学者书屋系列

搜索引擎技术

赵 杰◎著

哈尔滨工程大学出版社
Harbin Engineering University Press

学者书屋系列

搜索引擎技术

赵 杰 著

哈尔滨工程大学出版社

内容简介

本书比较系统地介绍了互联网搜索引擎的工作原理、信息预处理和查询过程及其用到的关键技术。全书共分7章，从基本工作原理概述开始，到一个小型简单专题搜索引擎实现的具体细节，进而详细讨论了歧义字段自动识别技术和命名实体自动识别技术；最后基于Agent与Multi-Agent技术，阐述了基于Agent的个性化信息检索系统的实现过程。本书层次分明，由浅入深；既有深入的理论分析，也有大量的实验数据，具有学习和实用双重意义。

本书可作为高等院校计算机科学与技术、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站点的管理、数字图书馆、Web挖掘等研究和应用开发的科技人员也有很大的参考价值。

图书在版编目(CIP)数据

搜索引擎技术/赵杰著.—哈尔滨:哈尔滨工程大学出版社,2007.11
ISBN 978-7-81133-125-7

Ⅰ.搜… Ⅱ.赵… Ⅲ.互联网络-情报检索-高等学校-教学参考资料 Ⅳ.G354.4

中国版本图书馆CIP数据核字(2007)第172793号

出版发行 哈尔滨工程大学出版社
社　　址 哈尔滨市南岗区东大直街124号
邮政编码 150001
发行电话 0451-82519328
传　　真 0451-82519699
经　　销 新华书店
印　　刷 黑龙江省地质测绘印制中心印刷厂
开　　本 787mm×960mm 1/16
印　　张 11.75
字　　数 150千字
版　　次 2007年11月第1版
印　　次 2007年11月第1次印刷
定　　价 25.00元

http://press.hrbeu.edu.cn
E-mail:heupress@hrbeu.edu.cn

前　言

Internet 自诞生以来不断发展，其内容不断丰富，整个网络逐渐堆积成一个前所未有的超大型信息库。Internet 作为一个信息平台在人们的日常生活和工作中发挥着越来越重要的作用，人们越来越多地通过 Internet 获取信息。在互联网发展初期，网站相对较少，网页数量亦较少，因而信息查找比较容易。然而伴随互联网爆炸性的发展，普通网络用户想找到所需的资料简直如同大海捞针，以至于迷失在信息的海洋中不知所措，出现了我们所说的“信息丰富，知识贫乏”的奇怪现象。搜索引擎正是为了解决这个“迷航”问题而出现的技术。

搜索引擎(Search Engine 简称 SE)是一个信息处理系统，它以一定的策略在互联网中搜集、发现信息，对信息进行理解、提取、组织和处理，并为用户提供检索服务，从而达到信息导航的目的，一般包括信息搜集、信息整理和用户查询三部分。从用户的角度来看，它就是一个帮助人们进行信息检索的工具。

本书介绍了搜索引擎的工作原理和主要实现技术。全书内容共分 7 章。

第 1 章对搜索引擎进行了简单介绍，主要介绍了搜索引擎的概念、分类及发展趋势等。

第 2 章 介绍 Web 搜索引擎的工作原理，主要介绍了搜索引擎的基本要求、爬虫、预处理、信息查询服务等相关内容。

第 3 章详细分析了数据预处理的相关技术，主要包括文本信息提取、去噪、分词、特征提取、文档表示、降维等方法。

第 4 章讲述了 Web 信息查询系统，重点介绍了查询方式和文档摘要。

第 5 章详细介绍了中文分词技术，主要包括自动分词系统的理论模

型 CWSM 和自动分词系统的评价准则，分析了歧义产生的根源，给出了歧义的三种分类。分析了中文识别的难点并对中文的识别技术进行了探讨，讨论了利用推理机制实现中国地名的自动识别方法。

第 6 章主要讨论了面向专题的信息搜集和处理，介绍了专题搜索引擎的构建、专题搜索引擎的自动文本分类和分词技术、基于向量空间模型的文本聚类等内容。

第 7 章讲述了基于 Agent 的智能信息检索技术，主要包括 Agent 技术、Agent 技术在信息检索中的应用、基于 Agent 的个性化信息检索系统等内容。

本书结合搜索引擎的主要技术和发展趋势而编写，力求重点突出，通俗易懂。

本书由牡丹江师范学院学科建设专项经费资助出版。

参加本书编撰工作的人员有：杨柳、柴宝杰、杨治秋等，在此对他们的辛勤工作和大力支持表示感谢。

由于作者水平有限，时间仓促，加之搜索引擎技术日新月异，书中存在的错误和不当之处，敬请读者指正。

作者

2007 年 8 月

目　录

第1章 绪 论

信息的生产、传播、搜集与查询是人类最基本的活动之一。考虑以文字为载体的信息，传统上有图书馆相应的编目体系和专业人员帮助我们很快找到所需的信息，其粒度通常是“书”或者“文章”。随着计算机与信息技术的发展，产生了信息检索学科领域，有了关于图书或者文献的全文检索系统，使我们能很方便地在“关键词“的粒度上得到相关的信息。

我们注意到，上述全文检索系统一般工作在一个规模相对有限、内容相对稳定的馆藏上，被检索的对象通常是经过认真筛选和预先处理的（例如人工提取出了“作者”，“标题”等元数据，形成了很好的“摘要”等），并且系统需要同时响应的查询数量通常都不会太大（例如每秒钟 10 个左右）。

1994 年左右，万维网（World Wide Web，简记为 WWW 或 Web）出现。它的开放性和其上信息广泛的可访问性极大地鼓励了人们创作的积极性。作为一个信息源，Web 和上述全文检索系统的工作对象相比，具有许多不同的特征，它们给信息检索领域带来了新的发展机遇和技术挑战。

首先，规模大。在短短的 10 年左右时间内，人类至少生产了 40 亿网页，而人类有文字上万年以来产生了大约 1 亿本书，而到 2004 年初中国网上大约有 3 亿网页，而中华民族有史以来出版的书籍大约不到 275 万种。尽管书籍的容量和质量是一般网页不可比的，但在对应的时间背景上考察其文字的总体数量，我们不能不为人类在 Web 上创造文字的激情而惊叹！

其次，内容不稳定。除了不断有新的网页出现外，旧的网页会因为各种原因被删除（有研究指出 50% 网页的平均生命周期大约为 50 天）。从原则上讲，读者数和作者数在同一个量级，形式和内容的随意性很强，

权威性相对不高,也不太可能进行人工筛选和预处理。

第三,与生俱来的数字化、网络化。传统载体上的信息,人们目前正忙于将它们数字化、网络化。这个特性是一把双刃剑:一方面便于我们搜集和处理,另一方面也会使我们感到太多,蜂拥而至,鱼目混珠。而作为要在 Web 上提供服务的信息查询系统,如搜索引擎和数字图书馆,通常要具备同时对付大量访问的能力(例如每秒钟 1 000 个查询),而且响应时间还要足够的快(例如 1 秒钟)。

1.1 搜索引擎的概念

搜索引擎,英文名称 Search Engine,一般是指通过超文本(超媒体)技术在 Internet 网络上建立的一种向网络用户提供网上信息资源检索和导航服务的专门站点或服务器。它通过搜集网上的信息,如网站,网页,URL 以及非 WWW 形态的 BBS,Telnet,FTP,Netsgroup 等,进行整理,组织,加工,处理,建立管理和存储这些信息的索引数据库,并提供基于该索引数据库的检索。用户输入关键词(Keywords)查询后,全文检索到的结果是与输入关键词相关的一个个网页的地址和一小段该网页内容的摘要(Abstract)。这些网页中应包含所输入的关键词或者相关的词汇。大多数搜索引擎支持最常见的关键词查询,并且检索功能强大。一般可以进行布尔逻辑检索,词组检索,位置检索,截词检索,检索词出现在特定位置检索等。

1.2 搜索引擎的分类

1.2.1 按数据检索方法分类

文档的索引与检索模型是搜索引擎的核心,检索模型的优劣直接影

响到搜索引擎的搜索效果。按文本信息检索模型搜索引擎可以分为:全文检索搜索引擎和目录分类式检索搜索引擎。

(1)全文检索搜索引擎 用户可以对各网站每个主页中的每个次页进行搜索,其查询全面而充分,但是由于信息太多反而会降低此种搜索引擎的命中率。此外,由于没有分类式搜索引擎那样清晰的层次结构,有时会给人一种繁多而杂乱的感觉,而且提供的查询结果重复链接较多。全文检索的关键是如何将原文档中所有基本元素的信息以适当形式记录到索引库中。在中文文档中,"基本元素"可以是汉字单字或词。

(2)目录分类式检索搜索引擎 目录分类式检索方法基于目录式分类结构(Directory)。目录分类式搜索引擎访问到新网站时首先将网站归于到某个分类下,再记录一些摘要信息(Abstract)对该网站进行概述性的简要介绍,故此类搜索引擎对用户提出的搜索要求只能在网站简介中查找。该类搜索引擎符合人们传统的信息查找方式,尤其适合于那些"希望了解某一方面、范围的信息,并不严格限于查询关键字"的用户。但其搜索范围与全文搜索引擎相比小得多,尤其是当用户选择类型不当时可能遗漏重要的信息源。最具代表性的目录式分类搜索引擎是 Yahoo。

1.2.2 按主要技术分类

从宏观上看,目前国内外各搜索引擎根据采用的主要技术不同可以分为以下四类。

(1)目录式(Directroy)搜索引擎 目录式搜索引擎(Yahoo 等)通过人工读取文档,以某种分类形式,如按学科,按字母顺序,按时间先后或这些方法的组合,组织 WWW 信息资源。其优点是结构清晰,便于人们浏览,其不足之处在于需要大量人力来搜集、组织信息,需要人工维护,且包含的内容不全,分类方式没有统一标准。

(2)基于网络爬虫(Crawler-based)的搜索引擎 基于网络爬虫的搜索引擎(Google, Altavista, WebCrawler, Lycos 等)又称索引式搜索引擎,是

一种最常见的搜索引擎。它通常包含三部分:查询接口,索引数据库以及网络爬虫。

爬虫首先会从一些初始的已知的 URL 开始,依次在 Web 上抓取这些网页并对抓取下来的网页进行分析,检查获取这些网页中的其他链接并将链接加入待爬 URL 队列中等待抓取,然后,索引器分析网页内容并将相应信息存入本地索引数据库中。索引数据库中如何存放组织数据往往要根据不同的分析结果和要求,针对索引和查询目的而进行设计。可以采用专用的也可以采用通用的数据库。查询接口通过索引数据库为用户的查询请求提供服务。

网络搜索策略及数据检索方法是基于网络爬虫搜索引擎的两个最重要的因素。Web 的搜索问题和经典的人工智能(Artificial Intelligence, AI)搜索图之间存在明显的映射关系,Web 中的文档相当于图中的结点,而到其他文档的超文本链接(Hypertext Links)相当于搜索图的边。数据检索方法通常有基于内容的全文(Full - text)检索和基于标题(Title - based)检索两类。前者的索引数据库往往较大。

(3)元搜索引擎(Meta-Search Engine) 元搜索引擎(Metacrawler, dopile, ixquick 等)的基本思想是,当搜索引擎收到来自不同用户的查询要求后,同时查询其他多个搜索引擎。目前网上有很多的搜索引擎,不同的搜索引擎往往有各自不同的用户查询接口和覆盖 WWW 不同部分的数据库,用户要得到更多的信息往往要多次使用不同的搜索引擎。所以,使用元搜索引擎,用户只需通过一次查询,就能得到相对更完整的信息。

(4)分布式(Distributed)搜索引擎 分布式搜索引擎根据地域、主题或其他的划分标准建立分布的检索服务器,检索服务器相互之间可以交换中间信息,且查询可以被重新定向,即如果一个检索服务器没有满足查询请求的信息,它可以将查询请求发送到具有相应信息的检索服务器上继续查询。

1.3 搜索引擎的发展现状

经过了多年的发展之后，现在的搜索引擎功能越来越强大，提供的服务也越来越全面，总的来说现在的搜索引擎主要有以下几种情况。

(1)目录型和检索型的搜索引擎相互结合 由于目录型和检索型的搜索引擎有各自的优点和缺点，目前它们谁也无法完全取代谁，于是很多搜索站点都同时提供这两种类型的服务。例如 Yahoo 是目录型搜索引擎的代表，但同时它也提供基于关键词的检索服务；而 Infoseek 则主要是一个检索型的搜索引擎，但它同时也建立了一个由人工编辑的小型目录。

(2)多样化和个性化的服务 现在绝大多数搜索引擎都提供多样化的服务，以吸引更多的用户，商业搜索引擎尤其注重这一点。以 Yahoo 为例，用户可以从它的首页上查看新闻、金融证券信息、天气预报、浏览黄页，可以进行网上购物、拍卖、找人，或者使用免费 E - Mail 和网上寻呼等服务。近期许多搜索引擎已开始提供个性化的服务，例如 Yahoo 的“My Yahoo”，Infoseek 的“Personalized start page”，Lycos 的“My Lycos”等，它们允许用户为自己定制起始页面，并选择感兴趣的内容和经常使用的服务放在该页面上。

(3)强大的查询功能 与最早的搜索引擎相比，现在的搜索引擎在查询功能方面已有了很大的改进。除了简单的 AND，OR 和 NOT 逻辑外，不少搜索引擎还支持相似查询，例如 AltaVista，Northern light，Lycos 等支持短语查询，AltaVista 的高级搜索功能支持 NEAR 逻辑等。域搜索也是一项很实用的功能，它允许用户把查询范围限制在网页的某个域中，例如标题、URL、图像标记或链接等，AltaVista，Northernlight，Infoseek 和我国的百度等搜索引擎都支持对网页的不同域进行搜索。

但是与搜索引擎快速发展的同时，其自身也存在着一些问题。

(a)提供的查询方式相当有限,与用户的交互性差,信息检索质量不高。

(b)仅支持单个关键词或者一组关键词及其逻辑运算符组成的查询,而并不支持自然语言搜索或语义搜索。

(c)不能利用历史信息进行搜索。用户的每次搜索都是从头开始,而不是从原有的查询结果中作进一步选择。

(d)呈现方式单一、呆板。多数搜索引擎只返回一个长长的搜索结果列表,其中可能数以万计的包含关键词的网页,但这些网页是否以及在多大程度上与用户的搜索意图相关,则不得而知。

1.4 搜索引擎的发展趋势

搜索引擎已成为一个新的研究、开发领域。因为它要用到信息检索、人工智能、计算机网络、分布式处理、数据库、数据挖掘、数字图书馆、自然语言处理等多领域的理论和技术,所以具有综合性和挑战性。又由于搜索引擎有大量的用户,有很好的经济价值,所以引起了世界各国计算机科学界和信息产业界的高度关注,目前的研究、开发十分活跃,并出现了很多值得注意的动向。总的看来,搜索引擎技术的未来发展趋势将主要体现在以下几个方面。

(1)注意提高信息查询结果的精度,提高检索的有效性　用户在搜索引擎上进行信息查询时,并不十分关注返回结果的多少,而是看结果是否和自己的需求吻合。对于一个查询,传统的搜索引擎动辄返回几十万、几百万篇文档,用户不得不在结果中筛选。解决查询结果过多的现象目前出现了几种方法:一是通过各种方法获得用户没有在查询语句中表达出来的真正用途,包括使用智能代理跟踪用户检索行为,分析用户模型;使用相关度反馈机制,使用户告诉搜索引擎哪些文档和自己的需

求相关(及其相关的程度),哪些不相关,通过多次交互逐步求精。二是用正文分类(Text Categorization)技术将结果分类,使用可视化技术显示分类结构,用户可以只浏览自己感兴趣的类别。三是进行站点类聚或内容类聚,减少信息的总量。

(2)基于智能代理的信息过滤和个性化服务 信息智能代理是另外一种利用互联网信息的机制。它使用自动获得的领域模型(如 Web 知识、信息处理、与用户兴趣相关的信息资源、领域组织结构)、用户模型(如用户背景、兴趣、行为、风格)知识,进行信息搜集、索引、过滤(包括兴趣过滤和不良信息过滤),并自动地将用户感兴趣的、对用户有用的信息提交给用户。智能代理具有不断学习、适应信息和用户兴趣动态变化的能力,从而提供个性化的服务。智能代理可以在用户端进行,也可以在服务器端运行。

(3)关联式的综合搜索 以往的搜索经验表明,很多人都遇到过在甲网站找图片,到乙网站找新闻再到丙网站找股票资讯的情况,这十分麻烦且浪费时间。那为何不考虑将这些图片、新闻、股票等等各种有关联的信息整合在同一界面,让互联网用户一次查询,全部满足呢?所谓关联式综合搜索,就是一种一站式的综合搜索服务,它使得互联网用户在搜索时只需输入一次查询目标,即可在同一界面得到各种有关联的查询结果。

(4)用分布式体系结构提高系统规模和性能 搜索引擎的实现可以采用集中式体系结构和分布式体系结构,两种方法各有千秋。但当系统规模到达一定程度(如网页数达到亿级)时,必然要采用某种分布式方法,以提高系统性能。搜索引擎的各个组成部分,除了用户接口之外,都可以进行分布:搜索器可以在多台机器上相互合作、相互分工进行信息发现,以提高信息发现和更新速度;索引器可以将索引分布在不同的机器上,以减小索引对机器的要求;检索器可以在不同的机器上进行文档的并行检索,以提高检索的速度和性能。

(5)重视交叉语言检索的研究和开发　交叉语言信息检索是指用户用母语提交查询,搜索引擎在多种语言的数据库中进行信息检索,返回能够回答用户问题的所有语言的文档。如果再加上机器翻译,则返回结果可以用母语显示。该技术目前还处于初步研究阶段,主要的困难在于语言之间在表达方式和语义对应上的不确定性。但对于经济全球化、互联网跨越国界的今天,无疑具有很重要的意义。

(6)本土化的搜索　世界上许多著名的搜索引擎都在美国,他们以英语为基础,完全按他们的思维方式和观点搜集和检索资料,这对于全球不同国家的用户来说显然是不适合的。各国的文化传统、思维方式和生活习惯不同,在对网站内容的搜索要求上也就存在差异。搜索结果要符合当地用户的要求,搜索引擎就必须本土化。我国的百度搜索引擎就是这方面的倡导者。

第 2 章　Web 搜索引擎的工作原理

2.1　搜索引擎的基本要求

搜索引擎是一个网络应用软件系统，如图 2－1 所示，对它有如下基本要求。

能够接受用户通过浏览器提交的查询词或者短语，记作 Q。在一个可以接受的时间内返回一个和该用户查询匹配的网页信息列表，记作 L。这个列表的每一条目至少包含三个元素（标题，网址链接，摘要）。

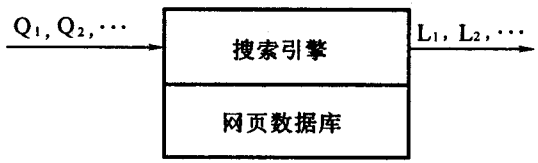

图 2－1　搜索引擎示意图

“可以接受的时间”，也就是响应时间。对于在 Web 上面的软件来说，这个时间不能太长，通常也就在“秒”这个量级。这是衡量搜索引擎可用性的一个基本指标，也是和传统信息检索系统的一个差别。更进一步的，这样的响应时间要求不仅要能满足单个用户查询，而且要能在系统设计负载的情况下满足所有的用户。也就是说，系统应该在额定吞吐率的情况下保证秒级响应时间。“匹配”，指的是网页中以某种形式包含有 Q 的内容，其中最简单、常见的形式就是 Q 在其中直接出现。但如果

一个搜索引擎就是以百分之百满足这种简单包含关系为目标，即使实现了也并不就达到了最好的效果。

“列表”，蕴含着一种“顺序”。绝大多数情况下，L 是相当长的，例如超过 1 万个条目(这是和图书馆全文检索系统的又一个不同，那里返回的列表通常较短，例如几十个条目)。这不仅是由于 Web 的信息量大，也由于搜索引擎的查询方式简单。简单，意味着抽象；抽象，意味着有更多的具体事物可能是它的体现。对于一个长长的列表，很少有用户有耐心都审视一遍(不仅是因为长，还因为大多数使用搜索引擎的用户通常都是“找到为止”，而不是“不全部找到不罢休”，加上这个列表中和一个用户关心的其实只占很少的比例)。有分析统计表明，用户平均察看返回结果不超过 2 页。现代大规模高质量搜索引擎一般采用三段式的工作流程：网页搜集、预处理和查询服务。如图 2－2 所示。

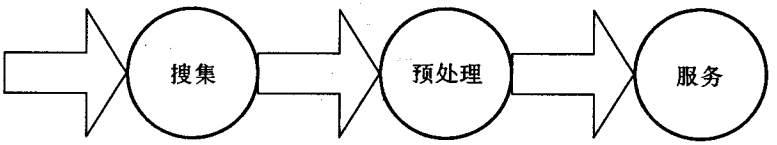

图 2－2　搜索引擎工作流程图

如果说软件系统是工作在数据集合上的程序的话，那么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。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