键字到网页 URL（统一资源定位器）的关系索引倒排文档，即建立索引数据库。检索器的功能是根据用户输入的查询词，在索引数据库中进行查询词与索引数据库的匹配算法，然后将查询结果按相关程度排序并输出到浏览器上。

除了考虑核心的技术以及采用高效的算法外，必须在用户体验上下功夫如结构化的显示搜索到的结果。比如，Google 所使用的 Ajax（异步 JavaScript）技术，用户输入查询时能够自动提示，还有 Google 查询后显示的数据，界面上字体等要比 Baidu 细致一个档次。这些细节的原因，某种程度关系到搜索引擎在市场的占有额。

四、垂直搜索引擎的相关技术

1. 页面解析与页面显示排序。

网页地址都是用 URL（Uniform Resource Locator 统一资源定位器）来表示，获取网页信息，必须找到 URL，读取该 URL 页面的 HTML、特定标签，高级的搜索引擎还能对 JavaScript 语句进行解析。这是因为许多网站直接用 JavaScript 构建出来，而且随着 Ajax 技术的流行，很多信息包含在 JavaScript 标签里，为了提高采集信息的准确率，提高搜索引擎的竞争力，搜索引擎必须提供 JavaScript 解析器。

页面排序是针对根据用户关键字，查询到的网页列表，采用何种策略将网页列表显示在用户面前，使用户最想知道的结果显示在最前面页数发生的概率最大。主要的算法有：PageRank 算法、HITS 算法。在排序上，有些搜索引擎（如百度），则加入收费这一方式，使排序成为搜索引擎的一大盈利模式。

2. 数据存储及分布式技术。

尽管垂直搜索引擎保存的网页数量相对通用垂直搜索引擎小很多，但是，作为一个优秀的商业垂直搜索引擎，必须在提高性能的同时减低成本，提升竞争力。可以采用数据压缩的技术对数据进行压缩存储，采用数据库技术，如索引等提高数据读取速度，也可以采用分布式技术，通过多台服务器相互合作，以提高数据采集和更新速度。

3. 网络蜘蛛的爬行策略。

网络蜘蛛（Robot 或 Spider）的搜索策略是指当网络蜘蛛搜索到一个文档之后，下一步应该转移到哪一个文档的方法问题。目前比较常见的搜索策略有以下几种策略：（1）IP 地址搜索策略；（2） 深度优先搜索策略；（3）广度优先搜索策略；（4）深度－广度结合搜索策略。

4. 中文分词技术。

在 Web 应用中，文本处理的速度往往是性能的关键，快速分词具有很大的现实意义。Web 文本分词是 Web 信息处理的基础，如信息检索、摘要形成、网页过滤等都需要对 Web 文本进行分词处理。Web 文本的正文主要由英文和中文构成，由于英文的单词与单词之间有空格，所以不存在分词问题。而中文的每一句中词与词之间是没有空格的，因而必须采用某种技术将其分开。

分词的方法很多，基本上分为两类：第一类是基于字符串的匹配，将汉字串与一个机器词典中的词条进行匹配，若在词典中找到某个字符串，则匹配成功。主要有正向最大匹配法、逆向最大匹配法、最少切分等方法。第二类是基于统计的方法，从概率角度出发，单字出现在词汇中联合概率是比较大的，因此当相邻的字越常出现，则越有可能是一个词。基于上述引，对处理的材料进行分析，得到相应的单字出现的概率，然后对相邻的字出现概率进行统计，若远大于单字出现的概率之和，则可能成为一个词。实际应用中，统计分词方法都是与字典结合着来使用的，这样既发挥匹配分词的切分速度快、效率高的特点，对利用了无词典结合上下文识别生词，并能消除歧义等优点。

五、垂直搜索引擎的发展空间

“确解用户之意，切返用户之需”“用户无法描述道他要找什么，除非让他看到想找的东西”，这是消费者（网络使用者）对搜索引擎提出的更高要求。以尽可能多地收集到与专业相关的信息为主要目标的垂直搜索引擎，比通用垂直搜索引擎在 Internet 上更加贴切消费者的要求。专业化的集中特定领域的垂直搜索引擎有效地弥补了综合性搜索引擎对专门领域及特定主题信息覆盖率过低的问题。市场需求的多元化，决定着搜索引擎服务的多元化；通用搜索引擎开拓市场上的成功，为垂直搜索引擎的市场战略提供了宝贵的借鉴经验，垂直搜索引擎的特点，决定着它在 Internet 上占有一席之地，必将成为搜索行业的一大力量。

（作者单位：广州市机电高级技工学校）

参考文献：

[1]王晓伟．垂直搜索引擎若干关键技术的研究[J]．浙江大学学报，2007，(5)．

[2]孙卫喜．搜索引擎分析[J]．高校实验室工作研究，2007，(3)．

[3]李副铭．垂直搜索引擎的研究与设计[D]．电子科技大学学报，2009，(9)．

[4]刘世涛．简析搜索引擎中网络爬虫的搜索策略[J]．阜阳师范学院学报，2006，(9)．

[5]邹海山，吴勇，吴月珠，陈阵．中文搜索引擎中的中文信息处理技术[J]．计算机应用研究，2000，(12)．

责任编辑　朱守锂