介绍业界热门的Lucene.Net、使用WebBrowser做爬虫
以及结合Solr开发ASP.NET搜索的第一本图书

DVD 附超值光盘

使用C#开发搜索引擎

罗刚 编著

Search Option 1

清华大学出版社

介绍业界热门的Lucene.Net、使用WebBrowser做爬虫
以及结合Solr开发ASP.NET搜索的第一本图书

DVD 附超值光盘

使用C#开发搜索引擎

罗刚 编著

清华大学出版社
北 京

内 容 简 介

从 C#基础开始，逐渐深入，是学习搜索引擎开发的首选。应众多公司的实际需求，本书介绍如何以 C#作为工具开发搜索引擎。全书以完成一个网站搜索\垂直搜索作为目标，从网络爬虫抓取数据开始，然后到中文分词、文本排重等文本挖掘技术和搜索结果展现。本书是市面上介绍业界热门的 Lucene.Net、使用 WebBrowser 做爬虫以及结合 Solr 开发 ASP.NET 搜索的第一书。

本书适合专业软件开发人员，也适合于希望学习搜索引擎工作原理的读者学习使用。本书对于在校学生学习复杂数据结构和应用动态规划等常用算法也有参考价值。

图书在版编目（CIP）数据

使用 C#开发搜索引擎 / 罗刚编著. —北京：清华大学出版社，2012.2
ISBN 978-7-302-27070-6

Ⅰ. ①使…　Ⅱ. ①罗…　Ⅲ. ①C 语言－程序设计　Ⅳ. ①TP312

中国版本图书馆 CIP 数据核字（2011）第 206926 号

责任编辑： 夏兆彦
责任校对： 徐俊伟
责任印制： 何　芊

出版发行： 清华大学出版社
网　　址： http://www.tup.com.cn, http://www.wqbook.com
地　　址： 北京清华大学学研大厦 A 座　　**邮　　编：** 100084
社 总 机： 010-62770175　　**邮　　购：** 010-62786544
投稿与读者服务： 010-62795954，jsjjc@tup.tsinghua.edu.cn
质量反馈： 010-62772015，zhiliang@tup.tsinghua.edu.cn
印 刷 者： 清华大学印刷厂
装 订 者： 三河市金元印装有限公司
经　　销： 全国新华书店
开　　本： 190mm×260mm　　**印　张：** 22.25　　**字　数：** 556 千字
（附光盘 1 张）
版　　次： 2012 年 2 月第 1 版　　**印　次：** 2012 年 2 月第 1 次印刷
印　　数： 1～4000
定　　价： 49.00 元

产品编号：035009-01

FOREWORD

前言

如果有计算机或者手机，在寻找解决问题的方法时，我们往往用搜索引擎寻找答案。在出门之前，我们往往先用搜索引擎查找乘车路线。在购物之前，往往先用搜索引擎找到最低的价格。搜索引擎已经成为大脑的外部记忆体。而宅男腐女们也更加依赖搜索引擎等计算机工具。

但这些也许仍然不够，也许你觉得语音输入搜索识别率准确率仍然有待提高。也许你希望能查找英文资料并能够自动翻译成为可读的中文资料。也许你希望能查找图像中的文字。很多网站搜索的表现也经常让人失望。大型搜索引擎虽然功能强大，但是为什么不能把搜索结果按时间排序？搜索引擎还有很多需要改进的地方，所以出现了越来越多的搜索引擎相关的开发岗位。这些岗位相对来说，一般都是高薪。

这本书不仅仅适合专业软件开发人员，希望学习搜索引擎工作原理的人也可以阅读。有很多人说，不是我不愿意，是我学不会技术。也许你需要一本更好的书，或者一位更好的老师。根据生活中的场景明白折半查找其实就是猜价格的游戏。有限状态机其实就是电话银行中缴费之前让人晕的一串数字输入。总之，学习技术，不仅仅很有用，还有可能比较有趣。

介绍搜索引擎原理及具体开发实现的书已经有几本，包括笔者出版的几本。但大部分是使用 Java 实现。C#在国内很流行，但却没有使用 C#开发搜索引擎的书籍。这本书就是为了填补这个空白。虽然使用.Net 总有一种寄人篱下的感觉，但是毕竟还有 Mono 这样的开源替代品。甚至已经有基于 Mono 开发的桌面搜索软件 Beagle。

这本书从需要用到的基本 C#语法开始介绍，然后再介绍如何使用它开发搜索引擎应用。C#语法已经越来越复杂，但这里只选取最需要的一部分。学以致用是这本书的写作原则。

Lucene 几乎已经成为全文搜索的同义词。随着.Net 开发平台越来越强大，作为 Lucene 在.Net 平台的移植版本，Lucene.Net 也越来越流行。Lucene.Net 来源于 2002 年的 NLucene，当时采用.Net 平台的首选语言 C#移植 Lucene，这个决定到现在看来仍然是正确的。但后来几经波折，这个开源项目在 2004 年一度陷于停滞，但是现在也终于被阿帕奇基金会接纳成为孵化项目，有望修成正果。

可能有人会好奇 Lucene.Net 是怎么从 Lucene 移植过来的。可以使用 Sharpen 这样的移植工具，这样可以把 Java 源代码转化为 C#源代码。当然还需要人工修改 C#源代码中的错误。虽然已经有包括 Autodesk 等公司和项目采用了 Lucene.Net，但是本书却是第一本介绍 Lucene.Net 的书，即使在全球范围内来看，也是如此。虽然这件好事来得晚了一些，但是该发生的终于还是发生了。

Lucene.Net 因为上手快，速度快，可扩展性好，赢得了很多开发人员的青睐。很多以前使用 SQL Server 全文搜索的开发人员转而使用 Lucene.Net。虽然 Lucene.Net 一般运行于 Windows 服务器，但是把它部署在 Linux 机器亦无不可。

通过 Lucene.Net 源代码学习各种算法也是一种不错的选择。例如学习使用堆实现的优先队列等。随着 Lucene 4.0 中灵活索引的推出，Lucene.Net 将来的版本性能会更好。

这也是第一本介绍如何使用 C#开发中文分词和文本排序、拼写检查等自然语言处理技术的书。使用 C#灵活的语法来实现中文分词使得代码可读性更好。

这也是第一本介绍如何使用 C#开发网络爬虫的书，因为 C#能够方便地调用浏览器内核，所以很容易解析动态网页。把网页转换成 DOM 树的表示形式，在 C#中也是轻而易举。

如果担心 Lucene.Net 功能仍然不够，可以使用支持分布式索引的 Solr。因为 Solr 有 Web 管理界面，所以可以在安装 Solr 之前就登录到 Solr 的管理界面使用它。Solr 的.Net 客户端接口也是精心设计的。在某些已经采用.Net 作为开发前端的大型网站中，采用这样的站内搜索将是绝配。

本书配套的光盘中提供了相关的源代码，有的来源于猎兔搜索多年的开发经验积累，有的是经典算法实现。其中很多都可以直接用于项目实践。

这本书从选题到出版已经过去了两年多时间，如果没有这本书，也许至少还要再等几年，才能出现一本内容类似的书。对于笔者推出的第一本介绍搜索引擎开发的书，有的读者有相见恨晚的感觉。希望这本书也能带给 C 语言开发人员一些新的想法。

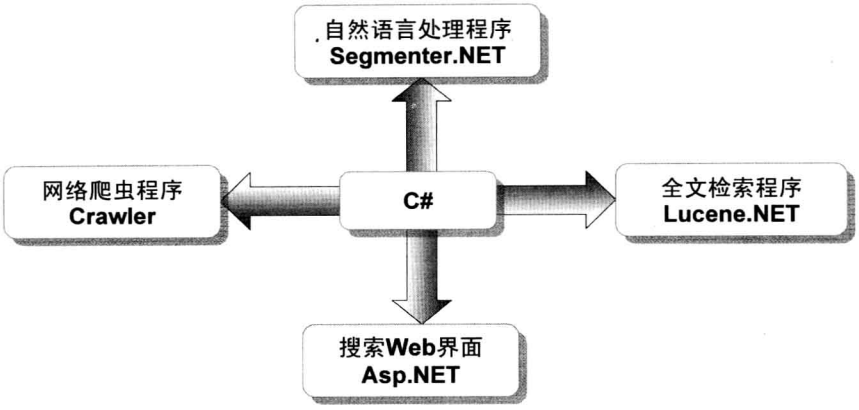
对于很多 C#开发人员来说，也许生活并不容易，因为很多 C#程序员收入相对较低。希望本书能让大家学习更轻松，工作更有成效，成为一本给力的软件开发类书籍。有问题可以直接和笔者交流，请发邮件到 luogang@gmail.com。最好有相关代码反馈。习惯使用 QQ 群的读者，可以加猎兔搜索 QQ 群：166015123。

编者

CONTENTS

目录

第1章 使用C#开发搜索引擎快速入门



搜索引擎经过最近几十年的快速发展，已经改变了人们的记忆方式。有研究表明，人们会忘记自己能在网络上找到的信息，而记住自己认为无法在网络上找到的信息。研究也发现，人们更容易记住在互联网的何处能找到这些信息，而不是记住信息内容本身。从某种意义上讲，由于有了搜索引擎，我们才可以把一些记忆任务交给机器来完成。

很多网站需要开发搜索功能。不仅如此，学会自己开发搜索引擎还将会为解决很多问题提供一种新方法。本书介绍使用流行的.NET（C#）编程语言开发搜索引擎。一件事情有更多人参与，就更容易做好。为了更好的协作，本章介绍的搜索引擎大部分采用开源软件实现。读者可以与猎兔搜索专业的技术开发人员一起改进相关实现。制作过程中所用的程序在所赠光盘中都能找到。

本章首先介绍搜索引擎的应用现状，然后介绍搜索引擎整体结构，并深入展开分析搜索的基本技术，最后复习下C#编程基础。

1.1 各种搜索引擎

搜索引擎有运行在大规模云计算的通用搜索引擎，也有一些行业搜索以及网站搜索。通用搜索引擎是大鄂，每一只都有自己独立的领地。行业搜索是领头雁，是各行业的旗帜。而网站搜索则像一只只小麻雀，麻雀虽小，五脏俱全。

1.1.1 通用搜索

目前通用搜索引擎的组织方式主要有网络综合搜索引擎和网络主题资源搜索引擎两种。其中网络综合搜索引擎能够广泛地采集各个互联网站点资源，并对其进行页面搜索，将索引结果存入索引数据库，供网络用户检索，提供互联网网络资源地导航功能的工具，如 Google、百度等。

这样的公司需要大量的服务器和专业开发人员，运营开销大。解决经济上可行性就是一个问题。通用搜索引擎的主要收入是在搜索结果页中展示与用户输入的关键词相关的广告。条幅广告更早出现。按点击付费的关键词广告比条幅广告的收费额度更低。点击一次广告可能只收几分钱，而条幅广告的计价单位至少在几百块（人民币）以上。那些曾经被忽视的中小企业，一度被认为是游离在广告市场之外的客户，突然成了时代的宠儿。地球上最大的动物鲸鱼吃的是小鱼小虾，只有这样才能摄入足够的食物。

通用搜索引擎企业是资本密集型企业，这样的公司往往前期有风险投资，有一定盈利后成为上市公司。

1.1.2 垂直搜索

垂直搜索是针对某一个行业的专业搜索引擎，例如搜房（http://www.soufun.com/），生活信息搜索（http://www.kooxoo.com），职位搜索（http://www.jobui.com），39 健康网上的搜索。垂直搜索是搜索引擎的细分和延伸，是对网页库中的某类专门的重要数据进行处理后，再对信息进行一次整合，定向分字段抽取出需种形式返回给用户。

垂直搜索需要从茫茫的互联网中获取行业信息，信息按行业过滤和分类是必不可少的。垂直搜索引擎和普通的网页搜索引擎的另一个最大区别是对网页信息进行了结构化信息抽取，也就是将网页的非结构化数据抽取成特定的结构化信息数据，比如网页搜索是以网页为最小单位，基于视觉的网页块分析是以网页块为最小单位，而垂直搜索是以结构化数据为最小单位。然后将这些数据存储到数据库，进行深一步的加工处理，如去重、分类等，最后分词、索引再以搜索的方式满足用户的需求。

整个过程中，数据由非结构化数据抽取成结构化数据，经过深度加工处理后以非结构化的方式和结构化的方式返回给用户。

垂直搜索引擎的应用方向很多，比如企业库搜索、供求信息搜索引擎、购物搜索、房产搜索、人才搜索、地图搜索、mp3 搜索、图片搜索……几乎各行各业各类信息都可以进一步细化成各类的垂直搜索引擎。

垂直搜索引擎大体上需要以下技术：

（1）定向的网络爬虫；

（2）网页结构化信息抽取技术或元数据采集技术；

（3）中文分词、全文检索；

（4）其他信息处理技术。

垂直搜索引擎的技术评估应从以下几点来判断：

（1）全面性：应该能从众多的来源采集信息。

（2）更新性：用户最好可以在几秒钟或几分钟内看到最新发布的信息。

（3）准确性：数据分类准确，不能包含重复冗余信息。

（4）功能性：功能完善，可以同时搜索文字信息，图片，视频，地理信息等。

垂直搜索的进入门槛很低，但是竞争的门槛很高。没有专注的精神和精湛的技术是不行的。行业门户网站具备行业优势但它们却没有技术优势，绝对不要想象着招几个人就可以搞定垂直搜索的全部技术。作为一个需要持续改进可运营的产品而不是一个项目，对技术的把握控制程度又是垂直搜索成功的重要因素之一。与专业的搜索技术提供商合作共赢是一种现实的解决方法。其中猎兔搜索是专业提供基于 Lucene 和自然语言处理商业支持的企业搜索公司。

1.1.3 站内搜索

站内搜索有三种流行的实现方式：

- **基于数据库的搜索** 比如 SQL Server 或者 MySQL 内部都有对全文检索列的支持。
- **基于爬虫抓取的站内搜索** Google 通过从外部抓取网页的方式提供免费的站内搜索。
- **站内搜索软件系统** 通过和数据库的同步利用 Lucene 建立独立的全文索引的站内搜索系统。

真正的全文检索应具备相关性排序技术和分词索引功能。分词、索引、排序这是全文检索的基本和核心，缺一不可。全文检索至少需要具备中文分词、索引、相关性排序功能。

所以简单考查一个站内搜索引擎的真伪只需要知道：能否实现相关性排序、国际标准的搜索语法、动态摘要、飘红、支持海量数据多并发快速查询、搜索耗时极短。

猎兔（http://www.lietu.com/）企业搜索的实现正是这样一种站内全文搜索的实现。

1.2 搜索引擎整体结构

一个最简单的搜索引擎由搜索和抓取两部分组成，完整的结构如图 1-1 所示。

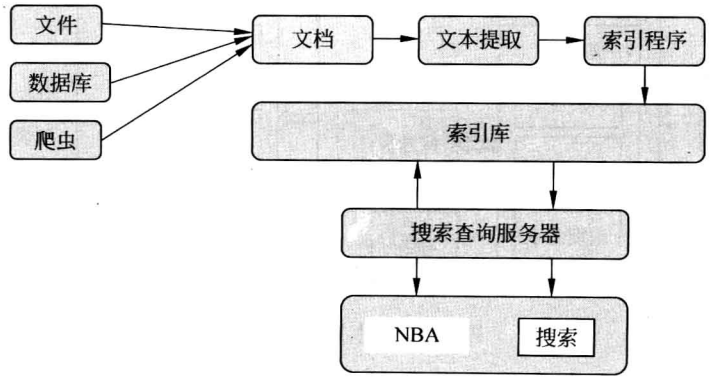


图 1-1 搜索引擎的基本结构

数据来源可以是 Web 或者数据库等，也可以是本地路径等。如果来源于互联网，就要用到爬虫。抓下来的文本在写入索引库时，需要切分成一个个的词。

1.3 搜索引擎基本技术

一个基本的搜索包括采集数据的爬虫和索引库的管理以及搜索页面展现等部分。本书的第 2 章到第 6 章将分别详细地介绍这些内容。

1.3.1 网络爬虫

网络爬虫（Crawler）的主要作用是获取互联网上的信息。网络爬虫利用主页中的超文本链接遍历 Web，通过 URL 引用从一个 HTML 文档爬行到另一个 HTML 文档。http://dmoz.org 是整个互联网抓取的入口。网络爬虫收集到的信息可有多种用途，如建立索引、HTML 文件的验证、URL 链接验证、获取更新信息、站点镜像等。网络爬虫建立的页面数据库，包含有根据页面内容生成的文摘，这是一个重要特色。

网站本身可以声明不希望被搜索引擎收入的内容。这可以有两种方式实现：第一种方式是在站点增加一个纯文本文件，例如http://www.lietu.com/robots.txt；另外一种方式是直接在 HTML 页面中使用 robots 的 meta 标签。在抓取网页时大部分网络机器人会遵循 Robot.txt 协议。

1.3.2 文本挖掘

搜索文本信息需要理解人类的自然语言。文本挖掘指从大量文本数据中抽取隐含的、未知的、可能有用的信息。

常用的文本挖掘方法包括：全文检索、中文分词、句法分析、文本分类、文本聚类、关键词提取、文本摘要、信息提取、智能问答等。文本挖掘相关技术的结构如图 1-2 所示。

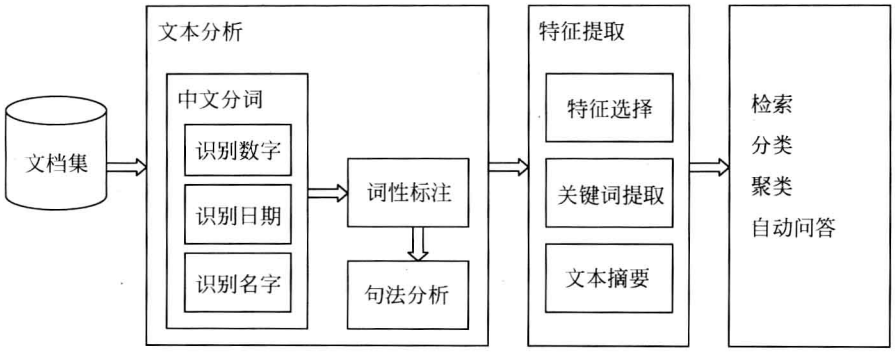


图 1-2 文本挖掘的结构

1.3.3 全文索引

早在计算机出现之前，为了方便查询，已经出现了人工为图书建立的索引，比如图 1-3 中的名词索引。

为了按词快速定位抓取过来的文档，需要以词为基础建立全文索引，也叫倒排索引（Inverted index），如图 1-4 所示。

倒排索引是相对于正向索引来说的，首先用正向索引来存储每个文档对应的单词列表，然后再建立倒排索引，根据单词来索引文档编号。

——按名词的拼音顺序检索

名词索引

半地下室 semi-basement

房间地面低于室外地平面的高度超过该房间净高的1/3，且不超过1/2者。
(摘自《住宅设计规范》（GB 50096-1999）3页 中国建筑工业出版社 1999.5第一版)

壁柜 cabinet

住宅套内与墙壁结合而成的落地贮藏空间。
(摘自《住宅设计规范》（GB 50096-1999）3页 中国建筑工业出版社 1999.5第一版)

比例 proportion

建筑构成各部分和各部分之间的相互关系，以及各部分与整体之间的比较关系。建筑比例是建筑构成中的一种量度尺度，有了具体的尺度才具有比例的真正意义。
(摘自《中国土木建筑百科辞典》（建筑卷）24页 中国建筑工业出版社 1995.5第一版)

变形缝 Deformation joint

为防止建筑物在外界因素作用下，结构内部产生附加变形和应力，导致开裂甚至破坏而预留的构造缝。变形缝包括
(摘自《中国土木建筑百科辞典》（建筑卷）27页 中国建筑工业出版社 1995.5第一版)

标准层 typical floor

平面布置相同的住宅楼层。
(摘自《住宅设计规范》（GB 50096-1999）2页 中国建筑工业出版社 1999.5第一版)

不对称均衡 asymmetrical balance

图 1-3　人工建立的名词索引

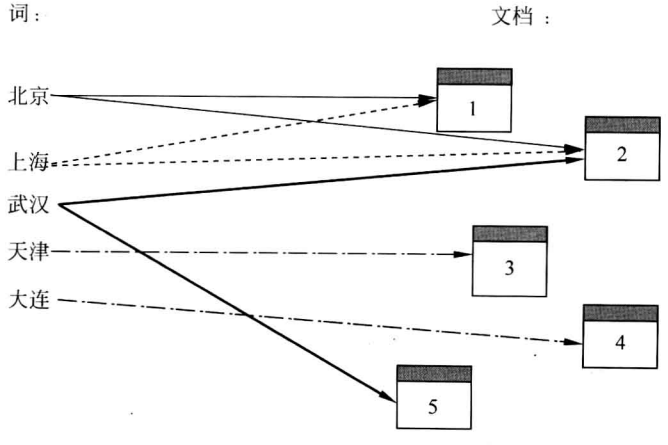


图 1-4　以词为基础的全文索引

例如要索引如下两个文档：

Doc Id 1：自己动手写搜索引擎

Doc Id 2：自己动手写网络爬虫

首先把这些文档中的内容分成一个个的单词：

Doc Id 1：自己/动手/写/搜索引擎

Doc Id 2：自己/动手/写/网络爬虫

按单词建立的倒排索引结构如表 1-1 所示。

表 1-1　Lucene.Net 中的倒排索引结构

词	（文档，频率）	在文档中出现的位置
动手	(1, 1), (2, 1)	(2), (2)
搜索引擎	(1, 1)	(4)
网络爬虫	(2, 1)	(4)
写	(1, 1), (2, 1)	(3), (3)
自己	(1, 1), (2, 1)	(1), (1)

每个单词（term）后面的文档编号（docId）列表叫做 posting list。在全文索引包 Lucene.Net 中，倒排索引结构存储在二进制格式的多个索引文件中，其中以 tis 为后缀的文件中包含了单词信息，frq 后缀的文件记录单词的文档编号和这个单词在文档中出现了多少次，也就是频率信息，prx 后缀的文件包含了单词出现的位置信息。

为了快速地查找单词，可以先对单词列表排序，例如，《新华字典》和《现代汉语词典》按拼音排序。查词典的时候要找一个词，可以把词典从头到尾翻看直到找到这个词为止。更快的方法是，直接翻看词典中间一页，看是否有这个词。如果比要找的词小，则往后找，如果比要找的词大，则往前找。这个方法叫做折半查找。从排好序的词表中查找一个词可以采用折半查找的方法快速查询。

“购物街”节目，里面猜价格那帮人真让人痛苦。“500”“高了”“400”“高了”“300”“高了”“200”“低了”“210”“低了”“220”“低了”……这个故事告诉我们学会折半查找多么重要。选择一个从 0～100 的数字，并且依赖你的猜测，我说：对了，高了或者低了。你选择从哪个开始猜呢？

50 当然!为什么?最好的情况是你猜中了。最差的情况是，你被告知 “高了”或者“低了”。现在思考下，如果你被告知“高了”，你就消除了 51～100。如果你被告知“低了”，你就消除了 0～49，换句话说，你消除了一半的可能性。

50　51　…　74　75　76　…　99　100

如果假设是“低了”， 接下来你猜什么呢? 75。因为它位于 50～100。如果你还没猜对，你又消除了一半的可能性!

一个上限值和下限值界定查询词所在的区间范围。下面是实现折半查找的代码。

```
//输入排好序的数组和待查询的值，返回在数组中的位置
static int BinarySearch(int[] array, int value){
    int low = 0;                        //查询的开始范围
    int high = array.Length - 1;        //查询的结束范围
    int midpoint = 0; //中间点
```

```
    while (low <= high)  {
        midpoint = (low + high) / 2;

        //检查是否和数组中的中间值相等
        if (value == array[midpoint])  {
            return midpoint;
        }
        else if (value < array[midpoint])
            high = midpoint - 1;
        else
            low = midpoint + 1;
    }

    //没有找到
    return -1;
}
```

1.3.4 搜索语法介绍

专业的搜索引擎一般都会实现一个搜索语法，基本的搜索语法有逻辑运算符：

- 与（+、空格）
- 或（OR、|）
- 非（-）

例如，搜“神雕侠侣”，希望是关于武侠小说方面的内容，却发现很多关于电视剧方面的网页。那么就可以这样查询：神雕侠侣 -电视剧

注意，前一个关键词与减号之间必须有空格，否则，减号会被当成连字符处理，而失去减号语法功能。减号和后一个关键词之间，有无空格均可。

除了逻辑运算相关的搜索语法，还有：

（1）把搜索范围限定在网页标题中——intitle

网页标题通常是对网页内容提纲挈领式的归纳。把查询内容范围限定在网页标题中，有时能获得良好的效果。使用的方式，是把查询内容中，特别关键的部分，用“intitle:”领起来。

例如，找小沈阳的小品，就可以这样查询：小品 intitle:小沈阳

注意，intitle:和后面的关键词之间，不要有空格。

（2）把搜索范围限定在特定站点中——site

有时候，如果知道某个站点中有自己需要找的东西，就可以把搜索范围限定在这个站点中，提高查询效率。使用的方式，是在查询内容的后面，加上“site:站点域名”。

例如，要从天空网下载软件查找 msn 聊天工具，就可以这样查询：msn site:skycn.com。

注意，“site:”后面跟的站点域名，不要带“http://”；另外，site:和站点名之间，不要带空格。

site 语法的另外一个用处是查看一个网站被搜索引擎收录的情况，例如通过 site:search.rayli.com.cn 可以看出 Google 中收录了 26 800 条瑞丽搜索的信息。

1.3.5 搜索用户界面

搜索用户界面一般由搜索首页，搜索结果显示页和高级搜索页面组成。图 1-5 是一个搜索首页，包含一个搜索输入提示的效果：

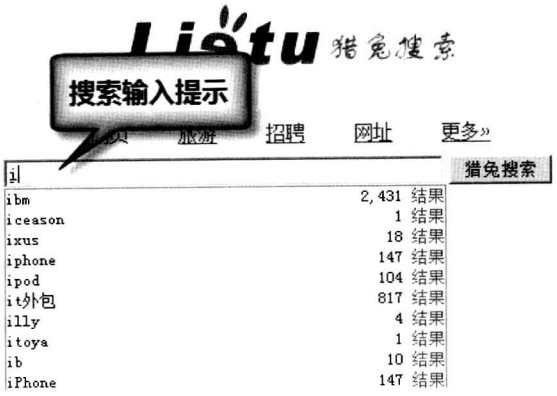


图 1-5 搜索首页

图 1-6 是一个搜索结果页面。

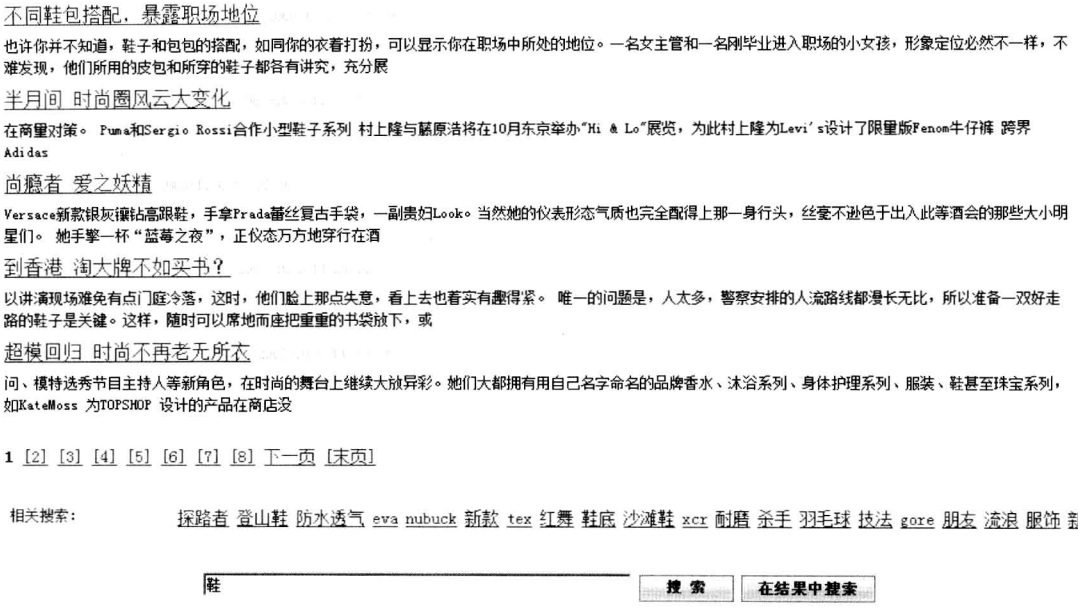


图 1-6 搜索结果页

在搜索结果页会对搜索词高亮显示，在搜索输入框下面会给出相关搜索词列表。如果用户搜索错误，在搜索输入框下面会给出可能的查询词提示“您是不是要找：XXX”，如图 1-7 所示。