信息检索技术
清华同方光盘协办

编者按：清华同方光盘股份有限责任公司为发展我国“信息检索技术”，在理论和实践上推动网络信息检索技术的发展与应用，以进一步加快图书情报技术网络化发展进程愿与本刊合作，协办本栏目的工作，为此编辑部代表广大读者对清华同方光盘股份有限责任公司支持我国图书情报领域计算机信息检索技术发展的举措，表示衷心的感谢！

关于搜索引擎与元搜索引擎的讨论

张俭恭 陈定权 吴振新

（中国科学院文献情报中心 北京 100080）

【摘要】 首先探讨了搜索引擎的一般原理以及结构，然后介绍了元搜索引擎的概念及其框架。在最后，提出了一种将一般搜索引擎和基于OPAC的图书目录检索系统集成于一体的元搜索引擎的构想，该构想可以在一定程度上解决异构数据之间的兼容问题。

【关键词】 搜索引擎 元搜索引擎 全文检索 【分类号】 G354

Research on Search Engine and Meta Search Engine

Zhang Jiangong Chen Dingquan Wu Zhenxin

(*The Documentation and Information Center of CAS, Beijing 100080, China*)

【Abstract】 This article elaborates some principle and architecture about general search engine, and then introduces the concept and framework of META Search Engine. and brings forward a new idea that integrates the general search engine with OPAC—based retrieval system . This method maybe resolves the problem about data heterogeneity.

【Keywords】 Search engine Meta search engine Full text retrieval

CNNIC的最新调查结果显示，截止到2001年6月30日，我国上网计算机数已达1002万台，比去年同期增长54%，是三年前的18.5倍；目前我国网民2650万，半年内增加了400万；CN下注册的域名数已达128362个，比去年同期增长28.7%；WWW站点数达242739个；国际线路总容量为3257M，各项指标与三年前相比，均有了大幅的增长。可以看出，Internet和WWW都在以迅猛的势头持续发展，并且越来越多的人利用网络途径获取信息，进行交流。

那么如何能够更有效地获取所需信息就成了一个非常值得研究的课题。虽然人们可以通过浏览诸如Yahoo等门户网站的分类目录来找到自己感兴趣的网站，然后再通过链接到相应的网站寻找自己的所需信息；但多数人则是通过搜索引擎来完成他们信息的搜寻过程。上网用户首先向搜索引擎提供一个由多个关键词组成的提问式，这时搜索引擎通过访问本身的数据库，在进行一些匹配运算以后，就会返回一个包含有用户提问关键词的相关网页列表。本文首先要讨论搜索引擎的一般原理以及一些实现方法和技术。

收稿日期：2001—11—05

另外，在实现搜索引擎的过程中，由于各个搜索引擎的信息搜集和索引建立有很大的不同，使得它们在收集的信息资源范围方面产生了巨大的差异，任何单个搜索引擎都只能涵盖一部分WWW资源，这对于用户就意味着使用任何一个搜索引擎都不可能达到信息查全的目的。为了克服这个缺点，在该领域又出现了一种新型的搜索引擎——元搜索引擎。本文进一步探讨了一些元搜索引擎的实现问题，并对元搜索引擎提出了一些设想和展望。

1 典型搜索引擎的实现原理

虽然对于信息检索，已经有不少很好的算法和技术，但由于互联网信息资源数量庞大、更新速度较快以及分布存储方式等特点，使得搜索引擎必须在原来传统的信息检索算法基础上加以扩展，通过一些新技术实现信息搜集、建立和更新索引等工作。针对网络上巨大的信息资源数量，搜索引擎还应该完成检索结果的区分和排序工作，把最符合要求和最相关的网页链接地址优先提供给用户。

那么最典型的搜索引擎结构是怎样实现这些目标的？图1给出了一个典型的搜索引擎原理的框架，它基本包括机器人、索引、检索三大模块。

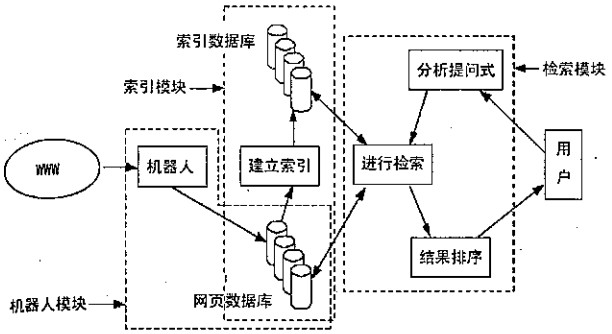

图1 搜索引擎模块划分

1.1 机器人模块

任何搜索引擎都会依赖一个机器人模块来完成它的信息获取工作，以期为将来的服务提供数据。而机器人就是一个可以浏览网页的程序，它很像真人的浏览过程，首先打开一个网页，然后再通过网页上的链接去浏览其它不同的网页，如此往复。工作的时候，机器人把开始确定的一组网页链接作为浏览的起始地址，然后将网页获取过来，抽取页面中出现的链接，并通过一定算法决定下一步要访问哪些链接；同时，机器人将已经访问的页面存储到自己的页面数据库里去。之后，机器人则继续重复这个访问过程，直至结束。在决定访问链接顺序的过程中，最常见算法有：深度优先、广度优先、有限深度/广度策略。当然，一般搜索引擎的机器人在实现的过程中，引入链长比(超链接数目与文档长度的比值)，只取链长比小于某一门限值的页面，即只采集内容页面，而不采集目录页面。在采集文档的同时，记录各文档的地址信息、修改时间、文档长度等状态信息，用于站点资源的监视和资料库的更新。在采集过程中，还可以构造适当的启发(heuristic)策略，来指导机器人的路径选择和采集范围，以减少文档采集的盲目性。

1.2 索引模块

当机器人访问完网页并将其内容和地址存入网页数据库以后，就要对其建立索引。索引模块总的来说是通过分析获取的网页，排除 HTML 等语言的标志符号，将出现的所有字或者词抽取出来，并记录每个字词的出现网址及相应位置，最后将结果存入索引数据库，就是一个很大的查询表，上面记录某个特定字词在互联网上出现的一组位置信息。

对于英文搜索引擎，由于是以单词为语言的基本单位，因此一般建立索引采用的都是词表法，即首先建立一个词表，然后将对应单词的出现位置记录下来。而检索的时候，就是以这些词语作为检索入口，并通过位置匹配可以实现多个词语的组合检索。但对于中文搜索引擎来说，由于语言的基本单位是汉字，在最底层往往采用的是字表法。和词表法相似，先建立一个汉字字表(一般采用 GB2312 汉字集)，然后对于网页中出现的汉字均记录在相应的字表项内。当检索的时候，采取字索引之间的位置匹配完成词语的检索。为了提高检索速度，一般还会在字索引的基础上建立一些词索引，有的是根据用户的提问动态生成已检索词的词索引，有的则是建立一个常用词表，然后生成这些词的索引。当然，无论是英文系统还是中文系统都会建立一个停用词表，以节省存储空间和提高检索效率。

1.3 检索模块

作为检索模块，首先分析用户检索时给出的提问式，再访问搜索引擎已经建立的索引，并通过一定的匹配算法，获得相应的检索结果。一般还会对检索结果进行排序，按照重要程度将结果有序地返回给用户。具体来说，当用户进行检索的时候，一般使用的是纯自然语言词汇或者是自然语言词汇组成的布尔逻辑式。对于前者，可以直接利用检索算法查询索引数据库中的词索引，或者是利用单字索引进行位置匹配，以获得检索结果。而对于后者，则首先要分析检索式的逻辑关系，分别对检索式中的各个检索词进行检索，最后再通过逻辑运算获得最终结果。由于网络上信息数量非常庞大，可能会产生一个相当大的结果集，那么如何精简结果以及如何将最重要的结果首先返回给用户就显得十分重要。最常用的方法是将结果按相关度进行排序，把引擎认为最相关的结果放在最前面。相关度计算有很多的算法，其中一个很重要的算法就是词频法，即通过计算网页中检索词的出现频率来决定该网页的相关程度，检索词出现次数越多则说明该网页越重要。虽然这种算法有很多缺陷，往往不能达到最好的效果，但由于计算网页中一个词的词频十分简单，使得该算法很容易实现。当获得检索结果以后，访问网页数据库，获得相关网页，并按照相应的格式和顺序生成结果网页，最终提供给用户，完成整个检索过程。

2 元搜索引擎的主要作用与框架结构

人们已经把搜索引擎作为在网络上查找信息一个非常重要的途径，从国外的 Yahoo、Excite、Altavisa 到中国的新浪、搜狐、中华网等，几乎每个门户网站都提供了搜索引擎的入口，所使用的搜索引擎可以是自己开发的也可以是从专业生产搜索引擎公司购买的。由于每个搜索引擎的实现方法、信息量以及收录站点等方面的不同，使得它们之间在处理内容上有很大的差异。当用户查找信息的时候，如果想要做到准确全面，他就必须访问不只一个搜索引擎。虽然这样的工作完全可以由用户自己来完成，但他们更希望能够只进行一次查询就可以获得多个搜索引擎有关查询的结果，而不是枯燥繁琐的重复劳动，这就是元搜索引擎的存在意义。它可以让查询一次完成，极大提高检索效率，节省用户的时间。

目前，在国外已经有 AskJeeves、Cyber411、DigiSearch、Dogpile、Highway61、IsIcuth、Mamma、MetaCrawler、ProFusion 等元搜索引擎，而在国内虽然中文搜索引擎已经有很多，但关于元搜索引擎的研究仍然很少，这就需要我们发展更多自己的中文元搜索引擎，以适应信息检索技术不断进步的需要。

所谓元搜索引擎，就是指在统一的用户查询界面与信息反馈的形式下，共享多个搜索引擎的资源库为用户提供信息服务的系统。元搜索引擎与搜索引擎的最大不同之处就在于它可以没有自己的资源库和机器人，它充当的是一个中间代

理角色，接受用户的查询请求，将请求翻译成相应搜索引擎的查询语法。在向各个搜索引擎发送查询请求并获得反馈之后，首先进行综合相关度排序，然后将整理抽取之后的查询结果提供给用户。这样由于信息源范围的扩大，不仅提高了检索效率，也大大增加了找到所需信息的可能性。

从结构讲，元搜索引擎主要包括 Web 服务器、结果数据库、检索式处理、Web 处理接口、结果生成等几个部分，其中用户通过 Web 服务器访问元搜索引擎，而元搜索引擎则通过 Web 处理接口访问其它外部的搜索引擎。

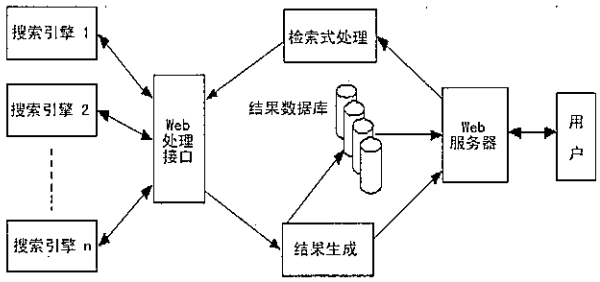

图 2 元搜索引擎结构框架

如图 2 所示，用户通过 WWW 服务访问元搜索引擎，并向 Web 服务器提出检索式。当 Web 服务器收到查询任务时，首先访问结果数据库，看在近期是否有相同的检索，如果有则直接返回保存的结果，完成查询；如果没有，那么就将检索式进行处理，分析并转化成与所要查找各搜索引擎相应的检索式格式，然后送至 Web 处理接口部分。Web 处理接口通过并行的方式同时查询多个搜索引擎，集中所有的查询结果。根据各引擎的重要性，以及所得结果的相关度，通过算法对结果进行抽取和排序，并生成最终结果网页返回给用户。与此同时，将此次结果保存在结果数据库中，以备下次查询参考。这就是整个元搜索引擎的服务过程。其中对于结果数据库中记录的处理，要指定一个生存期，也就是超过一定时间的检索结果要予以删除，以保证检索的时效性。

需要指出的是，首先由于大部分搜索引擎互不兼容，相互操作性差，而且用户接口不一致，使得检索式处理非常复杂。这不仅要求精确掌握各个搜索引擎在查询时调用 CGI 的格式，还要做到将当前检索式转化成相应格式。其次，由于不同搜索引擎反馈的结果页面格式相差很大，对于这些页面的处理难度也是相当大，一方面要解析页面找到查询结果，同时还要能够把这些结果的内容抽取出来，目前采用最多的是固定查找和智能判断相结合的策略。再者，作为一个元搜索引擎，如何能够将获取的信息按照相关度进行排序也是非常复杂的问题。因为不同搜索引擎在本身查询结果排序过程中采用的算法相差很大，甚至有一些未知的算法，而元搜索引擎必须结合这些使用不同排序算法产生的结果，并以统一的结果形式返回给用户。这些都是在研究元搜索引擎中遇到的难点，也是能否成功实现一个元搜索引擎的关键。

3 结束语

现在搜索引擎的技术越来越成熟，性能越来越好，可供选择的数量也越来越大，这更加促进了元搜索引擎发展。一方面，搜索引擎给用户提供的信息越来越多；另一方面，元搜索引擎让用户可以更快更容易地访问多个搜索引擎。两者之间互相促进，互为补充。

同时，对于一些像图书馆这样的信息单位，一般都缺乏高级的计算机网络人才，若想开发和维护一个独立的搜索引擎有很大困难，但如果能充分利用网上现有的搜索引擎来开发元搜索引擎，并结合专业特点进行优化，则完全可以建立符合用户要求的信息检索工具。

再者，由于目前的搜索引擎都是对网页进行浏览和检索的，并没有包括动态网页(也称为隐蔽网页)，因此建立在其基础上的元搜索引擎也存在这些缺点。那么，如果让元搜索引擎在访问其它搜索引擎的同时也检索动态网页，将是一个非常好的功能扩展。通过这种方法也可以在一定程度上解决异构数据之间的兼容问题。例如：在一个元搜索引擎中集成一般通用的搜索引擎的功能又能够集成基于 OPAC 的图书目录检索的功能。例如，在元搜索引擎中，当输入“华罗庚”检索词时，既可以在普通网页上检索出含有“华罗庚”的新闻、评论等页面，又可以在图书馆中检索出有关华罗庚自传、生平等书籍、华罗庚写的有关文章。

总之，目前元搜索引擎是搜索引擎之后在信息检索方面的又一个研究热点，它是以比较成熟的搜索引擎技术为基础，并且对其进行了扩展和综合。元搜索引擎在将来也会和搜索引擎相辅相成，共同发展，共同为用户的信息检索服务。

参考文献：

[1] 赖茂生. 计算机情报检索. 北京：北京大学出版社，1993，102－114
[2] 吴广印，胡亚莉. 非结构化网络数据库在数据情报服务中的应用. 现代图书情报技术，2001，(1)：16－19
[3] 邹 涛，王继成，张福炎. 基于 www 的资料搜集系统的设计与实现. 情报学报，1999，18(3)：195－201
[4] 邹 涛，王继成等. www 上的信息挖掘技术及实现. 计算机研究与发展，1999，36(8)：1019－1024
[5] Ricardo Baeza－Yates. Modern Information Retrieval. Harlow：ACM press，1999.

编辑部启事：

最近，本刊收到的稿件较多，但有部分稿件的作者在投稿时没有按本刊投稿要求撰写稿件，其中大量稿件缺少中英文题录、关键词、分类号、中英文摘要、分类号等有关项目。这直接影响了稿件初审的通过，造成了稿件审核时间长，不能及时发表的状况。故本刊再次重申，所有稿件的作者在稿件写成后，必须参照本刊已发表文章的格式，将所有项目配齐，然后投寄本刊。否则，稿件审核时间延误，由作者本人负责。

（本刊编辑部）