元搜索基于源搜索引擎的链接优化*

程仁贵[1], 黎 明[2]
(1. 福建武夷学院 计算机科学与工程系, 福建 武夷山 354300; 2. 重庆师范大学 研究生处, 重庆 400047)

摘 要:现有的 Internet 网站中一些大的搜索引擎,其功能很全,正因其功能强大,则难免在细节上出现一些不太完善的地方,如对搜索结果的链接没有实现有效性检测,导致了在网页中搜索结果出现了许多无效链接。针对这一缺点,本文先给出了元搜索引擎的定义、元搜索引擎的原理及搜索引擎与元搜索引擎的主要区别,同时还指出了搜索引擎的不足,最后提出了一个搜索引擎改进方法,给出了思路流程与方案解析,即在客户端实现过滤,这样既可以保持搜索的快速性,又能达到过滤链接的效果,并用 UML 时序图描述了搜索引擎的检索结果,同时进行有效性检查,建立这样的元搜索构想功能即是对如上搜索得到的网址进行检测过滤,以提示或去除其中无效的链接,使用户能够更加准确、快捷地获取所需要的资料信息。

关键词:搜索引擎;链接检测;元搜索;过滤
中图分类号:TP393.09 文献标识码:A 文章编号:1672-6693(2008)04-0060-04

1 元搜索技术

如今的信息时代,人们越来越多地关注如何从一大堆网络信息中快速有效地获取有价值的信息。搜索引擎的崛起,满足了这一要求。现在,搜索引擎已成为一个网络门户,起着网络导航的作用。搜索引擎技术正成为计算机科学界和信息产业界争相研究、开发的对象。

任何搜索引擎的设计,均有其特定的数据库索引范围、独特的功能和使用方法,以及预期的用户群指向。一种搜索引擎不可能满足所有人或一个人所有的检索需求。在某些情况下,如文献普查、专题查询、新闻调查与朔源、软件及 MP3 下载地址搜索等等,人们往往需要使用多种搜索引擎,对搜索结果进行比较、筛选。为解决逐一登陆各搜索引擎,并在各搜索引擎中分别多次输入同一检索请求(检索字串)等繁琐操作,元搜索引擎应运而生[1-2]。

1.1 什么是元搜索引擎

元搜索引擎(Metasearch Engine),是一种调用其它独立搜索引擎的引擎,亦称"搜索引擎之母(The Mother of Search Engines)"。元搜索引擎就是对多个独立搜索引擎的整合、调用、控制和优化利用。相对元搜索引擎,可被利用的独立搜索引擎称为"源搜索引擎"(Source Engine),整合、调用、控制和优化利用源搜索引擎的技术,称为"元搜索技术"(Meta-searching Technique),元搜索技术是元搜索引擎的核心[3]。元搜索引擎通过一个统一用户界面帮助用户在多个搜索引擎中选择和利用合适的(甚至是同时利用若干个)搜索引擎来实现检索操作,是对分布于网络的多种检索工具的全局控制机制。

1.2 元搜索引擎原理[4-6]

元搜索引擎分为并行处理式和串行处理式两大类。并行处理式元搜索引擎将用户的查询请求同时转送给它调用链接的多个独立型搜索引擎进行查询处理,串行处理式元搜索引擎将用户的查询请求依次转送给它调用链接的每一个独立型搜索引擎进行查询处理。

可将元搜索引擎看成具有双层客户机/服务器结构的系统,用户向元搜索引擎发出检索请求。元搜索引擎再根据该请求向指定的一个或多个搜索引擎发出实际检索请求;搜索引擎执行元搜索引擎检索请求后将检索结果以应答形式传送给元搜索引擎,元搜索引擎将从搜索引擎获得的检索结果经过整理再以应答形式传送给实际用户。

元搜索引擎是用户利用引擎进行网络搜索的中介。检索时,元搜索引擎根据用户提交的检索请求,调用源搜索引擎进行搜索,对搜索结果进行汇集、筛选、删并等优化处理后,以统一的格式在同一界面集

* 收稿日期:2008-06-24 修回日期:2008-08-10
资助项目:福建省教育厅资助项目(No. JA06066) 武夷学院校基金重点资助项目(No. XLZ06002)
作者简介:程仁贵(1968-),男,讲师,研究方向为计算机网络及应用、数据挖掘等。

中显示。元搜索引擎虽没有网页搜寻机制,亦无独立的索引数据库,但在检索请求提交、检索接口代理和检索结果显示等方面,均有自己研发的特色元搜索技术索引支持。对检索结果的显示,不同的元搜索引擎有不同的处理技术,由于元搜索引擎设定的检索结果排序依据、最大返回结果数量、相关度参数及优化机制等不同,调用相同源搜擎的不同元搜索引擎显示检索结果的数量多少、排序先后、结果信息描述选择亦有较大差异。

1.3 元搜索引擎与传统搜索引擎的区别

搜索引擎与元搜索引擎的主要区别在于搜索引擎拥有独立的网络资源采集标引机制和相应的数据库,而元搜索引擎一般没有自己独立的数据库,更多地提供统一联接界面(或进一步提供统一检索方式和结果整理),形成一个由多个分布的、具有独立功能的搜索引擎构成的虚拟整体,用户通过元搜索引擎的功能实现对这个虚拟整体中各独立搜索引擎数据库的查询显示等一切操作。元搜索引擎中各独立搜索引擎被称为“目标搜索引擎”,或者“成员搜索引擎”,它们各自保持其原来的局部数据模式和自己的检索指令;元搜索引擎给出一个全局外部模式,用以接受用户检索输入和结果输出。不过,有些元搜索引擎给出的全局外部模式不够完善。

2 搜索引擎链接失效、重复的不足之处

搜索引擎以其强大的数据库系统和搜索机器人自动抓取相关网页的能力让世人赞叹不已,用户对搜索引擎的依赖性也越来越高,对于搜索引擎结果的良好期望也越来越大。而目前即使像百度、Google 这样大型的搜索引擎抓取的结果也难以避免结果出现无效、重复的情况,用户不得不在一页页的隐藏失效结果的 Web 页面中去找出对他们有效的信息,起码是链接地址正确无误的信息。这无疑加重了获取信息过程中所占用的时间成本。而在今天,人们不仅需要高速度、高质量的信息,而且需要准确获得信息的通道。

如果搜索引擎结果能够提供更加准确无误、清晰的信息,让用户在最短时间里取舍判断,进而选取正确无误的结果链接去获得其想要的信息,将是吸引用户使用的一大优势。

现有的 Internet 网站中一些大的搜索引擎,其功能很健全,而正因其功能强大,则难免在细节上出现一些不太完善的地方,如对搜索结果的链接没有实现有效性检测,导致了对网页中搜索结果出现了许多无效的链接。例如利用百度或 Google 搜索引擎搜索网址时,当用户在搜索栏中键入某些关键字(Key)后,会出现一个搜索后所得的 Web 结果页面,Web 页面出现一系列相关网址,其中连同错误的链接也显示出来,在某种程度上降低了搜索效率,给用户带来了不便,用户点击搜索得到的结果后时常出现链接无效的错误页面。针对这一缺点,作者建立的元搜索构想功能即是对如上搜索得到的网址进行检测过滤,以提示或去除其中无效的链接,使用户能够更加准确、快捷地获取所需要的资料信息。

3 元搜索引擎的设计

3.1 系统架构

元搜索引擎的系统架构见图1。

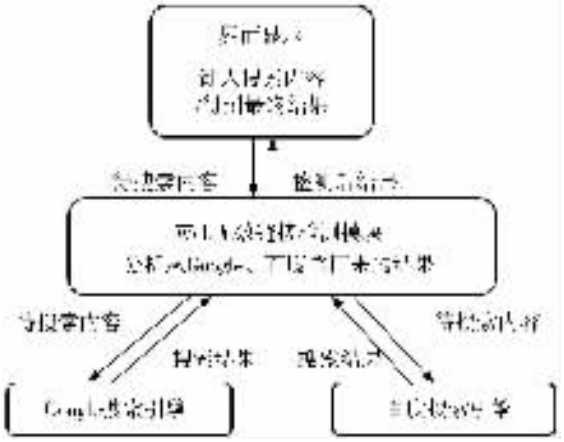

图1 系统结构图

3.2 思路流程与方案解析

3.2.1 简要思路 作者以元搜索网站的形式作为百度搜索引擎的代理,用户登录到作者的元搜索网站中进行搜索,搜索方法与其它搜索网站基本相同,作者将用户的查询请求转送给它调用链接的搜索引擎进行查询处理,取回用户搜索的结果后将对其进行结果过滤检测有效性处理,将提示或去掉错误链接后的正确搜索结果返回给用户。对于用户来说,上述过程是完全透明的,用户不用关心如何操作实现,只需在元搜索页面中键入其需要搜索的关键字,点击搜索按键进行搜索,即可得到检测过滤后相应的搜索结果。

用户可以把笔者的链接放入收藏夹中,进行搜索时,打开笔者的网页,这和用户收藏百度到收藏夹一样。

笔者通过对百度搜索结果页面的分析,采用 JAVA 语言编制程序进行过滤提示,结合 APPLET 及多线程编程方法,通过 HTML 代理页面作为搜索结果显示界面,将结果返回给用户。

3.2.2 设计方案 构建该系统,首先要对其操作步骤进行剖析,由此划分好该系统各部分的功能模块,

各个模块所负责处理的事件，模块间如何交互，如何通讯等。

首先，需要有一个自己的 HTML 代理页面（即后面所提到的 APPLET 界面），该页面是与用户交互的模块，用户在页面上输入其需要查询内容的关键字，选择相应的搜索引擎，点击查询按钮进行操作后，会返回一个代理搜索的结果页面给用户，里面的搜索结果会给出相应的有效性提示信息。

当用户提交搜索的关键字和所选用的搜索引擎时，这两个参数将被送到另一模块进行处理，这一模块就是搜索代理模块。该模块主要负责到百度去取搜索结果，即执行用户相应操作的搜索请求，而后将取得的结果返回。

那么，取得的结果又要继续怎样的操作呢？因为返回的结果中有很多链接，包括搜索相应关键字后得到的内容链接和其他附加链接，如在百度中的“ 百度快照 ”;“ 更多搜索结果 ”等等一系列的对于笔者要处理的结果无太大意义的链接，需先把它们去掉，也即是从结果中提取要检测的链接后，才能对有效性进行分析，这样就涉及到需要另一个模块来专门负责对这些结果，即从搜索引擎返回的 HTML 源文件进行分析，提取出各个 URL 地址，这一模块就赋予了它结果分析的功能。

还需要考虑一种情况，即当对这些结果进行有效性验证的同时，用户会出现什么情况。由于验证需要一定的时间，可能超过 10 s，也可能是几十 s 甚至更长，这样一来，有可能用户会对着一个空白的页面，直到程序检验完毕后才看见搜索的结果，为了避免这种情况发生，可以先把还未检测的结果返回在笔者的代理页面中，然后用异步刷新的方式，运用 JAVA 中的 APPLET 技术，对每个结果的有效性提示进行局部刷新，这样一来也可以避免因为不断刷新而使页面屏幕闪烁不定的问题。还应该在结果分析这一功能模块上加上另一功能与其同时异步执行——界面控制刷新。

上面笔者设计了三个主要的模块，一个是 APPLET 代理页面模块，一个是搜索代理模块，还有一个是结果分析及界面控制刷新模块，那么，分析后提取到的 URL 结果又交到哪里去处理呢？这就是另一模块的职责，无效链接过滤模块。这一模块主要负责对提取后的每个 URL 地址进行检测，通过发送 HTTP 请求到 URL 网站后，得到其相应的状态码来判断其有效性，然后将有效性测试结果返回给结果分析及界面控制刷新模块，再由它来对代理页面结果进行刷新操作，其过程见图 2。

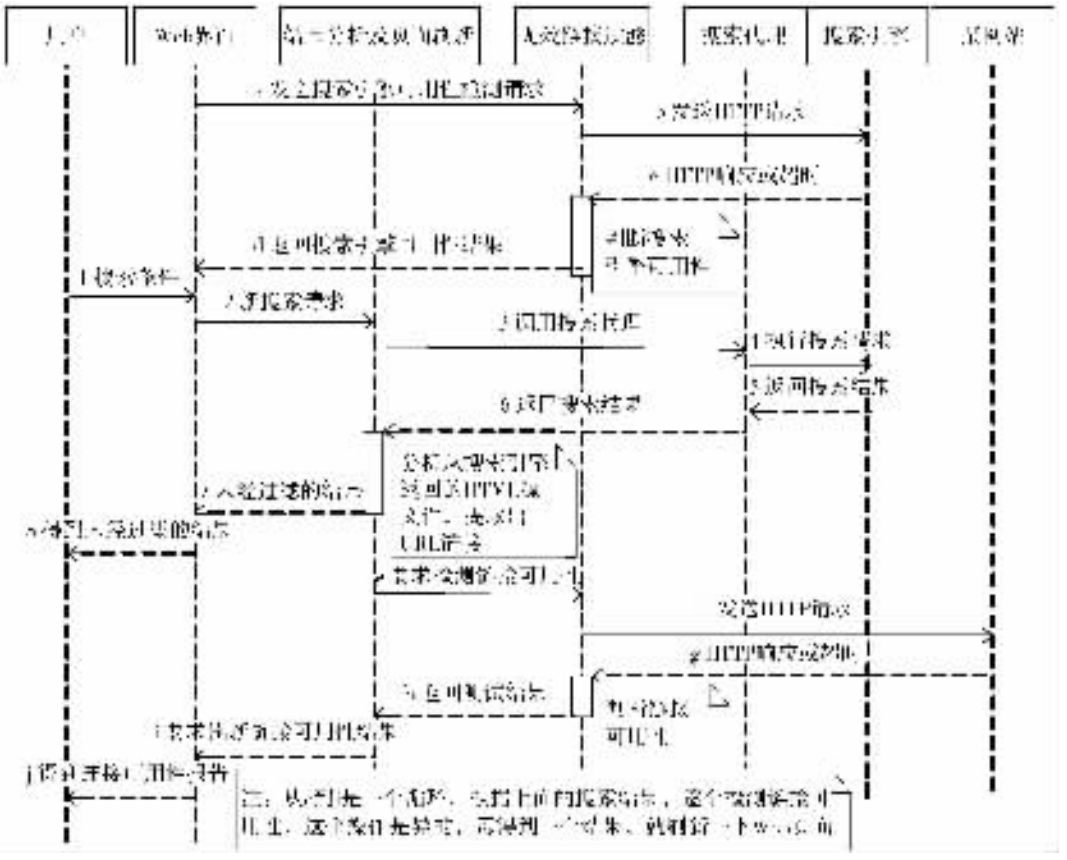

图 2　UML 时序图

4 总结

实现中可以进一步对每一个搜索引擎做一个 httpclient 封装类,以便以后扩展新的搜索引擎,对现有的修改也简单。

过滤的构思是在客户端实现的,由于 HTTP 协议是无连接、无状态的,如果应用于服务器端,则每一个用户请求就必须建立一次连接和进行一次链接处理,一旦同时使用的用户较多,服务器端的系统开销将会加重,其运作效率将会降低,所以采用在客户端的模式来实现,既可以保持搜索的快速性,又能达到过滤链接的效果。

参考文献:

[1] 马燕,邹显春,包骏杰,等. 一种互联网智能元搜索引擎模型的设计[J]. 重庆师范大学学报(自然科学版),2004,21(3):15-18.
[2] 张蕊. 元搜索引擎揭密[N]. 中国计算机报,2000(27).
[3] 严莉莉,王倩倩,孟杰,等. 基于聚类的个性化元搜索引擎设计[J]. 计算机技术与发展,2007,17(4):186-188.
[4] 陈大平. 集成搜索引擎与元搜索引擎比较研究[J]. 大学图书情报学刊,2005,23(1):42-43.
[5] 邢志宇. 集成搜索引擎与元搜索引擎[EB/OL]. (2003- 10-05)[2008-06-20]. http://www. sowang. com/ sousuo. htm.
[6] 许业卿,李金厚,黄明星. 一种基于 dotNET 的元搜索引擎的设计与实现[J]. 计算机与数字工程,2007,35(8):171-172.

Metasearch Based on Source Search Engine's Link Optimization

CHENG Ren-gui[1], *LI Ming*[2]

(1. Dept. of Computer Science & Engineering of Wuyi University, Wuyishan Fujian 354300;
2. Graduate Department of Chongqing Normal University, Chongqing 400047, China)

Abstract: Now some big search engines' function in internet are as clear as a bell, but they are hard to avoid not being perfect in their details, such as they can check the links' validity in the search result, and lead to appear a lot of error links In web page, and lower the search efficiency on a certain degree, so they bring the inconvenience to customers. In view of this weakness, the definition of meta-search, the principle of meta-search and the difference between search engine and meta-search engine are provided. Meanwhile, the weakness of search engines is referred in the paper. Finally, a search engine upswing method and the thought flow and scheme resolution are also given in the paper. That is to be filtrated in the client server. So it can not only keep rapidity of search, but also reach the effect of filtrating link, and decribe search result of search engine with UML time sequencing graph, and make the efficient check. The function of the meta-search which we have thought to build is to check or filtrate the links in the search result, to hint or delete these invalid links, and make the customers obtain the information accurately and fast.

Key words: search engine; link detection; META search; filtrate

(责任编辑 游中胜)