

《走进搜索引擎》

图书基本信息

书名：《走进搜索引擎》

13位ISBN编号：9787121049224

10位ISBN编号：7121049228

出版时间：2007-1

出版社：电子工业出版社

作者：梁斌

页数：272

版权说明：本站所提供下载的PDF图书仅提供预览和简介以及在线试读，请支持正版图书。

更多资源请访问：www.tushu000.com

《走进搜索引擎》

内容概要

《走进搜索引擎》由搜索引擎开发研究领域年轻而有活力的科学家精心编写，作者将自己对搜索引擎的深刻理解和实际应用巧妙地结合，使得从未接触过搜索引擎原理的读者也能够轻松地在搜索引擎的大厦中遨游一番。《走进搜索引擎》作为搜索引擎原理与技术的入门书籍，面向那些有志从事搜索引擎行业的青年学生、需要完整理解并优化搜索引擎的专业技术人员、搜索引擎的营销人员，以及网站的负责人等。《走进搜索引擎》是从事搜索引擎开发的工程技术人员难得的参考书，也可作为大中专院校相关专业的教学辅导书。

《走进搜索引擎》

作者简介

梁斌毕业于南京大学，获得软件工程硕士学位，曾经发表过多篇论文，获得1项国家专利，作者主要的兴趣方向包括数据挖掘、Web挖掘、搜索引擎和软件工程等，目前在清华大学信息科学与技术国家实验室从事搜索引擎相关研究工作。

书籍目录

第一章 引言

第一节 什么是搜索引擎

第二节 搜索引擎的发展简史

搜索引擎的发展历史

第三节 搜索引擎大事快览

第四节 国内著名搜索引擎

百度 (www.baidu.com)

中搜 (www.zhongsou.com)

天网 (e.pku.edu.cn)

搜狗 (www.sogou.com)

参考文献

第二章 搜索引擎概貌

第一节 搜索引擎的主要需求

查得快

查得全

查得准

查得稳

第二节 搜索引擎的大系统

搜索引擎的体系结构

第三章 搜索引擎的下载系统

第一节 爬虫的发展历史

世界上第一个爬虫

爬虫的发展历程

第二节 万维网及其网页分析

蝴蝶结型的万维网

万维网的直径

万维网的规模及变化特征

网页的特征

第三节 有关爬虫的基本概念

爬虫

种子站点

URL

Backlinks

第四节 网页抓取原理

telnet和wget

从种子站点开始逐层抓取

不重复抓取策略

网页抓取优先策略

网页重访策略

Robots协议

其他应该注意的礼貌性问题

抓取提速策略 (合作抓取策略)

第五节 网页库

第六节 下载系统回顾及未来发展

参考文献

第四章 搜索引擎的分析系统

第一节 知识准备

HTML语言

锚文本 (anchor text)

半结构化数据 (Semi-structured data)

第二节 信息抽取及网页信息结构化

网页结构化的目标

建立HTML标签树

通过投票方法得到正文

网页结构化过程回顾

第三节 网页查重

网页查重技术发展历史

网页查重实现方法

第四节 中文分词

什么是中文分词

通过字典实现分词

通过统计学方法实现分词

第五节 PageRank

PageRank的由来

PageRank的基本想法

PageRank的计算公式

PageRank的计算方法

第六节 分析系统结构图

参考文献

第五章 搜索引擎的索引系统

第一节 知识准备

信息

索引

倒排索引、倒排表、临时倒排文件、最终倒排文件

其他概念

第二节 全文检索

全文检索

第三节 文档编号

编号的本质

文档编号的方法

游程编码

第四节 倒排索引

经典的倒排索引

正排索引 (前向索引)

倒排索引

第五节 数据规模的估计

齐普夫法则

布尔检索模型下的索引规模估计

第六节 涉及存储规模的一些计算

正排表与倒排表的合并

多个临时倒排文件的归并

倒排索引分布式存储

倒排文件缓存

倒排索引词典统计信息的计算

第七节 倒排索引文件的创建过程

创建倒排表

计算统计信息

参考文献

第六章 搜索引擎的查询系统

第一节 知识准备

什么是信息熵

检索和查询的区别

检索词和查询词的区别

自动文本摘要(Automatic Text Summarization)

第二节 网页信息检索

早期的检索模型

向量空间模型 (Vector Space Models)

关键词权重的量化方法TF/IDF

搜索引擎采用的检索模型

多文档列表求交计算

检索结果排序

堆排序

第三节 中文自动摘要

自动摘要的发展历史

自动摘要的含义和实现

第四节 生成搜索结果页

生成搜索结果页

第五节 搜索结果页的缓存

搜索结果页的缓存

第六节 推测用户查询意图

查询分类

推测信息类、事物类的查询意图

第七节 查询系统的当前热点和发展方向

查询系统的当前热点

参考文献

第七章 搜索引擎的其他话题

第一节 搜索引擎问与答

为什么搜索引擎的搜索速度这么快

为什么搜索引擎能够返回那么多的查询结果

为什么搜索引擎总能返回最想要的结果

搜索引擎如何大规模存储网页的

什么是SEO

什么是元搜索引擎

搜索引擎认为的作弊行为是哪些

如何进一步学习和了解搜索引擎发展的最新成果

第二节 搜索引擎未来的发展

新兴的搜索产品

搜索技术的未来

参考文献

附录A 搜索引擎系统结构全视图

精彩短评

- 1、我不是做搜索引擎相关工作的,但是对这方面的东西很有兴趣,梁总这本书让我这个外行走进了这个领域
- 2、书里面的字也太大了把??段与段之间间隔太大了,是不是在凑页?这个就算了,内容讲的也就那样。。。
- 3、结构还是很清晰的,想了解搜索引擎的入门书籍。或许是我身在此行吧,所以读得认真些。如果不认真的话,这本书确实很难读进去。
- 4、本书的参考文献可以作为搜索引擎入门阅读指南
- 5、搜索引擎的介绍内容比较单薄,不过同类的书本来就不多,也没别的选择了
- 6、定位为入门科普书籍,讲的不是很深。而且书的每一页其实真正的字数很少,有大片的旁边栏是空白,美其名曰:读书笔记
- 7、这书写的都是理论的东西,作为搜索基础知识学还不错
- 8、呵呵,这本书的字好大,说明内容不是很多。。。当成入门书还是可以的。。。
- 9、介绍se原理中,算比较详细的一本
- 10、扫了一遍,像是本科普书
- 11、太入门级了,不深入。空白太多,字体太大,如果按照一般的书那种印刷方式,估计也就100页。而且书里面有不少缺字,错字,公式下标不明确,英文之间的空格也有时有时没有的,感觉好像没怎么校对似的。反正这本书不太值这么多钱。
- 12、质量有问题,有好多页重复印刷,根本看不清内容
- 13、一般性的讲解了搜索引擎的工作原理。操作性不强,偏重理论。
- 14、很好的书,讲的很浅显易懂。
- 15、只能说这本书作为一本科普读物还是不错的,如果说作为技术书,那还是有点不够格
- 16、入门必备.
- 17、帮同学买的书,翻了翻还不错
- 18、关于本书的一些问题,给大家一个解释。首先从事搜索引擎工作的圈子很小,能够进入这个行业有一定门槛,信息检索的技术从研究界而来,商业化以后,研究界的水平已经大大落后,而业界的技术一般均不公开,在这种背景下,普通人能够接触到的“最深入”的技术莫过于北大李晓明教授的搜索引擎一书,而业界的高端技术都是不传之秘,这些本书也不能公开的,如果想更深入的学习和研究不妨去搜索引擎公司锻炼一下。虽然如此,搜索引擎的技术可以被其他行业借鉴,有志从事搜索引擎的同学们可以预先学习,这是本书的主要宗旨,我的目标就是把北大李晓明的搜索引擎书中不够深入的地方,深入一些,系统一些。爬虫、TFIDF的物理意义解释,PageRank的计算,索引规模估计,索引创建,自动摘要部分,都是李这本书没有或不够深入的。这本书的大部分内容,都是从研究界论文中整理而来,只有TFIDF的物理意义解释是我本人独创的,为了方便读者理解(没有使用交叉熵,KL距离这些比较难懂的概念去解释),以及我举得一些例子。读者认为这本书太浅,没有含量,是我即高兴又难过,高兴在于读者的水平都很高,难过在于这些一流学者的研究成果没有得到应有的尊重,也许你接受过高等教育,但是你不会鄙视小学学过的入门知识吧。这本书已知有一些错别字,但并不多,影响阅读的错别字就更少了。这本书完全是我一个人写成,校对也做了很多遍,但错误在所难免,《编程之美》这本书大家都知道,作者团队十分庞大,但错误也是比较多的,做一件完美的事情,是每个人的愿望,但有时确实很难,如果您看到错误,可以与我联系,协助改正,或者在评论中写明也帮助其他读者。公式下标有问题我至今没有发现。最后就是这本书的排版,这可能是最大的批评,我想知识传播是有代价的,有些读者想如果我不是为了赚钱,完全可以写出来在网络上共享,如果有写作经历的人就知道写一本书的稿费实在有限,和搜索引擎业界待遇相比,差距极大,这也是业界无人出来写书的一个原因。我曾打算捐出稿费以表态度,但这样做无疑会提高道德的标准,对其他写书的人产生不好的影响,孔子有个学生很有钱,年关收账的时候将账目烧毁,乡亲们很感谢,孔子知道后批评了他,因为这样做提高的道德的标准,脱离了当时的物质发展基础,以后的地主收账就成为不道德的行为了。出版社是一个盈利的机构,他们需要赚钱,才能更好的传播知识,才有可能进入良性的循环,我长期在水木社区解答网友的问题,水木社区的很多网友我都是无偿送书的,至今送了不下50本。<http://www.newsmth.net/frames.html?mainurl=%2Fbbsdoc.php%3Fboard%3DSearchEngineTech>大家

《走进搜索引擎》

有问题也可以来水木找我。每位同学的批评对我来说都是莫大的帮助，我也在不断反思自己的问题，但是我的目标不会变，我要把搜索引擎的技术进行力所能及的推广，做一些有价值有意义的事情。我常常用搜索引擎搜索对我这本书的评价，大部分还是积极的，令我十分欣慰，为这些从中获取知识的人感到鼓舞，这种精神会让我继续努力，继续奋斗。

19、雷得要死。。

20、个人觉得这本书适合入门的互联网工程师，把很多知识点都简易介绍了，你不可能期待一本书可以让你完全懂得整个技术，入门也是很关键的一个门槛

21、比较清晰的介绍了搜索引擎，搞seo的一定要去看看。打四分也是从seo价值。

22、收到这本书后，用了一个星期把这本书看了一遍。感觉这是一本介绍搜索引擎入门很好的书，不过还是偏向于理论。看完这本书后，感觉对搜索引擎的结构有了一个大概的了解，当然还是比较初步的。不过，正如许多网友所说的那样，这本书的排版可能是最有争议的地方了。个人感觉，这本书空白的地方太大、太多了，完全可以更加紧凑一些。但是，这并不影响它是一本了解搜索引擎的好书！

23、出版社“刮”不知耻！：）当年，杨澜那本《凭海临风》也是，每页左右留很多空挡，然后整本书充满了照片，哈，文字不知有几许。

24、挺浅显易懂的

25、这本书买回来看了看都是一些深入的知识；不适合我们初学者！

26、泛泛而谈，没有实质

27、内容太入门级了，出版格式太差劲，好像在凑字数

28、刚看了一点，还不错！

29、作为一个入门的书籍还可以，偏重于一些理论。不满意的是，该书用大大的字号，加上许多空白来充页数，给人的感觉非常不好。还不是实在些，该多少页就排多少页，页数加多了最然能标个高价，很反感这种做法。

30、我买书的时候也看到评论中对这本书有些争论。不过这本是还是非常适合我这样的读者的。我总结一下它适合的人群，以帮助后来者选择：适合初学搜索引擎原理的人；不适合只想应用搜索引擎或已经熟知其原理的人。适合对书价不敏感的人。适合眼睛不好，新欢看大字体的人。

31、深入浅出，入门教材，这就称得上好书。谁不是从小白开始入行的???

我觉得这书让我收获不小

32、搜索引擎的四个系统的介绍，非常专业。

虽然是一本理论书籍，但是大多数段落我居然看懂了。特别是PageRank的部分写的太好了，总算理解这个是怎么回事了。

33、关键是看内容吧。

我只看过作者翻译的深入搜索引擎--海量文本索引技术，这本是很不错的。

34、这书刚出的时候读过，说实话感觉挺水的。。。

35、有点言过其实，但是还是国内相关书籍里面比较严谨的

36、反正我图书馆借的.....

37、书还凑活，细节之处不够细致。搜索引擎的算法介绍的比较少，是一本概论性质的书。

38、不如李的权威

李指的是谁？

39、adaptingtothosewhohaslittleconceptaboutsearchingengineering.

40、简单了解下原理,还算有趣

41、呃，书评不咋地~当入门看看了

42、07年买的书，前阵子利用一些零碎的时间，两周把书看了两遍。书写的浅显易懂，很多相当复杂的概念写的很容易理解。比如信息熵，TF/IDF，zipf法则等都讲的很清晰，易于理解。书中对很多问题提出了解决方案，更偏向于实践，书上的字很大，刚好适合阅读，含金量较高，建议有一定算法基础的同学都可以细读。是国内介绍搜索最好的两本书之一，另一本就是“搜索引擎：原理、技术与系统——华夏英才基金学术文库”了，顶梁老师一把。

43、空白实在太多了，行距超常的大。。。适合视力不好的人以及对价格不敏感的人看。。

内容还是比较通俗易懂，结合李晓明的那本看效果不错。

《走进搜索引擎》

- 44、 比较客观地说这本书还是不错的，不如李的权威，却更通俗易懂。
不知道作者除了书上介绍之外，还有什么来头
- 45、讲的内容挺全，就是太泛
- 46、拿到手的感觉就一个词，失望！居然花了40块钱买这么一本书。内容算是入门级的，这倒无可厚非，和书的标题还比较符合。印刷、排版以及校对水平都太差了，感觉是把字体弄大，行间距扩大，这才成了一本书，否则能不能到100页都成问题。里面居然有这样的句子：文献（xxx,19xx）blabla知道文献的作者，年份，文献的名称咋就不写出来呢？感觉就像是篇论文综述翻译了一遍。。。
- 47、读了两遍，虽然大致的跟现在有些不同，但是总体来说可以算是一本SEO的入门书籍
- 48、基础理论
- 49、2012.05.14《走进搜索引擎》92分钟。这并不是一本有关SEO的书，实际上，是从原理上对搜索引擎的四大系统：下载系统、分析系统、索引系统和查询系统做了详细的讲解！更好的理解搜索引擎的工作原理，当然就能更好的理解SEO的本质！
- 50、现在的书很多都是这样滴。特特别是一些名人写的书。
- 51、翻书一看就是巨大的字体，滥竽充数的味道很浓。而且
- 52、有心的人看了这书，一定会获益匪浅的
- 53、 大致看了一遍，要说这本书的唯一缺点，就是价格比同类书高了一点，不过该书内容绝对对得起它的价格，确实有特色之处，怪不得能得到王小川的推荐。这本书在我看过的同类书中我觉得是数一数二的。个人意见，供大家参考。
- 54、 李晓明那本书很经典，这本书比较通俗，不过价格偏高
- 55、是一本非常好的关于搜索引擎的书，了解搜索引擎从这里开始！
- 56、所以书最好是先翻一下再买的好！先读一下书评也不错！网上购书的弊端于此.....
- 57、后面基本全在说算法，对seo狗并没有什么卵用。
- 58、书写得还不错，比较通俗
- 59、没有一定软件开发基础的人是不知道在说什么的，比较难懂。有点后悔买。慢慢琢磨吧。或许真的能得到一些想要的内容。
- 60、大致了解了搜索引擎，确实让我走近了~~~~~
- 61、真的假的。。。pagerank那部分不像书里这么写还能怎写。。。顶多算是科普读物吧这书
- 62、多谢小米多看阅读商店7月31号的限免，以每个工作日下班地铁时间读完。作者应该有非常不错的工程经验，但是内容比较全面但不深入，有些地方泛泛而谈，像综述却不如好的综述
- 63、对，就是有骗钱嫌疑
完全的科普读物
嗯.....那本书好像也.....
- 64、内容偏理论，对于刚入门的人来说，还是有一定的价值
- 65、非常粗糙，远不如当时翻译的那本Managing gigabyte啊，不过给出的提示都比较靠谱，入门书吧
- 66、 作者倒是认真的，给大爷大妈们写了本介绍搜索引擎的“专业书”。
电子社居然两三百字就凑成一页，弄些图片来填充，每页还留有些“读书笔记”的硕大空挡，居然凑满了272页，卖你50大元没商量！还在封面上“刮”不知耻地写上“打造优质搜索引擎的第一书！”
- 我靠！
- 67、北大李晓明的《搜索引擎：原理、技术与系统》是本不错的书。
- 68、李晓明教授
- 69、没有实际的内容。
- 70、这本书偏向研究型而不是工程型，正好适合我的胃口，里面有很多思路值得借鉴，不错！缺点就是空白实在太多了，行距超常的大，等于买了一半白纸。。。
- 71、 这本书写的不是一般的差，有骗钱的嫌疑。
完全是写给外行的人看的，建议做it的人不要买，小学生可以当作科普读物。
我建议要了解搜索引擎的话，还是应该要看北大李晓明写的《搜索引擎：原理、技术与系统》
- 72、关于排版：卖了一半的白纸，不值；关于错别字：有好多，不好；关于内容：很细致，不错；建议：对于搜索引擎的作用，发展趋势，在信息化领域的地位，搜索引擎的分类等等框架性的东西需要

《走进搜索引擎》

再写一点，现有的很好，但是技术性强了一点，和信息化的大框架结合的不紧密

73、用一周的时间看了一遍，感觉对整个搜索引擎的系统架构有了很清楚的理解，具体每个部分的实现细节还需要仔细阅读并参考相关资料或数据。个人感觉书写的很系统，受益匪浅。另外排版也很好，这可能是大部分其他读者不认可的地方，因为我比较喜欢做读书笔记，所以大部分的空白都能派的上用场，可以把本书作为一个学习搜索引擎的提纲，然后在相关章节记录深入学习的体会和总结。非常好，感谢作者，希望有机会能认识，多和您讨教，谢谢您。

74、内容过于技术理论化，seo实操性不强！

75、看完出去装逼有谱

76、2014第一本. 搜索引擎入门, 包括下载, 分析, 索引, 查询四个系统. 作为一本书来说简单的地方太简单, 数值公式推导要不是点了相关技能点还是有困难的不过可以直接记结论. 相关引用倒是列的很全, 大段的留白居然用水印说是阅读笔记orz.

77、很新的书，包装不错，内容还没来得及看

78、这本书的定位是让有一定知识背景的人了解搜索引擎，从这个角度来看，非常成功。

不适合资深专业人员看。

79、快速读了一遍，入门书，结构还算清晰，内容简陋了些。

80、两本书的侧重点和知识结构不一样吧，都能有收获

书嘛，都贵.....

81、本来以为是代码方面的知识，谁知道是数学的，涉及许多的高等数学，没心机看完。

82、通俗易懂，很好的一本入门书籍，不错。。。

83、买回去了本书的确专业，但是能让我这个计算机外行看到“搜索引擎”的核心，很超值。

《走进搜索引擎》

精彩书评

- 1、比较客观地说这本书还是不错的，不如李的权威，却更通俗易懂。不知道作者除了书上介绍之外，还有什么来头
- 2、空白实在太多了，行距超常的大。。。适合视力不好的人以及对价格不敏感的人看。。内容还是比较通俗易懂，结合李晓明的那本看效果不错。
- 3、这本书的定位是让有一定知识背景的人了解搜索引擎，从这个角度来看，非常成功。不适合资深专业人员看。
- 4、大致看了一遍，要说这本书的唯一缺点，就是价格比同类书高了一点，不过该书内容绝对对得起它的价格，确实有特色之处，怪不得能得到王小川的推荐。这本书在我看过的同类书中我觉得是数一数二的。个人意见，供大家参考。
- 5、搜索引擎的四个系统的介绍，非常专业。虽然是一本理论书籍，但是大多数段落我居然看懂了。特别是PageRank的部分写的太好了，总算理解这个是怎么回事了。
- 6、这本书写的不是一般的差，有骗钱的嫌疑。完全是写给外行的人看的，建议做it的人不要买，小学生可以当作科普读物。我建议要了解搜索引擎的话，还是应该要看北大李晓明写的《搜索引擎：原理、技术与系统》
- 7、感觉这本书写的很好啊，逻辑非常清晰，为读者描述了搜索引擎的骨架，介绍了搜索引擎的四大系统以及互相协作。其他人的评论重点在吐槽板式，我看的电子版没有太注意这个问题。个人认为作为刚接触搜索引擎，想要长期发展的读者，这是一本很好的“渡船”书籍。对于技术高手应该从序及引言部分就能发现是否适合自己吧。书的定位明确，内容易懂，架构清晰，笔者认为是一本不错的书籍。
- 8、作者倒是认真的，给大爷大妈们写了本介绍搜索引擎的“专业书”。电子社居然两三百字就凑成一页，弄些图片来填充，每页还留有些“读书笔记”的硕大空挡，居然凑满了272页，卖你50大元没商量！还在封面上“刮”不知耻地写上“打造优质搜索引擎的第一书！”我靠！

版权说明

本站所提供下载的PDF图书仅提供预览和简介，请支持正版图书。

更多资源请访问:www.tushu000.com