

基于语义分析的垂直搜索网络蜘蛛

罗立宏¹, 陈 志²

(1. 广东工业大学 数字媒体系, 广东 广州 510006; 2. 华南理工大学 广东省计算机网络重点实验室, 广东 广州 510640)

摘 要: 通用搜索引擎数据量庞大, 但查询结果不够准确。分类目录正好相反。为了综合两者优势, 对垂直搜索引擎进行了研究和分析。着重研究了垂直搜索引擎的核心模块——智能网络搜索蜘蛛。提出了搜索分析的新概念——规则。研究了蜘蛛中定义支持同义词的语义词典的方法, 给出了按照规则分析和检索的实现方法和流程。程序需要定义多种规则, 让蜘蛛依照规则进行网页爬行和信息采集。最后给出一个项目实例, 证明了上述方法的可行性。

关键词: 计算机应用; 互联网; 搜索引擎; 垂直搜索; 网络蜘蛛; 语义分析

中图分类号: TP391.9 **文献标识码:** A **文章编号:** 1000-7024 (2008) 18-4662-04

Search spider of vertical search engine based on semantic analysis

LUO Li-hong¹, CHEN Zhi²

(1. Department of Digital Media, Guangdong University of Technology, Guangzhou 510006, China;
2. Guangdong Key Laboratory of Computer Network, South China University of Technology, Guangzhou 510640, China)

Abstract: General search engine has large volume of data, but its search results are not accurate enough. Directories classification is on the contrary. In order to integrate advantages of the two, vertical search engine is studied and analyzed. The core module—intelligent search spider is mainly focused on. A new concept about searching and analyzing is brought forward: Rules. The method is researched that defining semantic dictionary which supports synonyms. The algorithm and flow that realize searching and analyzing according rules are afforded. Kinds of rules must be defined in search spider program, depending on which the function web pages crawling and information data extracting work. At last a project example is presented to prove the feasibility of these methods.

Key words: computer applications; internet; search engine; vertical search; web spider; semantic analysis

0 引 言

由于互联网的信息量越来越庞大, 于是“搜索”越来越成为互联网用户离不开的功能。获得网站网页资料, 能够建立数据库并提供查询的系统, 我们都可以把它叫做搜索系统。按照工作原理的不同, 可以把它们分为两个基本类别: 通用全文搜索引擎和分类目录。全文搜索引擎的数据库是依靠一个叫“网络蜘蛛”或叫“网络爬虫”的软件, 通过网络上的各种链接自动获取大量网页信息内容并分析整理形成的。Google、百度都是比较典型的通用全文搜索引擎系统。

分类目录则是通过人工的方式收集整理网站资料形成数据库的^[1]。比如雅虎中国以及国内的搜狐、新浪、网易分类目录。另外, 在网上的一些导航站点, 也可以归属为原始的分类目录, 比如“网址之家”。

通用搜索引擎和分类目录在使用上各有长短。通用搜索

引擎因为依靠软件进行, 所以数据库的容量非常庞大, 但是, 它的查询结果往往不够准确; 而且通用搜索引擎对动态实时信息不敏感, 例如读者可以尝试用百度搜索当天或前一两天发布的某城市的房屋出租信息, 如用关键词“房屋出租 北京 2007年7月6日(输今天或者前一两天的时间)”搜索。可以发现搜不到太多有价值的信息, 尽管互联网上每天发布出来的房屋出租信息有成百上千。分类目录则依靠人工收集和整理网站, 能够提供更为准确的查询结果, 也可以靠人力做到对实时信息敏感, 但收集的内容却非常有限。有没有能综合两者优势的解决方案呢? 垂直搜索引擎的概念就是在这样的环境中被提了出来^[2-3]。垂直搜索与 Google、百度等通用全文搜索不同, 它是针对某一个行业的专业全文搜索引擎。它是搜索引擎的细分和延伸, 是对网页库中的某类专门的信息进行整合和整理后再以某种形式返回给用户的搜索引擎。它是针对某一特定领域、某一特定人群或某一特定需求提供的有一定

收稿日期: 2007-10-08 E-mail: luo_lihong98@163.com

基金项目: 国家自然科学基金项目 (90412015)。

作者简介: 罗立宏 (1975—), 男, 广东怀集人, 硕士, 工程师, 研究方向为数字媒体、信息检索、中文信息处理; 陈志 (1970—), 男, 广东广州人, 博士, 副教授, 研究方向为信息检索、计算机网络。

价值的信息和相关服务。其特点就是“专、精、深”，且具有行业色彩。它是与通用搜索引擎截然不同的引擎类型。垂直搜索引擎专注具体、深入的纵向服务，致力于某一特定领域内信息的全面和内容的深入，这个领域外的闲杂信息不收录。比如设计一个搜索引擎专门搜索家教供求信息，并能够搜集通用搜索不敏感的实时动态信息，这就是一种垂直搜索。

1 垂直搜索引擎基本原理

垂直搜索的原理和结构与通用搜索类似，但又有不同。垂直搜索也有网络蜘蛛、数据信息库和检索界面等模块。总体结构如图1所示。

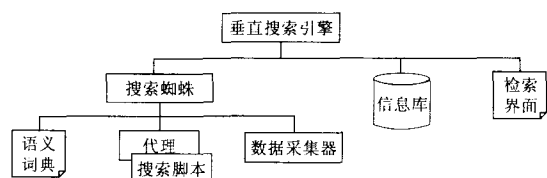


图1 垂直搜索引擎总体结构

其中蜘蛛负责寻找存在信息的网站，并深入网站内部寻找描述信息的页面。对于找到的信息描述页面，蜘蛛的数据采集器将对信息描述页面进行信息抽取。对于抽取出来的信息要检验合法性和有效性。通过检验的信息将保存进数据信息库。在蜘蛛的爬行过程以及页面分析过程中，都要对页面的全文进行语义分析。通过语义分析，程序可以理解当前打开的页面内容和含义，如：该页面是描述那方面的信息？该页面是一个搜索导航页还是信息描述页？信息的各字段分别是什么内容？该信息是否过期的？等。语义分析是一个必要的功能，而且语义分析做得好不好，将直接影响垂直搜索引擎的性能^[4]。很多通用搜索引擎在页面分析与处理中也有语义分析，但通用搜索的分析基本上都是针对全文特征概括^[5-6]或者检索语句理解^[7]。垂直搜索的语义分析则要做得更加深入，需要能够信息的详细内容进行准确定位和提取。语义分析需要一个词语库作为基础，因此需要为程序定义一个语义词典，不同行业的垂直搜索应该有不同词典。页面分析器对数据进行抽取和检验之后，就可以把信息按分析的结果分类存放进数据信息库，以备查询。最后应该建立一个检索系统给用户使用，检索系统可以是一个网站，也可以是企业内部使用的应用程序。

在搜索引擎中，搜索蜘蛛是最核心的模块。本文对这一模块的实现进行详细论述。

2 语义词典及语义分析

语义词典是搜索蜘蛛不可缺少的一个子模块。它是操作语义分析的基础数据。语义词典就是预先定义好的常用词和同义词库。每一个常用词以及它的同义词(如有的话)对应一个具有惟一性的标识符(一个数值)。如为了检测网页中是否提到电子邮件信息，应该使用一个同义词组{“邮件”，“邮箱”，“EMAIL”，“E-MAIL”}。在文本中搜索是否出现同义词组中的其中某个词。一个词语可能有多个含义，因此某些词语会拥有多个标识符。同义词组在程序中实时构造，判断两个词是

否同义就是判断两个词语是否拥有相同的标识符。语义词典可以用数据库定义，也可以用文件定义。

垂直搜索语义分析的目的不是要把文字内容像人阅读那样逐字逐句读懂理解。目前的技术水平也做不到这点。垂直搜索语义分析的目的在于确定以下几个问题：给定的页面是否是一个信息描述页；如果是，它描述哪方面的信息，是否属于引擎关心的领域；如果不是，页面上有没有可导航到信息页的超链接；如何确定某个超链接是要找的超链接；信息页中的信息是否合法，是否有效；如何进行提取以及如何进行分类？通过语义分析，网页被赋上一些属性。语义分析的结果，将影响程序下一步选择何种处理方法以及具体如何处理。

分词(切词)算法是通用搜索引擎经常用到的算法^[8]，分词算法是通过寻找匹配的最长词语，确定某段文字内容的分类。但仅仅靠分词，却无法应付垂直搜索的分析需要，因为垂直搜索做的分析要比通用搜索深入得多。为了做到准确的分析，需引入一个概念：规则。

规则是指一个判定某个页面或某段文本是否有某种属性而制订的条件。例如，要判断一个页面是不是某个论坛中的一个帖子，可以制订这样的条件：

- (1)网页文本中有“主题”、“作者”、“发表时间”的词语或者它们的同义词；
- (2)网页中有文本为“回复”(或它的同义词)的超链接，但这样的超链接又不得超过3个；
- (3)对于论坛的常见词汇，如“版主”、“楼主”、“发表”、“引用”、“注册时间”等，至少出现两个；
- (4)页中有时间数据文本。

符合以上几个条件的，可以判定该页是一个论坛帖子页，可以给该页赋上一个“论坛帖子”的属性，在信息采集时就可以用适合论坛的采集方式进行采集。而以上的几个条件就共同构成了一个判定论坛帖子的规则。为了判定页面的多种可能的属性，需要为程序制订很多的规则。规则的制订需要对各种页面的格式、网页制作者的编辑习惯、读者的阅读习惯进行仔细的分析。

3 搜索过程的实现

蜘蛛是专门搜索寻找新网页地址的模块，每个独立的非分类目录类型的搜索引擎都有这样一个搜索模块。蜘蛛是一个很形象又通用的名字，它经常还有其它类似的称呼，如“爬虫”、“网络机器人”等。它进行搜索的基本方法就是对一个已知页面的HTML进行分析，提取出所有超链接进行遍历，就好像在一只蜘蛛在蛛网上爬来爬去。与通用搜索相比，垂直搜索蜘蛛的任务有所不同，主要是两点：它不是要把找到所有超链接都进行遍历访问，而是把找到的超链接按规则过滤后再根据需要访问；它深入网站的程度要比通用搜索高得多，要深入到网站很深的层级去找到具体的信息。

因为垂直搜索是仅仅针对某个领域或者行业进行的搜索引擎，因此它不需要把无关的网站也进行访问查找。对于从HTML提取出来的超链接，如何知道它是否有用呢？譬如对于一个专门搜索招聘求职信息的垂直搜索，给定一个新的网站的超链接，如何知道这个新网站是不是一个招聘信息汇集

的网站呢?为了做这个判断,需要制定一个规则:

(1)该超链接的文本要出现“求职”字样,或者它的同义词(如“招聘”、“招工”、“人才”、“英才”等)。

(2)超链接文本的长度不超过 10 个汉字长度(满足条件 1 但文字很长的链接可能是新闻、求职经验之类的文章链接,不是信息集)。

依据这样的规则过滤,就可以从众多的超链接中取出搜索所需要的站点。需要说明的是,不同行业、不同类型的信息,要制定的过滤规则都是不一样的,因此垂直搜索并不能做成通用的模板,以像通用搜索那样包罗万象,而只能一个垂直搜索引擎专攻一个或少量几个特定领域。

另外,垂直搜索需要深入网站很深的层级去寻找信息。通用搜索的蜘蛛的目的只是找到这个网站,并判别一下网站的类别,就可以了,所以一般深入网站几个层级就可以了。垂直搜索则不然,蜘蛛需要沿着有用的超链接一直找下去,直到找到信息为止,过程中还可能要穿透网站的查找页和权限设置。

要穿透网站的查找页,蜘蛛必须有动态构造搜索脚本的功能。因为在查找页中,信息页的地址不是以超链接的形式给出,需要向查找页提交查找条件,才能取得信息页的地址。因此蜘蛛除了做文本的语义分析的同时,还需对页面的结构进行分析,检查页面是否包含查找控件,以及控件内部的文

本。控件查找可通过 HTML 标签识别实现,如查找 HTML 中是否有 FORM 表单,表单内部是否有 SELECT 标签,SELECT 内部的 OPTION 又是什么文本。然后蜘蛛依靠控件内给出的备选条件构造自己的查询条件,动态改写 HTML 代码,然后再按新的 HTML 加载。加载后,使 HTML 自动提交。符合查询条件的信息页地址就会取到。HTML 表单自动提交可用以下的代码简单实现:

```
<script language = "JavaScript" type = "text/JavaScript">
form.submit()
</script>
```

HTML 改写加载可用 MFC 的 IHTMLDocument 接口实现。蜘蛛的程序流程可描述如图 2 所示。

蜘蛛爬行寻找信息页面过程中,需要把每个有用站点的主页记录下来,存入一个来源站点地址表中。下次找寻信息可直接从表中取出站点地址进行信息查找分析,而不是从新爬行。因为爬行是非常耗时的,这样做可以大大提高效率。当作过一两次完整的爬行后,站点表中就有很多指定领域信息的来源站点地址。可以假设互联网上网站更替的过程是缓慢的,所以此后完整的爬行不需要天天做,而只要一个月甚至两个月做一次即可。所以流程图中最左边的分支不需要每次启动都执行,毕竟这不是垂直搜索的主要任务,这与通用搜索有很大区别。

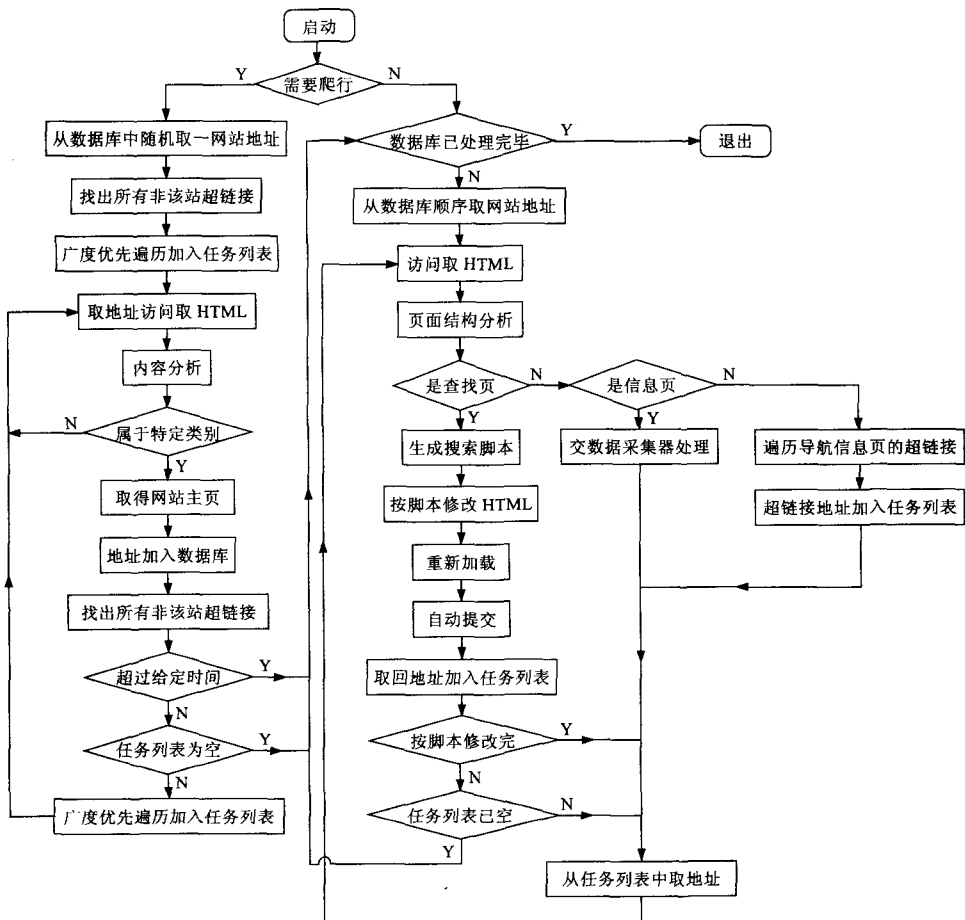


图 2 蜘蛛的流程

4 数据采集

蜘蛛寻找到的信息页面,最后由信息采集器进行数据采集。数据采集之前要根据页面文本的语义和规则判断信息页面的类型,页面类型可分为两种:格式化数据的和非格式化数据的。格式化数据是指数据经过整理、按规律填写好的页面数据。格式化数据在网页中多以表格形式出现。非格式化数据是指数据未经过整理,再网页中以一整段描述文字给出。

对于格式化的数据,如以下的一段 HTML:

```
<table class="systemfont" width="440" align="center" border="0">
<tbody>
<tr height="25">
<td width="68">工作地点:</td>
<td width="113">无锡</td>
<td width="87">月薪:</td>
<td width="154">1500 元/月</td>
</tr>
<tr height="25">
<td>学历要求:</td>
<td>初中及以上</td>
<td>外语语种要求:</td>
<td>不限</td>
</tr>
<tr height="25">
<td>性别要求:</td>
<td>男</td>
<td>招聘人数:</td>
<td>1 人</td>
</tr>
</tbody>
</table>
```

格式化的数据采集相对简单。因为表格数据有规律可寻,一般是文本都是“信息元素名”和“信息元素值”间隔出现。如上面的招聘信息,要查找该信息给出的“月薪”是多少,可以遍历页面的文本结构,查找包含“月薪”(或同义词)字样又不超过 10 个字符的文本元素。显然,表格中第 3 个文本元素“<td width="87">月薪:</td>”就是符合要求的文本元素,然后检查它后面一个文本元素,整理提取出文本“1500 元/月”就是查找的信息值。信息值取出后,可以写入数据库记录的相应字段中。

对于非格式化数据,一般都以整段的文本出现,如下:

```
<div>
<p><font size="6">本人有一套位于萍矿十字路口的新房子出租。三室二厅,高档装修,家具:煤气电视和洗衣机热水器一套俱全。条件:押金 3000 元,房租 700 元,费用不包。
```

```
有意者请打电话 13607998951,李。或网上留言</font></p>
</div>
```

非格式化未整理过,因此数据采集相对复杂,对每一类信息都需要定义一整套的信息提取规则。如要检索上述房屋出租信息中的居室数信息(即几居室的房子)。需要制定如下规则:①查找文本中是否存在“房”(或同义词“室”)字样;②若存在,检查该字符前面的字符是否为数字或数字的同义词;③若是数字,提取数字并转化为整数类型数据。

依照该规则检索文本,可找到有两处“房”字和一处“室”

字,但两个“房”字前面的字符都不是数字,“室”字前面是数字“3”的同义词“三”,对词语“三”转化求值,得出 3。最后得出信息中的居室数为 3。写入数据库记录的相应字段。

在记录信息的数据库中,不同的分类信息都要对应一个或多个表。这是用户检索时要查询的数据,是数据库中最主要的数据。数据表的设计根据搜索的信息种类不同而异,如招聘求职的分类信息,可以设计“招聘单位”、“职位类别”、“招聘人数”、“学历要求”、“工作地点”、“薪金范围”、“详细要求”、“发布日期”等。与通用搜索不同,垂直搜索记录的是通用搜索不敏感的实时动态数据,通用搜索记录的最小单元是页面,垂直搜索记录的最小单元是信息。因为信息是实时更新的,而且很多信息有效时间很短,所以垂直搜索的回访时间要非常快,一般一两天一次,甚至每小时一次。而一个站点内可能就有非常多的信息,所以数据库中的信息量也非常庞大。因此应该注意数据库的设计,以保证访问效率。如应注意对分类信息表进行纵横分割设计,使一个表不会有太多记录;每个信息量大的表要建立索引;SQL 语句要优化;尽量采用存储过程等。

大量数据记入数据库后,数据库就可以为前台检索界面执行查询操作。检索界面是面向用户的系统模块,负责接收用户查询条件,检索数据库,返回结果集。如 Google 和百度的主页其实就是搜索引擎的检索界面。检索界面可以是网页,也可以是桌面应用程序。面向互联网用户的检索界面一般做成网页,面向企业内部的检索界面一般做成普通应用程序。网站构建和数据库应用程序开发可用 ASP、VC+ADO 或其它类似技术实现。关于网页设计和数据库应用程序的书籍和资料非常多,因此此处不再详细论述。

5 一个实例

依靠以上设计方法,笔者曾为多家单位研制过各类信息的垂直搜索引擎,其中包括肇庆市人才信息中心的职业信息检索系统,广州亿居房产公司的房屋信息检索系统等。引擎系统的信息全面度、数据量、准确度、更新速率以及检索时间等指标均良好。图 3 为亿居公司的房屋信息系统界面。

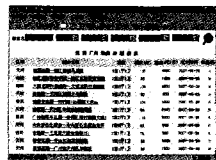


图 3 一个垂直搜索引擎系统实例

6 结束语

随着信息技术和因特网的发展,垂直搜索引擎在网络信息资源检索中的地位日渐重要。垂直搜索引擎可以综合通用搜索引擎和分类目录两者的优点,针对性强、查准率高,特别适用于搜索通用搜索难以搜到的动态网页和实时信息,而且高度自动化,不需要人工整理,很容易取得庞大数量的信息。本文论述了通过与通用搜索进行技术比较,分析了垂直搜索引擎结构,并给出了垂直搜索引擎蜘蛛的研制方法。其中详细阐明了语义分析的方法以及蜘蛛的搜索方法。

(下转第 4812 页)

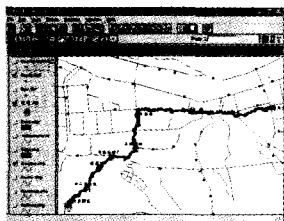


图3 GPS/DR 组合定位轨迹

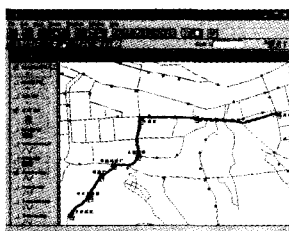


图4 修正后的组合定位轨迹

假设车辆的行驶轨迹是兰州理工大学家属院—中心实验室—弹簧厂—红旗机械厂—上西园桥—小西湖—小西湖公园—文化宫—西关什字。应用地图匹配算法对 GPS/DR 组合定位的数据进行了处理,取匹配修正权值 $w = 0.7$ 。

由表1的数据,可以明显看出,经地图匹配修正后,组合定位系统的定位误差明显减小,能正确的反映车辆的行驶路

线,提高了对车辆航线的跟踪质量。

3 结束语

本文研究了地图匹配技术在 GPS/DR 组合定位系统中的应用,给出了一种基于 D-S 证据推理的地图匹配算法,建立了 GPS/DR/地图匹配组合定位系统模型。通过 Matlab 仿真,结果表明,应用该地图匹配算法能够在很大程度上降低定位误差对地图匹配效果的影响,从而大大地降低地图匹配的误配率,对提高导航跟踪系统的性能起着积极的作用。

参考文献:

- [1] 王志刚.车载导航 GPS/DR/MM 组合定位技术的研究[D].武汉大学,2005:26-29.
- [2] 张威.基于 GPS/MM 组合的车辆定位技术研究[D].南京航空航天大学,2004:32-34.
- [3] Zhao Y L. Vehicle location and navigation systems[C]. Norwood, MA: Artech House, 1995.
- [4] 张守信. GPS 卫星测量定位理论与应用[M]. 长沙:国防科技大学出版社,1996:79-88.
- [5] 毕军.车辆 GPS/DR 定位系统、地图匹配及路径规划技术的研究[D].北京理工大学,2003:48-52.
- [6] 彭飞,柳重堪,张其善.基于代价函数的组合导航系统地图匹配算法[J].北京航空航天大学学报,2002,28(3):261-264.

表1 定位数据修正前后结果

	修正前	修正后
东向、北向位置误差	小于 25 m	小于 15 m
东向速度误差	小于 15 m/s, 其中大约有 85% 的误差在 5 m/s 以内	小于 12 m/s, 其中大约有 82% 的误差在 4.5 m/s 以内
北向速度误差	小于 12 m/s, 其中大约有 80% 的误差在 4 m/s 以内	小于 10 m/s, 其中大约有 78% 的误差在 3 m/s 以内
东向加速度误差	小于 3.5 m/s ² , 其中大约有 78% 的误差在 2.5 m/s ² 以内	小于 3 m/s ² , 其中大约有 80% 的误差在 2.5 m/s ² 以内
北向加速度误差	小于 4.5 m/s ² , 其中大约有 76% 的误差在 2.5 m/s ² 以内	小于 4.5 m/s ² , 其中大约有 83% 的误差在 2.5 m/s ² 以内
东向位置标准差	6.7027 m	5.3676 m
北向位置标准差	6.8958 m	5.2661 m
东向速度标准差	1.2980 m/s	1.0751 m/s
北向速度标准差	1.3866 m/s	1.2031 m/s
东向加速度标准差	0.7341 m/s ²	0.65312 m/s ²
北向加速度标准差	0.75055 m/s ²	0.70202 m/s ²

(上接第 4665 页)

参考文献:

- [1] 徐春艳.网络搜索引擎分类目录检索功能研究[J].图书馆学研究,2003(7):56-59.
- [2] Krol C. Specialization comes to search[J]. BtoB, 2005, 90(5):19.
- [3] 田野.垂直搜索火热为哪般[J].中国计算机用户,2005,37(9):11.
- [4] Chau M, Qin Jialun. SpidersRUs: Automated development of vertical search engines in different domains and languages[C]. Proceedings of the 5th ACM/IEEE Joint Conference on Digital Libraries, 2005:110-111.
- [5] 邹娟,周经野.一种基于语义分析的中文特征值提取方法[J].计算机工程与应用,2005,41(36):164-166.
- [6] Ali Selamat, Sigeru Omatu. Web page feature selection and classification using neural networks[J]. Information Sciences, 2004, 158:69-88.
- [7] 钱兵,王永成.面向搜索引擎的自然语言理解的设计与实现[J].计算机应用研究,2006,23(12):260-262.
- [8] 潘以锋.基于 Lucene 的网站全文检索系统的开发[J].广西教育学院学报,2006,85(5):63-66.