搜索引擎的设计研究

王小林,刘宏申
(安徽工业大学 计算机学院,安徽 马鞍山 243002)

摘 要:搜索引擎是 Internet 信息服务的主体,搜索引擎的设计是各网站建设的重要部分。介绍了搜索引擎的分类和各类搜索引擎的工作过程。在此基础上,指出了蜘蛛程序是由网页下载和网页内容分析及信息提取两部分组成,并结合用 C++Builder 作为开发工具给出了这两部分的源代码示例。最后介绍了蜘蛛程序设计要注意的问题。

关键词:搜索引擎;蜘蛛程序;C++Builder
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2007)02-0005-03

Study on the Development of Search Engine

WANG Xiao-lin, LIU Hong-shen
(School of Computer Science, Anhui University of Technology, Maanshan 243002, China)

Abstract: Search engine is the main part of Internet information services, and design of search engine is important part of construction of Websites. At first, types and the working process of search engines are presented. It is presented that a spider (also called a "crawler" or a "bot") have two parts: one part goes to every page or representative pages on every Website that wants to be searchable and read it, using hypertext links on each page to discover and read a site's other pages. Another part finds keys of every page or representative pages on every Website. An example of spider programs is presented in C++Builder. Some problems of designing search engines are discussed.

Key words: search engine; spider program; C++Builder

0 前 言

互联网以非常惊人的速度在普及。它之所以快速发展,主要还是源于它能为人们提供所需要的信息。为人们在互联网上寻找信息的搜索引擎起着非常重要的作用。提起 google、百度等谁都知道,但它们是如何工作的、又是如何设计出来的并不是一般人所了解的。文中在介绍搜索引擎及其重要的组成部分——蜘蛛程序工作原理的基础上,用 C++Builder 作为开发工具介绍了蜘蛛程序的开发过程。最后对搜索引擎设计中的一些问题进行了研究讨论,这对致力于搜索引擎开发或相关研究的人们有一定的技术指导作用。

1 搜索引擎的工作原理

搜索引擎[1]一词在国内外因特网领域被广泛使用着,是为网页(网站)访问者提供查询信息服务。一个搜索引擎一般由网页信息数据库、用户信息查询程序和向网页信息数据库填加信息的程序组成,如图 1 所示。

不同类型的搜索引擎主要区别在于网页信息写入程序的工作方式不同,按其工作方式可将搜索引擎分为两种 标准的搜索引擎和目录索引。

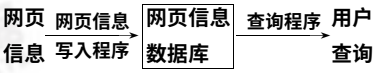

图 1 搜索引擎的组成

1.1 标准的搜索引擎

这是一种严格意义上的搜索引擎,它的网页信息写入是通过名为蜘蛛程序的程序在互联网上自动地提取各个网站的信息来建立自己的数据库的,因此是真正的搜索引擎。这里搜索引擎包括 ALTAVISTA, INKTOMI, INFOSEEK, GOOGLE 和百度等。

这类搜索引擎的网页信息搜集尽管是自动搜集,但操作方式还可以不同:一种是定期搜索,即搜索引擎定期主动派出"蜘蛛"程序,对一定 IP 地址范围内的网站进行检索,一旦发现更新或新的网站,它会自动提取网站的信息和网址加入自己的数据库。另一种是提交网站搜索,即由网站所有者主动向搜索引擎提交网址,然后搜索引擎在一定时间内专门向该网站派出"蜘蛛"程序,扫描并将有关信息存入数据库,以备用户查询。

收稿日期:2006-04-14
基金项目:安徽省教育厅自然科学基金项目(2006KJ018A)
作者简介:王小林(1964-),男,安徽安庆人,硕士,研究方向为数据库与人工智能。

1.2 目录索引

这类搜索引擎实际上不是严格意义上的搜索引擎,是将网站分门别类地存放在相应的目录中,即目录索引。这种目录索引不是由"蜘蛛"程序自动进行,而是完全依赖手工操作完成。用户(即网站所有者)提交网站后,目录索引编辑人员会亲自浏览该网站,然后根据一套自定的评判标准及编辑人员的主观印象,对其分门别类,将结果写入库中。

目录索引信息搜集不是自动进行,其效率低下。随着因特网的飞速发展,各类已有网站的内容在不断更新,而且新的网站层出不穷,因此要产生和维护一个信息充分详尽有效并能适时更新的信息库几乎是不可能的。鉴于此,文中主要介绍标准搜索引擎的设计问题。

2 蜘蛛程序的设计

作为搜索引擎的主要部分——蜘蛛程序的作用是从指定的网站或指定的超连接把网页内容下载到本机进行分析,提取该网页所感兴趣的内容。因此一个蜘蛛程序应该由两部分组成:网页内容的下载部分和网页内容的分析部分。文中用 C++Builder 作为开发工具介绍蜘蛛程序的这两部分的设计。

2.1 网页内容的下载

在现在的编程工具中,下载指定网页内容应该不是困难的问题,每个编程工具中都有相关的组件或控件,用户只要调用其中的方法就可下载指定网页的内容[2,3]。在 C++Builder 中,下载指定网页内容的组件是 NMHTTP[2],该组件有一个方法 Get 能将指定 URL(统一资源定位符)网页以 HTML 形式下载到本机,储存在该组件的 Body 属性中。设指定网页的 URL 存放在编辑框 Edit2 中,下面程序段完成网页内容的下载:

```
try
  {
  NMHTTP1->Get(Edit2->Text);
  }
  catch(...)
  {
  continue;
  }
.....
```

2.2 网页内容的分析和信息提取

网页内容的分析和信息提取是蜘蛛程序设计的难点,这直接关系到蜘蛛程序及搜索引擎的工作效率[3]。由上面分析可知,下载到本机的网页内容是 HTML 语言形式的网页内容,要从中提取出有用信息,需要用户对 HTML 语言(超文本标记语言)的语法有所了解,尤其要了解所感兴趣的信息点在 HTML 语言的出现形式。

根据信息点的作用不同,这里可把一页网页内容分成三部分:搜索引擎感兴趣的信息即搜索引擎要找的信息、超连接信息和其他信息。前两部分信息是搜索引擎要关心的信息,第二部分信息对搜索引擎之所以重要是因为搜索引擎要自动搜索信息必须能记录下在搜索过程中出现的新的网页的 URL 即超连接信息,这种信息在 HTML 中以特定标记出现,容易确定。对第一种信息在 HTML 中出现形式不确定,需要程序员有自己确定可能出现的形式,一般最可能含有感兴趣信息的地方是网页标题、超连接标题和网页正文等。下面程序段是搜索提取网页标题信息的程序段,程序先将下载的网页内容(在 NMHTTP 的 Body 中)放入字符串数组 HTM 中,然后在 HTM 中进行搜索:

```
htm->Clear();
htm->Add(NMHTTP1->Body);
for(i=0;i<htm->Count;i++)
    {
    s=htm->Strings[i];
    j=s.AnsiPos("<TITLE>");
    l=s.AnsiPos("<title>");
    if(j!=0)
     {
    s1=s.SubString(j+2+5,s.Length()-j);
    k=s1.AnsiPos("</TITLE>");
    s2=s1.SubString(0,k-1);
    }
.....
```

3 蜘蛛程序开发中一些需注意的问题

蜘蛛程序开发中有两个问题[1]决定着搜索引擎是否成功和实用:一是如何保证蜘蛛程序搜索时正常结束退出;另一个是对下载下来的网页内容分析提取需要的信息。

3.1 蜘蛛程序正常结束的问题

这个问题是由蜘蛛程序本身和网站的特点决定的。众所周知,网站的网页是通过各网页中的超连接连接起来的,如果把每个网页看作一个节点,把网页中的超连接看作节点间的联系的边的话,整个网站的网页和超连接构成的是图这种数据结构。对应网站的图有以下两个特点:第一是节点和边可能是数量巨大,像一个大学网站也有十几万个网页,网页间的联系即超连接杂乱无章,无顺序可言;第二是这个图没有明确的结束标志,因为一个网站中总有那么一些超连接是连

向站外的。蜘蛛程序搜索某个网站的网页实际上就是遍历与这个网站相对应的图。鉴于以上分析,对这种图的遍历要记录下 :已遍历的节点(即网页) ;下一步要遍历的网页,也就是已分析网页中包含的超连接。蜘蛛程序在网上搜索网页时间越久,分析的网页数便会增加,且产生的待分析网页数(即已分析网页中的超连接)也在增加,但后者增加的速度远远快于前者,如果前者以线性关系增加的话,后者就要以指数关系增加。这样就可能会出现待分析网页数增多到程序缓冲区存不下的状况。因此如何控制待分析网页数量,保证蜘蛛程序最后能完成搜索任务自动结束是蜘蛛程序设计时要考虑的一个难点问题。另外蜘蛛程序对一网站的网页分析可能要分几次完成,如何使得蜘蛛程序能从上次的断点再续也是一个现实问题。

3.2　网页有效信息的提取问题

如何有效地提取待分析网页的有效信息,是一个富有挑战性的问题。如上所述,将网页中的网页标题、超连接标题信息称为感兴趣的信息点,但是一个网页中的内容可能不仅包含这些文字信息,更多的是包含了多媒体信息,如图像、声音等信息,仅仅将网页中的上述信息(网页标题、超连接标题信息)提取出来是不全面的,机器如何提取和理解网页中的多媒体信息仍然是一个待研究的问题[1,4]。

在搜索引擎中经常碰到纯文档的网页,如何将该网页中的内容加以理解提取出若干个关键词,也是富有挑战性的问题。这实际上就是机器对纯文字信息理解,即自然语言理解问题,目前也是人工智能等研究的对象,还未有一个好的方法来解决[1]。

3.3　数据库内容查询速度问题

对于服务于各类网民的非专业搜索引擎如 google,时时面临着多个用户对庞大的数据信息快速搜索的请求问题 :一方面其后台数据库中的数据应该是海量的,另一方面网上同时对该搜索引擎提出搜索请求的用户可能很多,而且他们均希望服务器尽快返回他们需要的结果[5]。用一个或几个服务器来存储这些海量数据、完成这许多的用户请求且服务性能还不错是不现实的,因此用来支撑这些海量数据查询的服务器应该是几十个、几百个。为了得到较佳的服务性能,如何在这几十、几百个服务器间组织这些海量数据、平衡众多的服务请求也是一个复杂的问题,它涉及到操作系统、数据库系统中许多深层次的内容。

4　结　论

搜索引擎是一类服务软件,该软件包括前台部分和后台数据库两部分。一个性能优越、服务内容广且有效的搜索引擎的设计包含了许多高、尖、深的问题,如 :图像模式识别、声音模式识别、自然语言理解、海量数据存储访问技术等等,这些问题目前还有很多值得研究探讨的地方,因此搜索引擎的设计具有很大成分的研究性质。

参考文献 :

[1]　都云程,卢献华. 中文搜索引擎现状与展望[J]. 中文信息学报,1999,13(3) :61 - 64.

[2]　陈周造,陈灿煌. 精通 C + + Builder 5 程序设计高级教程[M]. 北京 :中国青年出版社,2001.

[3]　张晓辉,邵　华,常桂然. WWW 上的信息发现与搜索引擎技术[J]. 小型微型计算机系统,1998,19(6) :66 - 71.

[4]　桑梓勤,丁明跃,张天序. 环球网图象搜索引擎研究综述[J]. 中国图象图形学报,1998,3(6) :443 - 446.

[5]　任瑞娟,李洪建. 中文 WWW 搜索引擎比较研究[J]. 大学图书馆学报,1999(5) :55 - 61.

(上接第 4 页)

率较高。从建立关系模式到存储 XML 数据,整个过程自动完成,工作量大大减少。

参考文献 :

[1]　许卓明,刘　琴,董逸生. 基于关系数据库的 XML 存储技术评述[J]. 计算机工程与应用,2003(21) :197 - 200.

[2]　曾宇昆,王清明,杨卫东,等. XML 模式到关系范式的映射[J]. 计算机工程,2005(4) :37 - 39.

[3]　龚红炎,刘奕明,陈涵生. XML 与数据库结合技术的探讨[J]. 计算机工程,2005(2) :114 - 116.

[4]　严尚维,田绪红,孙爱东,等. 基于关系数据库的 XML 查询效率测试方法[J]. 计算机工程与应用,2004(2) :180 - 181.

[5]　方　翔. XML 文档到关系数据库的直接转换[J]. 计算机工程,2001(11) :65 - 66.

[6]　Deutsch A, Fernandez M, Suciu D. Storing Semistructured Data with STORED[C]// In : Proc of ACM SIGMOD Int' l Conf on Management of Data. Philadelphia, PA, USA : ACM Press, 1999 :431 - 442.

[7]　Shanmugasundaram J, Tufte K. Relational Databases for Querying XML Documents : Limitations and Opportunities [C]// In : Proc of 25th Int' l Conf on Very Large Data Bases. Edinburgh, Scotland, UK : Morgan Kaufmann Publishers, 1999 :302 - 314.

[8]　Tian F, Dabid J, Chen J. The Design and Performance Evaluation of Alternative XML Storage Strtegies[J]. ACM SIGMOD Record, 2002, 31(1) :5 - 10.