# 元搜索引擎及其主要技术

李广建 黄 崑

(北京师范大学管理学院信息技术与管理学系，北京 100875)

**摘 要** 本文介绍了元搜索引擎的涵义、特征及其体系结构、实现原理，并基于检索机制的划分标准分述不同元搜索引擎的类型和特点，分析元搜索引擎实现的主要技术，最后对开发中文元搜索引擎提出建议。

**关键词** 元搜索引擎 集搜索引擎 搜索引擎 信息检索

# Introduction and the Implementation Techniques of the MetaSearch Engine

Li Guangjian Huang Kun

(Beijing Normal University Dept. of Information Technology and Management, Beijing100875)

**Abstract** This article introduces the definition, features of the MetaSsearch Engine. At the same time it also includes the system structure and operation principles of MetaSearch Engine. The based the criteria of principles for retrievat, it describes the different categories of the MetaSearch Engine, as well as the major techniques applied on it. Finally it puts forward some suggestions on the development of Chinese MetaSearch Engine.

**Keywords** MetaSearch engine Multi—threaded engines Search engine Information retrieval

网络的飞速发展带来了网络信息资源数量的迅速膨胀，为了从纷繁芜杂的信息汪洋中挖掘出有用信息，人们越发离不开网络信息检索工具——搜索引擎的帮助。然而，任何一个搜索引擎都不可能100%的覆盖网上信息。据统计，国外的一些著名品牌的搜索引擎的信息覆盖率最高也不过30%，用户通常需要检索多个搜索引擎才能获得较全面的检索效果，这就为元搜索引擎的出现创造了生存的空间。

元搜索引擎是一种基于搜索引擎的搜索引擎，国外的英文搜索引擎中已经创下许多知名的元搜索引擎品牌，例如 Ixquick、Webcraler、ProFusion 等。元搜索技术正受到各大搜索引擎的关注，无论是如 Yahoo! 等老牌独立搜索引擎，还是如 Google 等搜索引擎的新秀，都已不再局限于独立搜索引擎技术，而是融合元搜索技术，以强化检索效果，提高检索质量。

## 1 元搜索引擎的界定及其特征

元搜索引擎是一种基于搜索引擎的搜索引擎，用于提供与查询需求相关的信息线索或者全文。元搜索引擎通过自己定制的检索界面，接收并处理用户的查询提问，在进行实际的查询时调用一个或者多个独立搜索引擎的数据库，搜索结果是来自独立搜索引擎的检索结果或者是这些结果集合的综合，结果呈现既可以是引用原始的独立搜索引擎的页面，也可以是由元搜索引擎重新定制后的形式。

元搜索引擎一般会采用品牌知名、检索效果较好的主流搜索引擎的数据库，一次提问同时检索多个数据库，提高了检索的效率，同时也起到了对检索工具的推荐和指南的作用。另一方面，元搜索引擎的检索模式还为各个搜索引擎的集成检索提供了的可能，具有一定的先进性和实用价值。

元搜索引擎区别于独立搜索引擎，主要有这样一些特征：

(1) 将一次提问提交多个数据库。元搜索引擎定制了调用多个独立搜索引擎的统一界面，将用户递交的提问提交给其它多个独立搜索引擎，因此，用户的一次查询可以同时检索多个独立搜索引擎。这期间，针对不同的独立搜索引擎将用户的提问作不同转换，以适应相应索引数据库的调用。也就是说，元搜索引擎需要对用户提问进行分门别类的处理，并根据不同独立搜索引擎的要求按不同的形式提交同一查询。

(2) 基于独立搜索引擎结果的二次加工。元搜索引擎的结果基于独立搜索引擎的查询结果，少数简单的直接调用原始的结果页面，但都实现了对独立搜索引擎查询结果的二次加工，如重复结果的删除、结果的再度排序等。

(3)标明结果记录的来源搜索引擎及其相关度。

收稿日期：2001—09—15

在定制结果输出形式的元搜索引擎中，检索结果一般都标明记录的来源搜索引擎及其相关度。

元搜索引擎的功能很大程度受独立搜索引擎的限制，因此不可避免的存在局限性。

(1) 检全率提高，检准率不易控制。元搜索引擎将一次提问同时检索多个搜索引擎，扩大了检索覆盖的范围，提高了检全率。但其结果主要来自独立搜索引擎查询结果中排名靠前的记录，在一定程度上默认了独立搜索引擎的检准效果，而目前独立搜索引擎自身在检全与检准提高方面存在着各种问题。因此，元搜索引擎在为用户提供更全面、综合的结果同时，难以控制各独立搜索引擎的无关输出。

(2) 检索功能简单。元搜索引擎一般只提供一个公共接口供用户输入查询词，实际查询在各个独立搜索引擎中实现。由于各家独立搜索引擎的检索语法不统一，以统一的查询入口执行多个数据库查询需要进行准确的提问分析和转换。对于简单的布尔逻辑检索和词组检索，元搜索引擎的检索效果很好；对于复杂的检索功能，效果并不是很好。因此，元搜索引擎一般只支持通用的检索句法，因此会抹煞独立搜索引擎中效果较好的高级查询功能。

## 2 元搜索引擎的运作原理

元搜索引擎一般包括用户提问处理、检索机制督导、结果加工处理和结果页面定制 4 个部分，如图 1 所示。

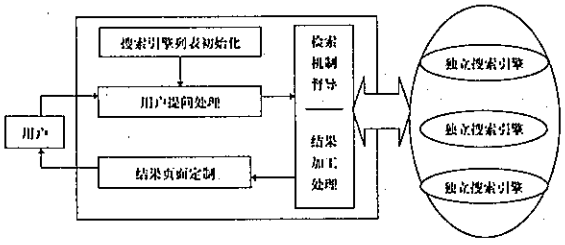


图 1 元搜索引擎体系结构简图

元搜索引擎在执行查询之前对要调用的搜索引擎列表进行初始化，初始化方式有系统默认和用户选择两种方式。系统默认方式，系统确定了用来检索的搜索引擎集合，用户无权变更；用户选择方式则允许用户自主选定需在哪几个搜索引擎中检索。

初始化结束后，元搜索引擎的查询入口接收用户查询词，针对不同的搜索引擎进行相应处理，将用户查询词转换成能检索不同数据库的提问表达式。

元搜索引擎的检索机制是元搜索引擎根据对各成员搜索引擎的检索结果测评分析而制定的一套规则，用于督导检索过程和结果输出过程。例如采用独立搜索引擎的何种检索功能（布尔检索、位置检索等）或检索类别（网页搜索、同站检索、类目检索等）。检索机制同时也限定了调用独立搜索引擎的搜索策略，采用串行搜索还是并行搜索。另外，在检索机制中还明确对结果反馈的要求，如在检索过程中是选择“结果最好”还是选择“速度最快”的标准来确定优先显示哪个搜索引擎的查询结果等等。

检索机制指导独立搜索引擎的检索过程，对从各个搜索引擎得到的结果作综合处理，这一结果处理过程对结果重复与否、结果之间相关大小等作出判断，最后遴选出满足条件的记录输出。

完成对结果的处理后，将最终结果以定制的界面呈现给用户。结果输出的页面定制形式在不同的元搜索引擎中有不同的体现，可以直接调用独立搜索引擎原始的反馈页面，也可以由元搜索引擎重新定制一个全新页面。

## 3 元搜索引擎的分类

元搜索引擎根据不同的标准可以划分为不同的类型，根据运作平台分，可以分为桌面型元搜索引擎和网络型元搜索引擎。

桌面型元搜索引擎是一种客户端元搜索软件，它与客户端环境充分结合，代理用户递交提问，一次性检索多个独立搜索引擎，并能获取实际的 Web 页面。目前已经有许多这类成型产品，如 copernic。桌面型元搜索软件的特点是结合客户端环境，更容易提供个性化的检索服务，但是多为收费软件。

网络型元搜索引擎是指提供检索服务的 WEB 元搜索引擎站点，它的使用更为普遍，用户通过浏览器就可以方便、免费地进行检索，例如：HotBot、SavvySearch、Mamma 等。其特点是使用便捷，通过浏览器就可以直接访问，操作简单。

根据检索机制可划分为集中罗列式元搜索引擎和统一入口式元搜索引擎，以下拟对这两种类型的元搜索引擎作详细介绍。

### 3.1 集中罗列式元搜索引擎

集中罗列式元搜索引擎按照一定的形式将所有

的独立搜索引擎集中呈现在页面上，并提供了一个公共的检索入口，但是实际上用户一次只进入一个独立搜索引擎检索。这种类型元搜索引擎的结果反馈页面多直接引用原始搜索引擎的结果页面。从表面上看，这类元搜索引擎与独立搜索引擎具有更多的相似点，其主要代表有 ezfind、Infodump 等。这类搜索引擎的特点是：

（1）一次检索一个搜索引擎；（2）检索结果直接调用原始独立搜索引擎的结果页面；（3）只支持原始独立搜索引擎支持的检索句法。

图 2 显示出了 ezfind 的检索界面。

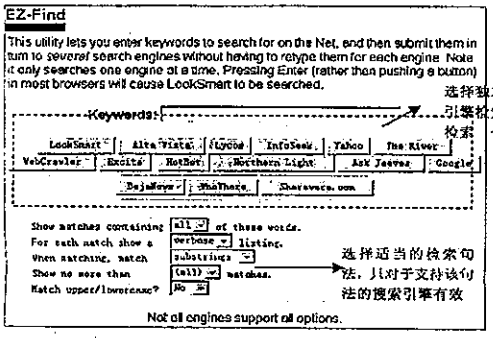


图 2　EZFIND 的检索界面

### 3.2　统一入口式元搜索引擎

统一入口式元搜索引擎为收录的独立搜索引擎建立了一个公共查询入口，用户发出检索请求后，提问式被分别提交给多个独立搜索引擎，最终反馈的结果是多个独立搜索引擎查询结果的综合。根据结果显示的不同，这类元搜索引擎又可分为直接调用原始页面型、混合综合型和分散综合型。

（1）直接调用原始页面型元搜索引擎。检索结果直接来自原始搜索引擎站点的结果页面，例如，ALL4ONE 的检索界面（图 3）就是一个典型的例子，该搜索引擎将查询内容分为 5 类，每一类中由系统默认调用 4 个独立搜索引擎来检索，以 The Web（网页检索）和 High—Tech News（高科技新闻）为例，前者检索使用 Altavista、Yahoo!、HotBot、Excite；后者则调用 CMP TechWeb、ZDNet、Cnet News、Wired News 一些新闻信息查询的站点。结果页面调用原独立搜索引擎的结果，如图 4。

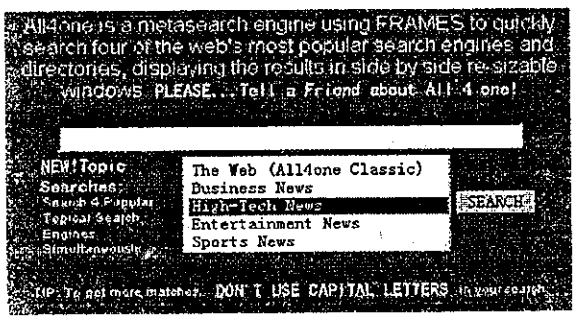


图 3　ALL4ONE 的检索界面

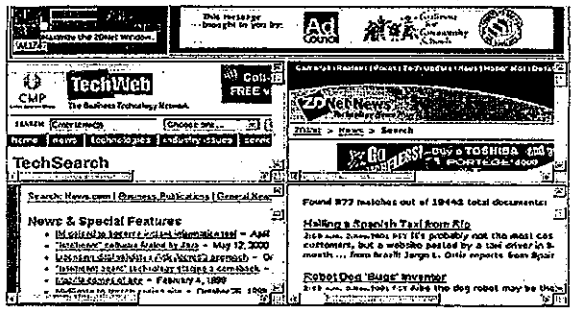


图 4　ALL4ONE 的查询结果界面

（2）混合综合型元搜索引擎。将各个独立搜索引擎中查找的结果进行综合，结果显示以记录为单位，记录描述包括该记录被检出的来源。例如 ixquick 的检索界面（图 5），它提供了 4 种查询范围：网页（Web）、新闻（News）、mp3、图片（Picture）。此外它还允许用户从系统挂接的 12 个搜索引擎中选择，进行新一轮查询。ixquick 的查询结果如图 6 所示。

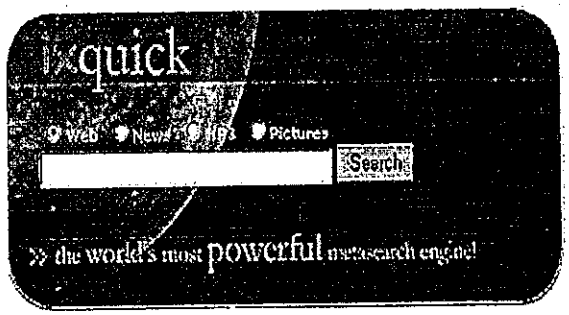


图 5　ixquick 的检索界面

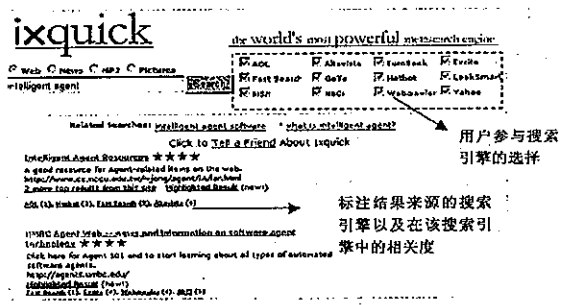


图 6　ixquick 的检索结果页面

（3）分散综合型元搜索引擎。这种类型与混合综合型元搜索引擎在结果显示上有所不同，它以独立搜索引擎为单位进行结果显示，在同一个独立搜

索引擎中查询得到的结果被集中列在该搜索引擎之下，例如 Dopile 的检索结果界面（图 7）。

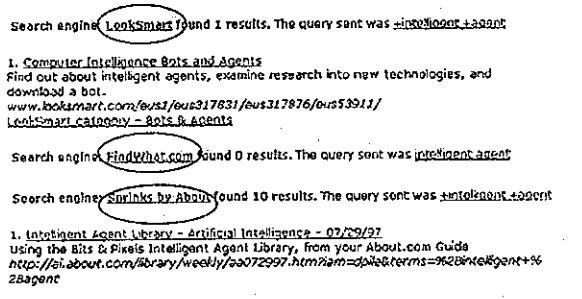


图 7 Dopile 的检索结果界面

## 4 元搜索引擎的主要技术

元搜索引擎的核心问题是解决如何调用其它搜索引擎的索引数据库、如何获取查询词在其他搜索引擎中的查询结果以及如何评价、排序、呈现结果等，解决这些问题的主要技术有用户提问转换、检索机制设计与优化、检索结果输出、分布式数据库调用等技术。

（1）用户提问转换。元搜索引擎定制查询界面供用户输入查询词，需要根据不同的搜索引擎转换成可以进行检索的查询表达式，不同的搜索引擎有不同的检索语法和操作符使用技巧，需要对提问进行处理。同时要对搜索引擎不能处理的检索方式进行排除，并选择一种合适方式来匹配。例如在 EZFIND 中，就提供了许多种检索句法，对于不支持的搜索引擎就不作处理，因此用户即便选定也不会起作用。

（2）检索机制设计与优化。对于搜索引擎初始化方式、各个搜索引擎结果平衡处理等问题，都需要在检索机制的设计初期进行规划，这主要受到检索反馈速度、检索结果满意度等因素的影响，目前，搜索引擎初始化主要有用户参与、系统默认或自动随机处理等方式。检索结果的处理例如如何衡量不同搜索引擎结果之间的相关程度。简单的处理方式是以搜索引擎为单位，在选定的独立搜索引擎下面显示比较靠前的结果（如 dogpile）；复杂的处理方式是以记录为单位，通过判定某一记录在多个独立搜索引擎中被评价的指数，如果多个独立搜索引擎都检出该结果，那么该记录将被排列在整个显示的前面，同时后面标注出是在哪些搜索引擎中检出的。

（3）检索结果输出。结果输出处理一般有两种形式：直接引用原始结果页面技术与结果页面定制技术。直接引用原始结果页面是元搜索技术中较为简单易行的方式，通过 CGI 技术，利用表单提交来调用数据库，在自制的页面中将表单提交的对象修改为独立搜索引擎调用数据库的脚本文件。在这种情况中，一般无须进行结果去重，只需完成表单提交的转换即可。结果页面定制技术则要对结果进行更多的加工处理，主要包括：①记录遴选处理，选择相关程度高的记录并加以显示，同时删除可能被多个独立搜索引擎同时检出的记录。②结果的再度排序，元搜索引擎并不是完全选取在独立搜索引擎中得到的结果的前几条记录，经过了再度排序，尤其在混合综合型的结果输出当中，根据检索机制对相关度判断的标准来比较各个搜索引擎得到的结果。

（4）分布式数据库调用技术。直接调用独立搜索引擎结果页面的元搜索引擎不需与独立搜索引擎的索引数据库直接交换数据，只需直接引用独立搜索引擎的结果页面；而查询多个搜索引擎的元搜索引擎则需在满足独立搜索引擎的数据库访问权限的情况下，实现对其索引数据库的访问和二次开发。各独立搜索引擎的数据库分布在不同的地域，要实现对异地、异构数据库的访问，需要使用一系列诸如分布对象技术等相关的核心技术。同时，不同数据库调用结果响应时间长短不一，这也会直接影响到结果页面的呈现。

## 5 中文元搜索引擎发展的建议

在中文搜索引擎领域中，元搜索引擎尚不成熟，以元搜索形式出现的搜索引擎站点数量甚微，在中文元搜索引擎的开发与利用上还有待加强。笔者认为，以下几个方面应引起重视：

（1）充分利用现有资源，促进各个搜索引擎服务提供商之间的合作。中文搜索引擎目前仍处于独立发展的状态，然而众多搜索引擎的数据库在范围覆盖上存在着重复和差异，因此互相引见和参见可以实现功能更强大的检索服务。例如，在 Yahoo! 的每个检索界面还同时提供了检索 Google、Ask Jeeves 等的检索入口，这是值得借鉴的。

（2）元搜索引擎的开发应能起到对独立搜索引擎的推荐作用，同时应注重对结果的二次加工质量。目前，元搜索引擎正由初期的直接引用原始结果页面向综合评价反馈型发展，这一点已成为元搜索引擎区别于独立搜索引擎的一个重要特色。

（3）元搜索引擎要获得好的资源检索效果，应当尽量收录一些大型、权威、主流、覆盖面大的独立搜索引擎，从更深广的角度来覆盖网络信息资源，以便从多方面、多角度的为用户的查询提供更多的线索和结果。

（4）元搜索引擎的综合处理结果直接影响用户满意度，是反映元搜索引擎质量高低的一个标准。在对多个搜索引擎的结果进行综合评价时，要能够较为客观的衡量各个不同记录之间的相关度，在不降低检索质量的同时提高检索效率，更好的满足用户的需求。因此，设计出优越的检索机制来综合处理各独立搜索引擎的结果是一个关键环节。

（5）中文元搜索软件的开发也是发展元搜索引擎的一个方面，国内在这方面已经有过一些成型的产品，虽然在界面和功能上不及国外的成熟软件，但是毕竟有了起步。另外中文信息贫乏本身也是一个不容忽视的客观问题，不是纯粹靠技术力量可以弥补的。

在国内中文检索工具中，利用元搜索引擎技术的尚不多见。元搜索引擎可以对独立搜索引擎在检索范围上的局限性作出一定程度的弥补和改善。对用户而言，元搜索技术可以使用户提交的检索请求一次性递交给多个独立的搜索引擎进行查找，获得多个独立搜索引擎的检索的结果，节省时间，提高效率。目前，国外非常重视元搜索引擎的研究和开发，因此，我们应对这一领域有充分的重视。

## 参考文献


1 Nicholas Tomaiuolo. Are metasearchers better searches?. Searcher. 1999 (1): 30～34

2 王　芳，张晓林. 元搜索引擎. 现代图书情报技术. 1998 (6): 18～21

3 张　蕊. 元搜索引擎揭密. 中国计算机报，2000 (27) B1，B3

4 Meta－Search. availble from URL http://www.lib.berkeley.edu/TeachingLIB/Guides/Internet/MetaSearch.html

5 http://www.searchenginewatch/Major Search Engines/index.html

6 http://www.ixquick.com

7 http://www.infodump.com/index.htm

8 http://info.theriver.com/TheRiver/ezfind.htm


（责任编辑：徐　波）

（上接第 170 页）网络的光盘数据库信息服务模式。在互有信息交流的信息机构之间建立一定的联系，借助电话、传真和因特网上的先进传输工具共享各自拥有的光盘数据库。既可资源互补，又避开基于广域网的光盘数据库信息服务中的版权之争。

（4）提供网络在线服务，充分利用光盘的内建 Internet 联机功能。利用光盘软件中内建 Internet 联机功能，可以直接从 Internet 上获取最新的信息。例如，几乎所有 1996 年以后版本的光盘百科全书都提供了这个功能，用户可以每个月从 Internet 上获取最新的时事信息，弥补光盘软件时效的不足。

（5）提供应用软件服务。应用软件如：用于图像处理的 Photoshop、Paintshop pro6；用于同页制作的 Flash、Dreamweaver、Firework、Frontpage；用于制作屏保的 Screensavershot；用于屏幕抓图的 Hypersnap－DX；用于制作效果字的 Cool3d；用于文字处理的 Access、Word、Powerpoint、Excel、Ultraedit；用于浏览文字和图像的 Adobe Acrobat、ACDsee；用于制作动画的 GIF；常用工具软件 Winzip、Cutftp 等。在电子阅览室提供这种服务，可以为没有计算机的学生提供更多的学习机会和学习条件。

## 参考文献


1 Jaap Jasperse，“Are We Losing a Generation of Manuscripts?” Online and CD－ROM Review21：1 (Jan. 1997)：30

2 Peter Kibby and Joseph Franzino，“CD－ROM and Multimedia State of the Art 1997”，Computers in Libraries17 (May 1997)：65～67

3 周　晨. 浅议馆藏光盘资料的著录. 河北科技图苑，2000 (3)：48～50

4 李　静. 谈提高光盘数据库的利用率. 图书馆，1999 (2)：55～57

5 袁勤俭，等. 谈光盘数据库信息服务模式. 中国图书馆学报，1999 (4)：66～70

6 刘　耕. CD－ROM 光盘检索的广域联网. 现代图书情报技术，1997 (2)：24～26

7 北京大学图书馆光盘联合目录. http://www.lib.pku.edu.cn/is/cdnew.htm

8 http://www.lib.tsinghua.edu.cn/
http://lib.nju.edu.cn/
http://www.lib.tju.edu.cn/
http://www.lib.sjtu.edu.cn/
http://210.32.205.34/


（责任编辑：徐　波）