

中文元搜索引擎调研报告

[作者] 龙宇巍, 王永成, 许欢庆

[单位] 上海交通大学计算机系

[摘要] 为了帮助人们对中文搜索引擎的利用, 本文在对中文搜索引擎详加调研的基础上, 提出了一个调研报告, 报告中分门别类地详细介绍了中文搜索引擎的现状, 这也可当作一种现有中文搜索引擎的索引, 供关心中文搜索引擎者参考。

[关键词] 元搜索引擎, 中文搜索引擎, 索引

编者按: 此文中使用的“元搜索引擎”一词其指称与学术界一般使用该词的指称有所不同。我们尊重作者用词, 未作改动。

1、引言

当今的世界是信息的世界, 网络上的信息资源飞速膨胀, 如何在浩瀚如海的信息空间里快速查找并获取所需的信息已成为信息时代重要的问题之一。网络搜索引擎在网络信息资源查找中起到了重要的作用, 它可以帮助人们从数以亿计的网络信息中找到自己想要的信息。目前常见的综合型搜索引擎优点是用户可以查到范围很广的信息, 不足之处在于由于其涉及领域太广, 因此在某些特定领域的查询上则不够深入和专业化。针对这种状况, 人们提出了对某一专题的定题搜索引擎, 它可以在某些小范围的领域取得比综合型搜索引擎更满意的结果, 满足了某些特定用户的需要。

数据采集 robot 是搜索引擎的信息搜索代理, 它顺着超链接在网络上爬行且将搜索到的网页存储到数据库中。搜索引擎依靠 robot 的帮助获取大量的网页来进行索引以供用户查询。传统的 robot 爬行算法是给定少量 URL 作为种子集, 然后沿着 URL 在网络上爬行, 下载所遇到的网页。而对于定题搜索 robot, 由于 web 中信息极为庞杂, 主题相关信息只占其中少部分, 这样搜索难以做到有的放矢, 效率难以提高。经过研究, 考虑改进定题搜索的数据采集算法, 利用知名综合型搜索引擎已有的成果, 实现元搜索引擎。元搜索先对综合型搜索引擎进行主题相关的检索, 分析返回页面, 下载结果 URL, 将得到的页面并进行分析、存储、提纯, 得到一个庞大的初始结果集。这个结果集综合了许多搜索引擎的查询结果, 并经过相应的处理, 已经比较完备。然后再利用此结果集对超链进行一定的分析判断后, 按照传统的沿超链递归方法爬行, 对结果集进行扩展。

目前我们已实现了一个中文定题搜索引擎, 搜索引擎的 robot 使用了多个与知名搜索引擎对应的元搜索引擎。人们常用的中文综合型搜索引擎有数百种, 为了保证定题搜索的查全率, 我们所使用的元搜索应该涵盖任何中文搜索引擎的所有内容, 为了确定应该实现哪些元搜索, 对此进行了详细的调研工作, 最后确定了总计 16 个综合型搜索引擎应该包括到我们的元搜索引擎系统里。换句话说, 普通用户使用这 16 种搜索引擎, 基本上可以获得网上能找到的中文信息。下面是详细的调研结果。

2、调研报告

根据调查,目前或曾经所使用的中文搜索引擎超过 100 种,但是其中有许多不是网页搜索引擎,而是如软件、游戏、音乐、图像、商业、股票、地图等搜索引擎,不符合我们需求。又有非常多的网站是在互联网热潮时匆匆建立的,然后在互联网冬天时很遗憾地倒闭或被购并,现在已基本上消失了,所以可用的搜索引擎只有 40 来种。而有些网站又并没有自己的东西,仅仅是调用了其它著名搜索引擎的搜索结果;还有一些网站所能搜索到的结果太少或者由于设计和繁简体编码等原因难以实现。最后我们发现适合用作元搜索的搜索引擎只有 10 来种。以下是我们对所有这些搜索引擎的分析报告。对那些可用的搜索引擎,使用了“钢铁”和“足球”两个关键词进行搜索测试,确定网站的规模,以供参考。

2.1 非网页搜索引擎,这里列出了主要的共 31 个

- 1) 地图搜索引擎 <http://www.chinesemap.net/> 地图搜索,已无法使用
- 2) 迷路啦搜索引擎 <http://www.milula.com/> 地图搜索,已无法使用
- 3) SoGua 搜索引擎 <http://www.sogua.com/> 音乐搜索
- 4) 九天音乐网 <http://www.9sky.com/> 音乐搜索
- 5) 音乐极限 <http://www.chinamp3.com/> 音乐搜索
- 6) 碧海银沙歌词查询 <http://www.yinsha.com/lyrics/> 歌词搜索
- 7) 歌词新概念 [http://www.51lrc.com/51lrc\[CD-*2\]Purple/index.asp](http://www.51lrc.com/51lrc[CD-*2]Purple/index.asp) 歌词搜索
- 8) 中文 Lyrics 歌词搜索 <http://www.lyric4u.com/> 歌词搜索,已无法使用
- 9) 中文大黄页 <http://www.chinabig.com/zhs/srch/> 电话号码查询
- 10) 计算机世界 <http://www.ccw.com.cn/> 查出的所有页面都是 IT 相关
- 11) 电子图书搜索网 <http://www.ebooksou.com/> 查询电子图书
- 12) 环境与搜索信息检索 <http://www.enviroinfo.org.cn/> 网站内部环境相关文章查询
- 13) 房产地图搜索 <http://www.house-map.com/> 房地产相关信息搜索
- 14) 搜索频道 TVB <http://www.chat.tvb.com.cn/baidu/out/TVB> 娱乐新闻搜索
- 15) 中国政网搜索 <http://www.search.gov.cn/> 各地区政府部门官方主页搜索
- 16) 华军软件园 [http://www.newhua.com/hj\[CD-*2\]ssyq.htm](http://www.newhua.com/hj[CD-*2]ssyq.htm) 软件搜索
- 17) 迈博健康资讯 <http://www.medboo.com/medsearch/> 医疗信息搜索
- 18) 三九健康网 <http://www.999.com.cn/NAVIGATOR/> 医疗单位搜索
- 19) 医搜通网 <http://www.medsou.com/> 医疗新闻,信息,单位搜索
- 20) 中国建筑搜索引擎 <http://member.chinamasonry.com/cgi-bin/search2.cgi> 搜索建筑网站
- 21) 鞋业搜索 <http://www.shoeseek.com/> 搜索鞋业网站
- 22) 中华美食网 <http://cn.5eat.lycosasia.comsearch/> 菜谱及餐厅搜索
- 23) 中国食品网 <http://www.tech-food.com/search/> 食品相关信息搜索
- 24) 搜酷—安徽网址大全 <http://www.soucool.com/> 搜索安徽的常用网站
- 25) 电影搜索网 <http://www.filmscouts.com/> 搜索电影下载
- 26) 中文商务网 <http://www.ccbcc.com/> 商务信息搜索,已无法使用
- 27) e68 产品搜索引擎 <http://www.e68.net/> 商务信息搜索
- 28) 文新网 <http://www.cbook.net/readbook/> 网络电子图书搜索
- 29) 考研加油站 <http://search.kaoyan.com/> 全国著名院校主页导航

- 30)中国油漆网 <http://www.chinesepaint.com/>油漆相关信息搜索
31)学生资源搜索引擎 <http://base.6to23.com/>学生个人主页搜索

2.2 网页搜索引擎

可用、独立并且已经实现了元搜索引擎的站点，这一类网站有 16 个

- 1)Google <http://www.google.com/>“钢铁”搜到约 205000 个，“足球”搜到约 1350000 个结果
 - 2)百度 <http://www1.baidu.com/>“钢铁”搜到约 1190000 个结果，“足球”搜到约 5050000 个结果
 - 3)网易 <http://search.163.com/>使用百度的技术，但结果不尽相同，“钢铁”搜到 559852 个结果，“足球”搜到 1951386 个结果
 - 4)新浪 <http://www.sina.com/>“钢铁”搜到约 235000 个结果，“足球”搜到约 1530000 个结果
 - 5)北大天网 <http://3.pku.edu.cn/>“钢铁”搜 194542 个结果，“足球”搜到 981453 个结果
 - 6)雅虎中国 <http://cn.search.yahoo.com/>“钢铁”搜到 1769376 个结果，“足球”搜到 6173606 个结果
 - 7)慧聪 <http://www.isearch.sinobnet.com/index.htm>“钢铁”搜到 448728 个，“足球”搜到 1881783 个
 - 8)3721 网络实名 <http://www.3721.com/>“钢铁”搜到 47553 个结果，“足球”搜到 318135 个结果
 - 9)中华网 <http://searcher.china.com/>使用慧聪的技术，但不是直接调用，结果不尽相同，“钢铁”搜到 272112 个结果，“足球”搜到 1881783 个结果
 - 10)中文纳讯搜索引擎 <http://naxun.sjtu.edu.cn/>这是新闻为主的网站，搜索结果每日更新，并且限制输出不超过 100 个。
 - 11)OpenFind 简体中文测试版 <http://cn.openfind.com/>“钢铁”搜到 347193 个结果，“足球”搜到 1490076 个结果
 - 12)北极星 <http://www.beijixing.com.cn/bjx01/start.shtm>“钢铁”搜到 4589 个结果，“足球”搜到 11641 个结果，网站反应较慢
 - 13)蓝帆科技 <http://www.search163.com/>对站点进行搜索，“钢铁”搜到 212 个站点，“足球”搜到 1001 个站点
 - 14)焦点网 <http://search.focus.com.cn/search/index.html>对站点进行搜索，“钢铁”搜到 89 个站点，“足球”搜到 597 个站点
 - 15)搜鼠 <http://www.sosoo.net/>对站点进行搜索，“钢铁”搜到 96 个站点，“足球”搜到 383 个站点
 - 16)搜狗搜索引擎 <http://search.sobao.com/>对站点进行搜索，“钢铁”搜到 204 个站点，“足球”搜到 709 个站点
- 没有自己的技术，而是调用的其它搜索引擎结果的网站，这一类重要的网站有 13 个，还有包括在中国搜索联盟中的其它 37 个加盟网站。
- 1)搜易 <http://www.souyi.com/>调用的多个搜索引擎 Google、百度、新浪、yahoo 等
 - 2)深圳之窗 <http://www.szptt.net.cn/>调用多个搜索引擎 Google、Yahoo、搜孤，21CN 等
 - 3)中新网 139 <http://www.china139.com/default.html> 网页搜索调用的 Google

- 4) 中文搜索引擎指南网 <http://www.sowang.com/> 直接调用的百度，结果与其完全相同
- 5) 21CN 搜索 <http://search.21cn.com/> 直接调用的百度
- 6) 上海热线 <http://search.online.sh.cn/> 直接调用的百度
- 7) 中国同学录 <http://www.5460.net/> 直接调用的百度
- 8) 搜狐 <http://dir.sohu.com/> 使用了百度的技术，但不是直接调用，搜索结果与网易相同
- 9) Lycos 中国 <http://search.lycos.com.cn/> 使用了百度的技术，但不是直接调用，搜索结果与网易相同
- 10) 广州视窗 <http://search.gznet.com/> 使用了百度的技术，但不是直接调用，搜索结果与网易相同
- 11) TOM <http://Tom.com/> 调用的 3721 网络实名
- 12) 263 搜索 <http://search.263.net/> 使用了慧聪的技术，但不是直接调用，搜索结果与中华网相同
- 13) 中国搜索联盟 <http://www.chinasearch.com.cn/> 网页搜索调用的慧聪
- 14) 中国搜索联盟还包括许多加盟网站，它们本就不是搜索站点，是新闻网站或者政府网站，不过是加盟以后给慧聪付费，调用慧聪搜索引擎，附加了搜索功能。它们“钢铁”搜到 448728 个结果，“足球”搜到 1881783 个结果。这些网站名字与 URL 如下。中国搜索联盟还在不断增加中。

- (1) 新华网 <http://www.xhnet.com/>
- (2) 人民网 <http://www.peopledaily.com.cn/>
- (3) 中国日报网 <http://www.chinadaily.com.cn/gb/worldrep/index.html>
- (4) 国际在线 <http://online.cri.com.cn/homepage.html>
- (5) 中青网 <http://www.cycnet.com/>
- (6) 中国广播网 <http://www.cnradio.com/>
- (7) 中国新闻网 <http://www.chinanews.com.cn/>
- (8) 千龙新闻网 <http://www.beijingnews.com.cn/>
- (9) 三晋热线 <http://www.sx.cinfo.net/>
- (10) 千龙网 <http://www.qianlong.com/>
- (11) 北方网 <http://www.enorth.com.cn/>
- (12) 长城在线 <http://www.hebei.com.cn/node2/index.html>
- (13) 中国山西 <http://www.shanxi.gov.cn/>
- (14) 东北网 <http://news.0451.net/>
- (15) 东方网 <http://www.eastday.com/>
- (16) 中国浙江 <http://www.cnzj.org.cn/>
- (17) 中安网 <http://www.anhui.news.com/>
- (18) 福建东南新闻网 <http://www.fjsen.com/>
- (19) 大众网 <http://www.dzdaily.com.cn/>
- (20) 荆楚新闻网 <http://www.99sky.com/>
- (21) 红网 <http://www.rednet.com.cn/>
- (22) 南方网 <http://www.southcn.com/>
- (23) 桂龙新闻网 <http://www.gxnews.com.cn/>
- (24) 华龙网 <http://www.cqnews.net/>
- (25) 四川新闻网 <http://www.newssc.org/gb/newssc/root/index.html>
- (26) 金黔在线 <http://www.gog.com.cn/jqzx/>

- (27) 中国宁波网 <http://www.cnb.com.cn/gb/node2/node12/index.html>
- (28) 每日甘肃 <http://www.gansudaily.com.cn/20030702/default.htm>
- (29) 青海新闻网 <http://www.qhnews.com/>
- (30) 北京新闻网 <http://www.beijing.org.cn/>
- (31) 上海浦东 <http://www.pudong.gov.cn/gb/node2/node5/>
- (32) 中国彩虹 <http://www.jilin.gov.cn/2002new/>
- (33) 陕西通 <http://www.sxtong.org/1107/default.asp>
- (34) 龙虎网 <http://www.longhoo.com.cn/gb/longhoo/index.html>
- (35) 天润互动 <http://www.c315.com/>
- (36) 京报网 <http://www.beijingdaily.com.cn/>
- (37) 中国国际电子商务网 <http://www.ec.com.cn/publish2/index.shtml>

由于倒闭或者被购并,或者公司经营目标转向而现已无法使用的搜索引擎,列出了曾经有点名气的 16 个,还有超过 100 个无名的没有列出

- 1) 找到啦搜索引擎 <http://search.zhaodaola.com/>
- 2) 友发搜索引擎 <http://www.youfa.com/>
- 3) 如意搜索引擎 <http://www.ruyi.com.cn/>
- 4) 奔腾搜索引擎 <http://www.search.bentium.net/>
- 5) Asiayeah 搜索器 <http://www.asiayeah.com/>
- 6) 哇塞中文网 <http://www.hksrch.com/>
- 7) 八爪鱼搜寻机 <http://www.octor.com/>
- 8) 中国通 <http://www.chinatone.com/>
- 9) Go 搜索 <http://www.go.com/>
- 10) 搜星网 <http://www.soseen.com/> 以上都是已经消失的。
- 11) 中公网—网典搜索引擎 <http://www.wander.com.cn/> 变成某蓝牙公司的主页
- 12) 悠游搜索引擎 <http://www.goyoyo.com.cn/> 变成 Myjob 公司主页
- 13) 指南针 <http://www.compass.com.cn/> 变成一个科技公司主页
- 14) FM365 搜索 <http://www.lyric4u.com/> 可以进入到网站主页,但是无法进入搜索引擎

界面

- 15) LycosAsia 香港 <http://hk.lycosasia.com/> 可以进入到网站主页,但进行搜索时出错

- 16) 亦凡搜索 <http://www.gotofind.com/opendir/> 可以进入到网站主页,但进行搜索时出错

上面列出的大都是曾经有一定的规模和影响的搜索引擎,最终仍然没能坚持下来的。还有很多很多没有太大名气的已经消失的搜索引擎,这里没有一一列出,详细可见网页 <http://202.120.227.52/service/search.htm>,页面中列出了超过 250 个搜索引擎的,现在只有约四分之一的链接还有效。

由于搜索结果太少,且结果在其它重要搜索引擎中都可以查出来,这一类网站共有 6 个。

- 1) 天虎导航 <http://data.sc.cninfo.net/cgi-bin/search/tyfo> 只能查询站点,“钢铁”只查出 3 个结果,“足球”只查出 89 个结果
- 2) 搜友搜索 <http://soyoo.nbnet.com.cn/> 只能查询站点,“钢铁”查出 1 个结果,“足球”查出 61 个结果
- 3) AsiaTop <http://www.asia-top.com/s> 只能查询站点,“钢铁”搜到 2 个结果,“足球”只搜到 92 个结果

4)木子 <http://search.muzi.com/>只能搜索到网站,“钢铁”搜到9个结果,“足球”搜到148个结果

5)coo 台湾索引 <http://search.coo.com.tw/>只能搜索站点,“钢铁”搜到3个结果,“足球”搜到123个结果

6)深圳热线 <http://www.szonline.net/>它的搜索是站内搜索,搜到的都是站内新闻或报道,没有太多有价值的东西

其它原因导致元搜索引擎难以实现的,这一类网站有5个

1)番薯藤 <http://www.yam.com/>可进出网站主页,但搜索引擎需要繁体输入,简体无法搜索

2)添达香港搜索器 <http://www.hksrch.com/>可进入网站主页,搜索引擎需要繁体,简体无法搜索

3)FastSearchwww.alltheweb.com这是一个国外网站,关键词查询时会自动转换为UTF—8编码,导致元搜索引擎难以实现

4)香港世页 <http://ipoinc.com.hk/ipo2/gb/index-gb.html>网站实现机制比较特殊,搜索时先会取到第一页,然后后面的页面不是用查询的方式表示,而是动态生成网页的编号如<http://ipoinc.com.hk/ipo2/tmp/page2454988.html>、[page2454901.html](http://ipoinc.com.hk/ipo2/tmp/page2454901.html)、[page2454902.html](http://ipoinc.com.hk/ipo2/tmp/page2454902.html)等形式,还会在机器上建立Cookie,以后就根据Cookie取页面,所以用普通的办法无法取到后面的值,元搜索引擎难以实现。且搜索结果不多,网站规定了结果集上限为500个。

5)浩瀚搜索 <http://search.fjii.com/URL>中有一个奇怪的code参数,变化的规律无法摸清,应当是通过一些复杂的密码变换得到,技术上无法实现元搜索引擎。不过网站查询结果本身比较少,“钢铁”搜到11个站点,“足球”搜到50个站点。

3、结论

本文中列出了常用的中文搜索引擎共124个,覆盖范围比较广,甚至包括了许多名气不大或者曾经辉煌而现已不存在的。除此之外,还有100多个已经消失的就没有一一列出,读者可在<http://202.120.227.52/service/search.htm>上看到这些搜索引擎的名字,它们的URL大多已无法使用。基本可以肯定,国内的中文搜索引擎很少会有超过这个范围的了。结果中还包括了国外的知名搜索引擎的中文版本,Google,OpenFind 简体中文测试版,Yahoo 中国等。

在对这124个中文搜索引擎进行详细研究、层层筛选过后,发现只有16个网站有价值作为元搜索引擎。具体说来,124个中文搜索引擎中,有31个不是网页搜索引擎;有13个有名的搜索引擎是使用了其它公司的技术,与它们的搜索结果相同;还有37个站点加入了中国搜索联盟,使用慧聪公司的技术,搜索结果与慧聪完全相同;剩下43个站点,其中有16个是网站已经不存在或者变成其它公司的主页或者不再提供搜索引擎服务。有6个站点只能搜索到少量的结果,对我们的检索不能起到帮助作用;还有5个站点是由于网站的设计机制或者由于繁体简体编码的不同导致我们难以实现。

我们列出的16个元搜索引擎,搜索到的内容基本涵盖了目前国内中文搜索引擎所能找到的所有网页。由于没有详细的数据,我们无法肯定说国内存在的所有网页都可以被我们搜到,因为可能有些页面太偏僻,没有被任何搜索引擎列入索引。但是,我们可以肯定的说,中文搜索引擎所能找到的网页,使用以上的16个元搜索引擎,基本上都可以找到。

参考文献

- 1、王永成等. 中文信息处理技术及其基础. 上海交通大学出版社, 1991. 12
- 2、朱茂盛, 王斌, 程学旗. 元搜索引擎及其实现. 计算机工程, 2002 年 11 月