<<搜索引擎>>

图书基本信息

书名：<<搜索引擎>>

13位ISBN编号：9787111282471

10位ISBN编号：7111282477

出版时间：2009-10

出版时间：机械工业出版社

作者：（美）克罗夫特

页数：520

前言

This book provides an overview of the important issues in information retrieval, and how those issues affect the design and implementation of search engines. Not every topic is covered at the same level of detail. We focus instead on what we consider to be the most important alternatives to implementing search engine components and the information retrieval models underlying them. Web search engines are obviously a major topic, and we base our coverage primarily on the technology we all use on the Web,l but search engines are also used in many other applications. That is the reason for the strong emphasis on the information retrieval theories and concepts that underlie all search engines.The target audience for the book is primarily undergraduates in computer science or computer engineering, but graduate students should also find this useful. We also consider the book to be suitable for most students in information science programs. Finally, practicing search engineers should benefit from the book, whatever their background. There is mathematics in the book, but nothing too esoteric. There are also code and programming exercises in the book, but nothing beyond the capabilities of someone who has taken some basic computer science and programming classes.

<<搜索引擎>>

内容概要

本书介绍了信息检索（1R）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。
对于网络搜索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

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。

书籍目录

章节摘录

插图：After documents have been converted to some common format, they need to bestored in preparation for indexing. The simplest document storage is no document storage, and for some applications this is preferable. In desktop search, for example, the documents are already stored in the file system and do not need to be copied elsewhere. As the crawling process runs, it can send converted documents immediately to an indexing process. By not storing the intermediate converted documents, desktop search systems can save disk space and improve indexing latency.Most other kinds of search engines need to store documents somewhere. Fast access to the document text is required in order to build document snippetsz for each search result. These snippets of text give the user an idea of what is inside the retrieved document without actually needing to click on a link.Even if snippets are not necessary, there are other reasons to keep a copy of each document. Crawling for documents can be expensive in terms of both CPU and network load. It makes sense to keep copies of the documents around instead of trying to fetch them again the next time you want to build an index. Keeping old documents allows you to use HEAD requests in your crawler to save on bandwidth, or to crawl only a subset of the pages in your index.Finally, document storage systems can be a starting point for information extraction （described in Chapter 4）. The most pervasive kind of information extraction happens in web search engines, which extract anchor text from links to store with target web documents. Other kinds of extraction are possible, such as identifying names of people or places in documents. Notice that if information extraction is used in the search application, the document storage system should support modification of the document data.

编辑推荐

《搜索引擎:信息检索实践(英文版)》：经典原版书库。

版权说明