<<搜索引擎>>

图书基本信息

书名：<<搜索引擎>>

13位ISBN编号：9787111288084

10位ISBN编号：7111288084

出版时间：2010-6

出版时间：机械工业出版社

作者：W.Bruce Croft,Donald Metzler,Trevor Strohman

页数：309

译者：刘挺,秦兵,张宇,车万翔

前言

本书综述了信息检索中的重要问题，并介绍了这些问题如何对搜索引擎的设计与实现产生影响。
本书并不是按照相同的详细程度描述每个主题，相反，我们侧重于那些对于实现搜索引擎组件以及组件背后的信息检索模型最重要的部分。
网络搜索引擎显然是一个重要的话题，我们主要覆盖了在网络上使用的搜索技术，但搜索引擎在其他场合中也有应用，这就是为什么我们重点强调各种搜索引擎背后的信息检索理论与概念的原因。
本书的目标读者群主要是计算机科学或计算机工程的本科生，但研究生也会发现本书是有用的，此外，本书也适合多数情报科学专业的学生。
最后，无论读者是什么背景，通过阅读本书都可以对他们动手开发搜索引擎有所帮助。
本书中涉及数学知识，但并不深奥。
’书中也有代码和程序设计的练习，但对于那些已经完成了基本计算机科学与程序设计课程的人来说，完全可以掌握。
每章末尾的练习使用了被称为Galago的基于Java的开源搜索引擎。
Galago既是为本书所设计的，也借鉴了Lemur和Indri项目的经验。
换句话说，这是一个功能齐全的能够支持真正应用的搜索引擎。
许多编程练习都是针对Galago组件的使用、修改和扩展。
内容在第1章，我们对信息检索及它与搜索引擎的关系做了一个高层次的回顾。
在第2章，我们描述了搜索引擎的架构，这一章全面介绍搜索引擎的各个组件，但没有涉及细节问题。
在第3章，我们关注于爬取、文档信息源和其他用于获取被检索信息的技术。
第4章描述了文本的统计特征，以及用来处理和识别重要特征的技术，并为建立索引做准备。
第5章描述了怎样为有效的搜索建立索引，以及怎样利用索引处理查询。
在第6章，我们描述了怎样处理查询，并把它们转换为更好的形式，以表达用户的信息需求。

内容概要

本书介绍了信息检索（IR）中的关键问题，以及这些问题如何影响搜索引擎的设计与实现，并且用数学模型强化了重要的概念。
对于网络搜素引擎这一重要的话题，书中主要涵盖了在网络上广泛使用的搜索技术。

本书适用于高等院校计算机科学或计算机工程专业的本科生、研究生，对于专业人士而言，本书也不失为一本理想的入门教材。

作者简介

作者：（美国）克罗夫特（W.Bruce Croft） （美国）Donald Metzler （美国）Trevor Strohman 译者：刘挺 秦兵 张宇 等克罗夫特（W.Bruce Croft），马萨诸塞大学阿默斯特分校计算机科学特聘教授、ACM会士。
他创建了智能信息检索研究中心，发表了200余篇论文，多次获奖，其中包括2003年由ACM SIGIR颁发的Gerard Salton奖。
Donald Metzler，马萨诸塞大学阿默斯特分校博士，是位于加州Santa Clara的雅虎研究中心搜索与计算机广告组的研究科学家。
Trevor Strohman，马萨诸塞大学阿默斯特分校博士。
他开发了Galago搜索引擎，也是Indri搜索引擎的主要开发者。

<<搜索引擎>>

书籍目录

章节摘录

插图：2.查询转换查询转换组件包括一系列的技术，这些技术用于在生成排好序的文档之前和之后改善初始查询。
最简单的处理涉及一些对文档进行文本转换的技术。
在查询文本上，需要进行词素切分、停用词去除和词干提取这些工作，以生成与文档词项具有可比性的索引词。
拼写检查（spell checking）和查询建议（query suggestion）是查询转换中的技术，生成与用户初始查询相似的输出。
在这两种情况下，向用户提供初始查询的一些候选查询，这些候选查询可能纠正了拼写错误或者是对用户所需信息的更规范的描述。
这些技术通常会导致为网络应用搜集大量的查询日志（query log）。
查询扩展（query expansion）技术是对查询进行推荐或者增加一些额外的词项，但通常都是在对文档中词项的出现情况分析的基础上进行的。
该分析通常是用不同的信息源，如整个文档集合、检索到的文档或者用户计算机上的文档。
相关反馈（relevance feedback）是一种查询扩展技术，利用用户认为相关的文档中出现的词项对查询进行扩展。
3.结果输出结果输出组件负责对相关组件得到的排好序的文档的结果进行显示。
可能包含的任务有生成网页摘要（snippets）来对检索到的文档内容进行概括；强调（highlighting）文档中重要的词和段落；对输出结果聚类以找到文档相关的类别；以及将相应的广告增加到结果显示中。
在涉及多种语言的应用系统中，结果可能会被翻译成同一种的语言。

编辑推荐

《搜索引擎:信息检索实践》：计算机科学丛书

版权说明