<<搜索引擎>>

图书基本信息

书名：<<搜索引擎>>

13位ISBN编号：9787030342584

10位ISBN编号：7030342585

出版时间：2012-5

出版时间：科学出版社

作者：李晓明、闫宏飞、王继民

页数：330

字数：472750

<<搜索引擎>>

内容概要

搜索引擎：原理、技术与系统（第二版）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。

搜索引擎：原理、技术与系统（第二版）层次分明，由浅入深，上篇和中篇涉及内容提供了源代码下载地址；既有深入的理论分析，也有大量的实验数据和程序，具有学习和实用双重意义。

搜索引擎：原理、技术与系统（第二版）可作为高等院校计算机科学与技术、软件工程、信息管理与信息系统、电子商务等专业的研究生或高年级本科生的教学参考书和技术资料；对广大从事网络技术、Web站点管理、数字图书馆、Web挖掘等研究和应用开发的科技人员有很高的参考价值；书中提供了大量源代码，除了用于构建搜索引擎之外，对于学习编程，提高编程技巧，以及实现一个大规模应用开发也有一定的参考价值。

<<搜索引擎>>

书籍目录

章节摘录

版权页： 插图： 第二节 网页搜集 搜索引擎这样一个软件系统应该是何种工作方式？
如果说软件系统是工作在某个数据集合上的程序的话，这个软件系统操作的数据不仅包括内容不可预测的用户查询，还要包括在数量上动态变化的海量网页，并且这些网页不会主动送到系统来，而是需要由系统去抓取。
首先，我们考虑抓取的时机：事先还是即时。
我们都有经验，在网络比较畅通的情况下，从网上下载一篇网页大约需要1秒钟左右，因此如果在用户查询的时候即时去网上抓来成千上万的网页，一个个分析处理，和用户的查询匹配，不可能满足搜索引擎的响应时间要求。
不仅如此，这样做的系统效益也不高（会重复抓取太多的网页）；面对大量的用户查询，不可能想象每来一个查询，系统就到网上“搜索”一次。
因此我们看到，大规模搜索引擎服务的基础应该是一批预先搜集好的网页（直接或者间接）。
这一批网页如何维护？
可以有两种基本的考虑。
定期搜集，每次搜集替换上一次的内容，我们称之为“批量搜集”。
由于每次都是重新来一次，对于大规模搜索引擎来说，每次搜集的时间通常会花几周。
而由于这样做开销较大，通常两次搜集的间隔时间也不会很短（如早期天网的版本大约每3个月来一次，Google在一段时间曾是每隔28天来一次）。
这样做的好处是系统实现比较简单，主要缺点是“时新性”（freshness）不高，还有重复搜集所带来的额外带宽的消耗。
增量搜集，开始时搜集一批，往后只是：①搜集新出现的网页；②搜集那些在上次搜集后有过改变的网页；③发现自从上次搜集后已经不再存在了的网页，并从库中删除。
由于除新闻网站外，许多网页的内容变化并不是很经常的（有研究指出50%网页的平均生命周期大约为50天（Choetal.2000，Cho2002》，这样做每次搜集的网页量不会很大（例如，我们在2003年初估计中国每天有30万～50万变化了的网页），于是可以经常启动搜集过程（如每天）。
30万网页，一台PC机，在一般的网络条件下，半天也就搜集完了。
这样的系统表现出来的信息时新性就会比较高，主要缺点是系统实现比较复杂；这种复杂还不仅在于搜集过程，而是还在于下面要谈到的建索引的过程。
上面讲的是系统网页数据库维护的基本策略。
在这两种极端的情况之间也可能有一些折中的方案，J.Cho博士在这方面做过深入的研究（Choetal.2000，Cho2002），他根据一种网页变化模型和系统所含内容时新性的定义，提出了相应优化的网页搜集策略。
其中一个有趣的结论是：在系统搜集能力一定的情况下，若有两类网页（如“商业”和“教育”），它们的更新周期差别很大（如“商业”类网页平均更新周期是“天”，而“教育”类网页平均更新周期是“月”），则系统应该将注意力放在更新慢的网页上（Choetal.2000），以使系统整体的时新性达到比较高的取值。
在具体搜集过程中，如何抓取一篇篇的网页，也可以有不同的考虑。
最常见的一种是所谓“爬取”：将Web上的网页集合看成是一个有向图，搜集过程从给定起始URL集合S（或者说“种子”）开始，沿着网页中的链接，按照先深、先宽或者某种别的策略遍历，不停地从S中移除URL，下载相应的网页，解析出网页中的超链接URL，看是否已经被访问过，将未访问过的那些URL加入集合S。
整个过程可以形象地想象为一个蜘蛛（spider）在蜘蛛网（Web）上爬行（crawl）。
后面我们会看到，真正的系统其实是多个“蜘蛛”同时在爬。

<<搜索引擎>>

编辑推荐

《搜索引擎：原理、技术与系统（第2版）》保留了第一版上篇的大部分内容，即搜索引擎的基本原理，过去这么些年并没有什么变化；删除了第一版中的第九，第十二和十三章，增加了第十，第十一和十三章，分别介绍基于搜索引擎技术开发并从2002年一直运行至今的“中国web信息博物馆”、“中国数字财富库藏”及开放域问答系统。
同时，较大幅度修订了第一版中的部分小节内容。
《搜索引擎：原理、技术与系统（第2版）》分三篇共13章，内容包括引论、Web搜索引擎工作原理和体系结构、web信息的搜集、对搜集信息的预处理、信息查询服务等。

版权说明