###### 一、信息检索技术评估指标

###### 二、TREC和CWIRF信息检索评估

###### 三、搜索引擎技术评估

#### 第五节小结

##### 下篇Web信息资源的组织与应用服务

##### 第十章大规模Web历史网页仓储系统的构建

###### 第一节国外Web历史网页保存现状

###### 一、Internet Archive

###### 二、PANDORA

###### 三、其他相关Web保存项目

###### 第二节中国Web信息博物馆的系统设计

###### 一、WehInfoM all的设计目标

###### 二、Web InfoMall的体系结构

###### 第三节历史网页的存储

###### 一、数据的组织

###### 二、存储结构

###### 三、数据管理与压缩

###### 四、存储性能

###### 第四节数据访问

###### 一、PageID的索引

###### 二、URL的索引

###### 三、数据服务

###### 四、性能与优化

###### 第五节网页的格式保存

#### 第六节小结

##### 第十一章大规模We非网页信息仓储系统的构建

###### 第一节网络资源库藏相关工作

###### 一、Ibiblio

###### 二、Internet Archive

###### 三、Wikimedia

###### 四、中国互联网数字资源财富库藏

###### 第二节CDAL系统概况

###### 第三节CDAL系统设计

###### 一、系统体系结构

###### 二、可扩展的存储组织方案

###### 第四节网络资源描述信息获取

###### 一、Ontology概述

###### 二、描述信息获取机制

###### 三、改进查询的方法

###### 四、改进排序的方法

###### 第五节基于局部聚类思想的共现词汇算法

一、基本定义  
二、FDC共现词汇算法  
第六节小结  
.....

第十二章中文网页自动分类与聚类  
第十三章开放域问答系统  
参考文献  
附录术语  
..... (收起)

[搜索引擎](#) [下载链接1](#)

标签

搜索引擎

系统设计

数据库

已购买

IT产业

评论

....是自己看不懂

-----  
国内还算不错的书，浅显易懂

[搜索引擎](#) [下载链接1](#)

## 书评

国内的著作，特别是冠以学术的东西，不论是可读性还是内容的质量都很糟糕，但这本却是例外。

300多页的内容把搜索引擎的原理讲的很清晰，此书成于2005年，搜索引擎领域的发展发生了极大的变化，但是基本的原理还是想通的，需要解决的问题还是一致。比如分词，检索还有存储，书...

---

主要是由北大李晓明那个实验室所发表的论文组成，很多地方偏学术，但是在国内这本书应该是最好的搜索引擎方面的书籍，推荐大家作为搜索引擎入门的书籍，要了解最新的搜索引擎技术还是要多读SIGIR,WWW等会议的相关论文。

读完这本书，可以进一步学习

---

最近埋头苦看各种搜索引擎原理的书籍，当然我是一个入门者，所以从入门者的角度来说几句吧~

首先我的背景是给老外打工，所以几乎都是英文，挑选这本书仅仅是偶然，其实我想找的是另外一本

《信息检索实践》，在误点的情况下下载了本书，然后读完了，通读一遍的感觉是里面还不...

---

因为以后要从事搜索开发的工作，所以公司推荐了这本书。书挺薄的，前后一个月看完吧，总体感觉还行。这本书把搜索引擎相关的各项技术基本都做了介绍，比较全面，算是为数不多、质量不错的入门书籍。说说缺点吧，这本书应该是北大n多论文拼出来的，有一种前后不太连贯的感觉；因...

---

适合搜索引擎入门时阅读：内容还算比较全面，涉及到SE的各个方面，但很多技术的确有点老了，毕竟这本书出的比较早  
建议配合TSE 代码阅读

---

谷歌搜索引擎优化seo？能不能自动优化？

万水千山总是情，你在那头，我在这头，默默的祝福，深深的思念，就让清风捎去我的问候，一切都未曾改变，你永远是我的牵挂！

走过山山水水，路过春夏秋冬，克服千险万阻，为你搜集一件无忧衫，前身是吉祥如意，后背是平安喜气，袖子是快...

是阅读该书及TSE源代码非常好的参考资料，可以作为想从零了解搜索引擎的朋友的入门资料。  
可以参考某人的csdn博客中的笔记：<http://blog.csdn.net/column/details/inside-tse.html>

北大天网实验室出的一本书，主要结合了天网的实践，并有一套称为TSE的C++代码。全书分为三部分。除了第三部分涉及更多的高级问题，理论性较强，书中描述也不太详细之外，前两部分对于非专业人士了解IR系统的“原理，技术与系统”很有帮助。该书对网页抓取，文本分析，索引建...

[搜索引擎 下载链接1](#)