<<搜索引擎与信息检索教程>>

图书基本信息

书名：<<搜索引擎与信息检索教程>>

13位ISBN编号：9787508453941

10位ISBN编号：7508453948

出版时间：2008-4

出版时间：水利水电出版社

作者：袁津生 等编著

页数：278

<<搜索引擎与信息检索教程>>

内容概要

随着搜索引擎技术的发展和不断完善，越来越多的人开始对搜索引擎原理和技术进行研究，越来越多的人喜欢上了搜索引擎。

本书从教学的角度出发，全面地阐述了搜索引擎的技术和信息检索技术，包括：搜索引擎的基本原理与技术、搜索引擎的数据结构和搜索引擎的爬虫、信息获取与信息检索技术、分类与聚类技术以及Web信息检索技术。

本书适合高等院校计算机科学与技术专业及相关专业的高年级学生和研究生阅读参考，也适合相关领域的工程技术人员参阅。

<<搜索引擎与信息检索教程>>

书籍目录

章节摘录

第7章 搜索引擎概述Internet上的信息量之大、范围之广、用户之多都比以往任何时候表现得突出，然而如何从浩瀚的信息海洋中得到所需要的信息就显得更加重要。
网络搜索引擎的出现从某种程度上解决了这个问题，它是目前比较有效的网上信息获取方法，多数网上用户使用搜索引擎来获得所需的信息。
据CNNIC的统计，用搜索引擎搜索仅次于电子邮件的应用。
目前，网上比较有影响的搜索工具中，中文的有：Google、百度（Baidu）、北大天网、爱问（iask）、雅虎（Yahoo）、搜狗（Sogou）等搜索引擎；英文的有：Yahoo、AltaVista、Excite、Infoseek、Lycos、Aol等。
另外还有专用搜索引擎，例如，专门搜索歌曲和音乐的；专门搜索电子邮件地址、电话与地址和公众信息的；专门搜索各种文件的FTP搜索引擎等。
本章主要介绍搜索引擎的概念、搜索引擎的发展史、搜索引擎的分类以及一些著名的搜索引擎。
1.1 搜索引擎的概念搜索引擎并不真正搜索互联网，它搜索的实际上是预先整理好的网页索引数据库，真正意义上的搜索引擎，通常指的是收集了Internet上几千万到几十亿个网页并对网页中的每一个词（即关键词）进行索引，建立索引数据库的全文搜索引擎。
当用户查找某个关键词的时候，所有在页面内容中包含了该关键词的网页都将作为搜索结果被搜出来。
在经过复杂的算法进行排序后，这些结果将按照与搜索关键词的相关度高低依次排列。
现在的搜索引擎已普遍使用超链分析技术，除了分析索引网页本身的内容，还分析索引所有指向该网页的链接的URL、Anchor Text，甚至链接周围的文字。
所以，有时候，即使某个网页A中并没有某个词，比如“信息检索”，但如果有网页B中有链接“信息检索”指向这个网页A，那么用户搜索“信息检索”时也能找到网页A。
而且，如果有越多网页的“信息检索”链接指向网页A，那么网页A在用户搜索“信息检索”时也会被认为更相关，排序也会越靠前。
搜索引擎的原理可以分为四步：从Internet网上抓取网页、建立索引数据库、在索引数据库中搜索排序、对搜索结果进行处理和排序。
（1）从Internet上抓取网页。
利用能够从Internet上自动收集网页的Spider系统程序，自动访问Internet，并沿着任何网页中的所有URL爬到其他网页，重复这过程，并把爬过的所有网页收集回来。
（2）建立索引数据库。
由分析索引系统程序对收集回来的网页进行分析，提取相关网页信息（包括网页所在URL、编码类型、页面内容包含的关键词、关键词位置、生成时间、大小、与其他网页的链接关系等），根据一定的相关度算法进行大量复杂的计算，得到每一个网页针对页面内容中及超链接中每一个关键词的相关度（或重要性），然后用这些相关信息建立网页索引数据库。

编辑推荐

版权说明