文章编号：1004-7271(2000)03-0235-05

中文水产搜索引擎的研究与探索

卢卫平，吴维宁，林 成

（上海水产大学网络管理中心，上海 200090）

摘 要：搜索引擎是人们检索Web信息资源的有效途径。专业搜索引擎能更好地满足专业信息检索的需要。水产搜索引擎普遍存在数据库规模较小但简洁实用等特点。大多数以分类集合网站网址为主，采用主题目录索引方式实现搜索。“猎渔搜索”是“中国水产网”开发的中文水产搜索引擎，其主题目录的编排采用主题分类法，共分为16个类目，并选择NT+IIS+SQL Server作为开发平台，由后台索引数据库和前台搜索界面组成。

关键词：中文搜索引擎；因特网；信息检索；水产

中图分类号：S911 **文献标识码**：A

The research and probe of Chinese fisheries search engine

LU Wei-ping, WU Wei-ning, LIN Cheng

(*Network Management Center, SFU, Shanghai 200090, China*)

Abstract: Search Engine is a most useful method for searching infomation resources in Internet. A Subject-specific Search Engines can meet the need of some customizing professional searching. Fisheries search engine has some common characters those are small database, but convenient and fisheries oriented. Most of them collected related websites, and then searched them with subject directory index. “LieYu Search Engine” is a Chinese fisheries search engine developed by “China-Fishery.net”. The subject directory is classified by 16 different subject catalogs. The database interface is NT + IIS + SQL Server and contains two parts which are background index database and foreground searching interface.

Key words: Chinese search engine; Internet; information search; fisheries

搜索引擎(Search Engine)为人们检索Web信息资源提供了一条有效的途径。据中国互联网络信息中心(CNNIC)1999年6月的调查，69%的网络用户是通过搜索引擎查找并登陆到目标站点的，在得知新网站的途径中占第一位。65.5%的用户选择搜索引擎作为最常使用的网络服务之一，仅次于电子邮箱的使用，其重要性由此可见一斑[1]。本文结合互联网上水产信息分布和水产搜索引擎特点等研究，着重探讨中文水产搜索引擎的开发。

1 水产搜索引擎的发展现状与特点

1.1 搜索引擎的含义及发展趋势

搜索引擎又称检索引擎，是指运行于因特网(Internet)上，以Internet信息资源为对象，以信息检索的方式为用户提供所需信息的数据库服务系统。它包括信息收集、信息过滤、信息存取、信息索引、信息检

收稿日期：2000-05-10

作者简介：卢卫平(1959-)，男，浙江磐安人，讲师，双学士学位，现从事互联网站建设和电子商务研究。

索等组件[2]。搜索引擎有多种类形,但 Internet 上数量最为庞大,内容最为丰富的信息源是 WWW,75% 以上的 Internet 用户是从WWW 上获取信息的。本文讨论的搜索引擎即指 WWW 型搜索引擎。

中文搜索引擎是指 Internet 上提供中文检索导航等服务的搜索引擎,其关键组件是能够在大量中英文数据上进行高效全文检索的信息管理系统[3]。它包括两方面的含义:一是要能够检索中文信息资源;二是提供检索结果的中文化导航服务。

1.1.1 搜索引擎的基本工作原理

搜索引擎可以作为一个 Web 网站,也可以是网站中的部分功能,其运作的基本原理是通过人工或搜索软件在 Internet 上主动搜索 Web 服务器信息并将其索引,其索引内容经过过滤、分类整理后,存储于可供查询的索引数据库中。当用户输入关键词或通过主题目录查询时,该引擎会告诉用户包含该关键词信息或该类目下的所有网址,并提供通向该网站的链接。

我们通常所说的搜索引擎一般包括关键词索引和主题目录索引两大部分,分别属于不同性质的两种系统,其工作原理也有一定的差别[4-6]。如关键词索引数据库中的网页资料绝大多数由机器人自动生成。而主题目录索引数据库一般由专业人员对互联网上的信息进行搜集、筛选、组织和评论,编制等级式的主题指南或主题目录,以供检索和查询。现在很多的搜索引擎都是将关键词索引和主题目录索引结合起来,以充分发挥两者的优势[7]。并且随着技术的进步,无论是关键词索引还是主题目录索引都逐步以机器人搜索、智能分类和人工筛选相结合的方式完成搜索队列的编排,以追求查全率和查准率的和谐统一。

1.1.2 搜索引擎的发展趋势

搜索引擎历史不长,但发展迅速,技术逐趋成熟,近来呈现出以下趋势[8,9]:

一是从综合性向专业化方向发展。随着 Internet 上网站和网页数量的快速增长,综合性搜索引擎的搜索功能越来越受到局限。据研究,目前最著名的搜索引擎只能检索到互联网上不到20%的网页[10]。由于涉及的学科面很广及受专业知识限制,引擎编辑很难顾及到某一专题,其目录的分类编排也无法科学合理,不能满足专业人员的需要。而专业性搜索引擎就不同,它不求各门学科信息最多,但求本学科、本专业最全。目前基本上处于刚起步阶段,发展余地较大。

二是从单一性向多元化方向发展。独立的搜索引擎(Single Search Engine),只能在自己搜集的信息或者数据库中查找用户所需的资料和信息,导致查询范围有限。多元搜索引擎(Meta Search Engine)提供单一搜索引擎之上的二次检索技术,有利于充分利用现有的独立搜索引擎资源,取长补短。

三是从以英文为主向多语言化方向发展。目前功能强大、信息齐全、查询快捷的优秀搜索引擎基本上都是英文的,这和 Internet 上流通的信息资源近 90%为英文是相适应的[11]。但随着各国 Internet 的迅猛发展,以各自母语为载体的 Web 信息也迅速增长,相应地,支持各自语言 Web 信息查询技术的搜索引擎也应运而生。近来,中文搜索引擎的快速发展就是一个很好的佐证。

1.2 水产信息在 Internet 上的分布

Internet 上水产信息的数量和种类都很多。首先,相对集中地分布在水产专业网站上。日本渔业信息服务(Fish Info Service)(http://www.fis-net.com)搜集了 3000 多家水产相关网站,相信这还只是其中的一小部分。尽管在浩如烟海的 Internet 资源里,水产网站只是"沧海一粟",但它们信息集中,专业性强,服务有针对性,是网络水产信息资源的主力军;其次,在许多农业及相关网站上都设有水产栏目,提供较多的水产信息。因为水产本身就属于大农业范畴,综合类的农业网站少不了水产的位置,如中国农业信息网(http://www.agri.gov.cn)上就有水产市场子网站,但相对于其他农业主流信息,水产信息就显得较为单薄;另外,还有一些综合类和商务类网站,也提供一些特定的水产信息,如阿里巴巴(http://www.china.alibaba.com)网站中就有水产、渔业设备及用具等分类栏目。尽管水产信息在这些网站中所占的份量很小,但由于此类网站数量相当庞大,其提供的水产信息资源还是相当可观的。

Internet 上水产信息基本上以英文为主,这和美国及欧洲国家 Internet 最为普及,Internet 流行英语表达是相适应的。此外,日文水产信息也较丰富,各国基本上都有支持各自母语的水产网站。近年来,中

文水产信息发展迅速，除了台湾、港澳及东南亚地区有众多的以 BIG5 码支持的繁体中文水产网站外，我国以 GB 码支持的简体中文水产网站也在不到 1 年的时间里，从 20 来家发展到 100 多家。

1.3 水产搜索引擎的特点

水产搜索引擎普遍都存在数据库规模较小，功能并不强大但简洁实用，针对性强等特点。大多数水产搜索引擎以分类集合网站网址（URL）为主，采用主题目录索引方式实现搜索，不支持全文检索技术，如水产发现（The Fish Finder）搜索引擎（http://www.thefishfinder.com），只要经引擎编辑验证了网址的正确性和网站内容的相关性后，就可登录到它的数据库中。有的只涉及到水产的某一专题，如养殖循环系统和水质搜索引擎（Aquaculture Recirculation Systems and Water Quality）（http://www.drydenaqua.com/Ring_Aqua/webring_aqua.htm），收集的网址仅限于养殖循环系统和水质处理等内容。有的只提供本国或本地区的渔业信息检索，如挪威 Havbruk Magazine 网站（http://www.havbruk.no）的搜索引擎，可查找挪威相关主题的内容。

少数水产搜索引擎同时支持关键词检索和主题目录检索，如水产搜索（Aquasearch）（http://www.aquasearch.net），可以检索互联网上有关水产生产、销售、经营以及管理、教育、科研等方面的网站；日本渔业信息服务（Fish Info Service），同时支持英文、日文、西班牙文检索，关键词检索只提供站内资料检索；国际养殖搜索引擎（International Aquaculture）（http://members.magnet.at/aquaculture/Webring.htm），以网站环（Webring）的形式汇集了世界上有影响的水产相关网站，其主题目录编排很有特色，分别按主题（Topics）、品种（Species）、区域（Continents）三条主线划分为 16 个类目。

中文水产搜索引擎在我国发展相对滞后，目前只有“中国渔业信息网”（http://www.fish.net.cn）开发了一个关键词搜索引擎——“渔网搜索引擎”，检索的信息只限于该网站内及国内主要渔业网站的中文渔业信息，不能检索国外网站的相关信息。

2 中文水产搜索引擎的开发

Internet 上的水产信息资源和水产搜索引擎大多为英文，对广大中文用户存在一定的语言屏障。一方面，水产网站和水产信息数量急剧增加，以水产搜索（Aquasearch）为例，仅 2000 年 3 月 24 日到 4 月 5 日新登记网站数为 159 个，平均每天增加 10.5 个[http://www.aquasearch.net/]。另一方面，提供专业网络导航的中文水产搜索引擎几乎是空白，给中文用户查找水产信息带来很大困难。因此，开发能快速、准确、全面及中文化提供互联网上水产信息导航的专业搜索引擎显得尤其重要。以下结合“中国水产网”（http://www.china-fishery.net）的“猎渔搜索”（LieYu Search Engine），讨论中文水产搜索引擎的开发。

2.1 “猎渔搜索”的开发思路与特点

设想中的中文水产搜索引擎应该具备以下功能：①同时提供主题目录索引和关键词索引功能，充分发挥机器人自动搜索快捷和人工分类编目准确的优势；②支持中英文网页全文检索技术，并全中文输出查询结果；③构建科学合理的水产主题目录树结构，提供不同的分类路径入口；④揉合多元搜索技术，充分利用现有独立搜索引擎的索引数据库，经二次检索输出，满足专业需要。

“猎渔搜索”（LieYu Search Engine）是“中国水产网”为满足水产从业人员和兴趣爱好者检索 Internet 上水产及相关网站信息需要而自行开发的中文水产搜索引擎，采用“综合设计，分段调试，逐步推出”的策略开发，现已推出基于人工的水产及相关网站的主题目录检索，并将逐步开发基于机器人自动搜寻并支持全文检索的关键词检索。该引擎现具有以下几个特点：①全中文输出界面，适合以中文为母语的华人使用；②人工搜集编排，可检索国内外水产及相关网站；③主题目录树状结构，类目设计简洁明了；④同时支持基于网站索引数据库的中、英文关键词检索；关键词检索支持精确查询和模糊查询；⑤在线前台自由登录，后台审核编辑，确保登录网站信息的质量。

我们认为，开发中文水产搜索引擎的难点，并不在于网络搜索技术，许多综合性搜索引擎的核心技术和搜索软件足够我们借鉴使用，而在于水产信息资源的甄选编排、分类整理，在于主题类目的确定和

关键词的抽取;在于及时跟踪反映网页资源变化,减少无效重复链接。

2.2 主题目录的编排

主题目录编排的好坏是水产搜索引擎质量高低的关键所在。目录编排应遵循的基本原则是:①基本符合学科分类原则,符合专业人员的使用习惯,方便查找;②提供多途径入口,醒目易懂,非专业人员可以通过不同入口方便地找到所需信息;③可扩充,易修改,能适应信息资源拓展的需要。

主题目录的编排有多种方式,如学科分类法、分面分类法、体系分类法等。"猎渔搜索"采用的是主题分类法。主题分类法的特征是一个主题充当一个类目,像主题词表一样按字顺或重要程度排列,而不是以逻辑顺序排列。一个类目又可分为若干细目,同位类的细目也是按字顺或重要程度排列。网络资源数不胜数且高速增长,任何分类目录都不可能包罗所有的网页,因此,只能选取一些热点事物作为主题类目。主题分类法一般设置 14 至 18 个一级主题类目,层次不超过 4 级。最末一级就是列成表的超文本链接点,每个链接点伴有对网页内容的简要介绍[12]。

"猎渔搜索"在对主题目录进行编排时,参考了《水产科学叙词表》中的"范畴索引"[13]、《水科学与渔业情报系统数据库标引与用户指南》[14],以及互联网上相关搜索引擎如 Aquasearch, International Aquaculture Webring, Commercial Seafood Webring 等的目录编排,把水产主题目录分为:政府机构、协会组织、教育科研、图书馆资料库出版社、公司企业、仪器设备、养殖、捕捞、加工、渔业资源、渔业经济、饲料药品添加剂、水产品、因特网、休闲娱乐、其它等 16 个大类。每个类目下,又区分不同情况,细分为不同的子类目。

2.3 数据库的开发

主题目录搜索引擎的数据库开发,有多种解决方案可供选择。但专业搜索引擎应以简单、实用、可扩展为主要考虑因素,数据库结构简单,算法也相对简单,查询速度有一定保障,对于如今的网络速度与带宽有现实意义。"猎渔搜索"选择 NT + IIS + SQL Server 作为开发平台,是基于"中国水产网"目前采用 NT 服务器作为系统平台,而 SQL Server 是运行在 NT 上最流行的关系型数据库管理系统,它对简单商用数据的处理具有灵活、安全、可扩展、可管理、事务处理支持、易开发等特点。

"猎渔搜索"由后台索引数据库和前台搜索界面组成。索引数据库由 SQL Server 提供支持,主要包括网站索引表和网站登录表等部分。网站索引表是其关键部件,主要存放通过不同方式收集整理归类好的网站信息数据,包括如下 10 个信息字段:网站中文名称、网站英文名称、网站地址、关键词、网站简介、网站内码或语言、网站评价、登录时间、网站类型(所属目录)、摘要。目前主要采用人工收集整理,今后可发展为采用机器人来完成。网站登录表临时存放由上网者自由登录的网站信息数据,结构同网站索引表,登录信息经后台编辑审核确认后,导入网站索引表供检索用。

搜索界面包括 Web 查询页面和搜索程序。Web 查询页面提供友好人机对话窗口;搜索程序采用 NT 上的活动服务器脚本(Active Server Page)编写,连接后台数据库进行查询,动态输出客户查询的结果,支持精确查询,模糊查询,并同时支持中英文查询。

3 结束语

建立一个优秀的中文搜索引擎是一项庞大而复杂的系统工程,在技术、设备、人力、财力的投入上都有比较高的要求。如"新浪搜索"的开发,仅人员就投入了四、五十人,对大量的网络资源进行人工分类[15]。专业搜索引擎虽然没有那么庞大复杂,但也毫不轻松。目前"猎渔搜索"的开发还只是初步的,这是因为其主题目录的信息内容还有待于不断扩充,信息的甄别更新机制还需要继续建立完善,基于机器人自动搜寻并支持全文检索的关键词检索功能尚未开发完成,但在网上推出后,已经得到了用户的积极响应,不少网站主动在线登录信息。我们相信,对中文水产搜索引擎的研究探索将是积极有意义的,有助于水产从业人员更有效地利用 Internet 上的水产信息资源,有助于水产网站更好地推介宣传,也有助于推动中文水产搜索引擎更快的发展。

参考文献：

[1] 王　鹏，晓　艳．搜遍神州——中文搜索引擎大比拼[J]．互联网周刊，1999，(34)：10－11．

[2] 王　忠，周士波．Internet英文搜索引擎评析[J]．图书情报工作，1999，(4)：33－37．

[3] 都云程，卢献华．中文搜索引擎现状与展望[J]．中文信息学报，1998，13(3)：61－64．

[4] 宁　静．Internet检索工具略述[J]．图书馆论坛，1999，(3)：23－27．

[5] 任瑞娟，李洪建．中文WWW搜索引擎比较研究[J]．大学图书馆学报，1999，16(5)：55－57，61．

[6] 张　颖，周志农．因特网三大检索工具的比较研究[J]．图书情报工作，1999，(10)：39－42，58．

[7] 强自力．网络分类目录及其分类法[J]．大学图书馆学报，1999，16(4)：37－39．

[8] 黄建年．网络搜索工具的发展趋向[J]．图书情报工作，2000，(2)：34－36．

[9] 李名智．中文搜索引擎：现状、问题及对策[J]．大学图书馆学报，1999，16(6)：44－45．

[10] Steve Lawrence，C Lee Giles．Accessibility and Distribution of Information on the Web[J]．Nature，1999，400(6740)：107－109．

[11] 朱　慧，李瑞勤．从Infoseek展望WWW搜索引擎的未来发展趋势[J]．图书情报工作，1999，(12)：42－45．

[12] 陈笑辉，范晓虹．Yahoo的分类体系结构及原理探微[J]．图书情报工作，1999，(9)：33－36，59．

[13] 中国水产科学研究院科技情报研究所．水产科学叙词表[M]．北京：中国农业科技出版社，1991．155－192．

[14] 盖明举．水科学与渔业情报系统数据库标引与用户指南[M]．海洋出版社，1990．14－39．

[15] 陈丽英，严援朝．以简单沟通与世界互动[J]．互联网世界，1999，(7)：4－5．