高等学校“十一五”精品规划教材

信息检索

主　编　程　娟
副主编　裴　雷　黄　强

中国水利水电出版社
www.waterpub.com.cn

高等学校“十一五”精品规划教材

信息检索

主　编　程　娟
副主编　裴　雷　黄　强

中国水利水电出版社
www.waterpub.com.cn

内 容 提 要

当今社会既是一个信息时代，又是一个网络时代，如何在纷繁复杂的信息资源中查找到最适合的信息是本书出版的目的。本书从信息的概念和信息的类型，通过大量信息检索示例，由浅入深地引导学生掌握信息检索这门课程。

本书共分为9章，内容包括信息素质教育、走进图书馆、信息检索概述、数据和事实检索与利用、专利文献信息检索、最新知识资源的检索与利用、数据库园地、网络免费信息的获取、看不见的网络及其检索利用。同时还在每一章的结尾设置了“知识要求”、“关键术语”、“本章小结”、“复习与思考”，为学生学习信息检索、掌握信息检索技能提供了较好的学习方法。

图书在版编目（CIP）数据

信息检索／程娟主编．—北京：中国水利水电出版社，2009

高等学校“十一五”精品规划教材

ISBN 978-7-5084-6344-5

Ⅰ．信… Ⅱ．程… Ⅲ．情报检索-高等学校-教材 Ⅳ．G252.7

中国版本图书馆CIP数据核字（2009）第032676号

书　　名	高等学校“十一五”精品规划教材 **信息检索**
作　　者	主编　程娟　副主编　裴雷　黄强
出版发行	中国水利水电出版社（北京市三里河路6号　100044） 网址：www.waterpub.com.cn E-mail：sales@waterpub.com.cn 电话：（010）63202266（总机）、68367658（营销中心）
经　　售	北京科水图书销售中心（零售） 电话：（010）88383994、63202643 全国各地新华书店和相关出版物销售网点
排　　版	中国水利水电出版社微机排版中心
印　　刷	北京市兴怀印刷厂
规　　格	184mm×260mm　16开本　17印张　403千字
版　　次	2009年4月第1版　2009年4月第1次印刷
印　　数	0001—3000册
定　　价	**35.00**元

凡购买我社图书，如有缺页、倒页、脱页的，本社营销中心负责调换

前　言

这是一本不同于以往的信息检索教材，旨在从大学生走向社会的需求出发，其互动性的交流、信息素养、信息检索技能等方面，均体现了现代信息检索理论与应用的全新概念和现代信息检索技术的有益尝试。

本教材的编撰来自一个团队，这个团队的组成人员既有在校的信息管理专业的博士研究生，又有本领域中的权威专家和学者。他们学习最前沿、最新信息检索技术的同时，也在为本专业的学生讲授信息检索课程；还有多年从事应用型大学文献检索教学实践、研究和应用的情报工作者。他们将所学的知识、教学的体验融合在一起，并不断交换教师与学生的角色，探讨最佳传授方式与效果。

本教材特点：集实用性、新颖性、通用性为一体。全书共分九个章节。导论部分由程娟编写；第一、二、八章由裴雷编写；第三章由沈红兵编写；第四、五、六章由杨妙君、程娟、黄强编写；第七章由卢炎香编写；第九章由肖雪编写。张娟与谭辉军参与了本教材的文字整理工作。全书的筹划、大纲设置、编写组织和统稿工作，由程娟、裴雷、黄强负责。

本书得到了武汉大学信息管理学院马费成教授的大力支持，在此表示真诚的谢意。

由于时间仓促及水平有限，对可能出现的缺点乃至错误，敬请广大读者批评指正。

编者

2009 年 3 月

目　　录

导　论

工欲善其事，必先利其器。

——《论语·魏灵公》

当你准备学习信息检索技能的时候，你首先可能会问："信息检索技能是什么?"大部分的朋友可能会第一时间微笑着回答："好像我并没有用专门的检索技巧，也能找到自己需要的信息。"确实，现在的信息检索工具越来越人性化，使用也越来越方便，检索自己所需的信息变得比较容易了。首先你是否知道，你可能已经下意识地使用了一定的检索技巧呢？其次，对于特殊的检索需求，比如需要查找"空调压缩制冷技术"的时候，你是否依然能够迅速找到你所需的信息，或者你检索到的信息是否准确呢？第三，在你查找到的大量相关信息中如何筛选质量更好、相关度更高的信息？如果你觉得一下子不能回答所有的问题，那你就是本书的"目标读者"——本书将向你介绍什么是信息检索技能，你为什么需要掌握检索技能，以及你需要掌握什么样的检索技能。

确实，检索技能在我们的生活中几乎无处不在。当你在学习的时候，你需要查找自己感兴趣的相关知识和信息；当你在工作的时候，你也会查找与你的工作技能相关、竞争对手相关以及工作对象相关的"专业知识"；即使在你的生活休闲中，你可能也会关注与旅游、法规、市场价格等与生活密切相关的信息和知识……总而言之，当你面临创新和决策的所有过程时，你都需要获取信息；而这一信息获取过程除了少量自己原有的知识积累和经验总结，大量的信息需要从你所处的知识环境中获取——而这种从知识环境中准确描述和获取自己所需知识和信息的过程就是信息检索过程，而准确表达信息需求，获取相关信息的能力就是信息检索技能。

在实际的检索操作过程中，运用不同的表述视角，信息检索技能也有不同的描述。关于信息检索，目前国内外的专家有这样几种代表性的观点：

（1）以美国数学家 Calvin N. Moors 为代表的"时间通信论"，他们认为信息检索是一种通信方式，与传统通信把信息从一个地点传递到另一个地点不同，通过检索能把信息（文献、图片、数据等）从一个时间点传递到另一个时间点；

（2）"信息检索科学论"，他们认为信息检索是包括系统设计、概率论、算法优化、智能识别、统筹科学和分析科学相结合的复杂理论，同时也包括检索策略与服务应用研究的研究体系，这是信息科学专业的理解范式；

（3）"情报服务论"，他们认为信息检索就是从大量信息中查询与用户需求相适应的文献和知识的服务过程，这是专业信息咨询顾问的视角；

（4）"全息检索论"，我国学者王永成教授认为，不论你是否专业信息检索者，也不论

你采用何种途径，“可以从任意角度，从存储的多种形式的信息中高速准确地查找，并可以任意要求的信息形式和组织方式输出，也可仅输出人们所需要的一切相关信息的电脑活动”，这一观点认为用户从已有的信息需求出发，到满足信息需求的过程称为信息检索。

此外，还有学者对信息检索是专业技术，还是基本技能，持有争端。专业技术论者认为，信息检索广泛借助检索工具来实现，因而信息检索技能应该是对信息检索工具的全面掌握，掌握了计算机检索工具，就具备了良好的信息检索能力。但是，信息检索工具种类繁多、设计多样，计算机出现以前是书目卡片、索引、题录等，如果你们关注古代的地方志，必定有一卷是索引卷，章回、人名、地名和事件等一一在列，而《四库全书》也有相当篇幅的索引卷，当前国外也有《化学文摘》等类似工具书；计算机出现以后，数据库、搜索引擎、光盘系统、主题网站等各种形式相继出现，要全面掌握这些工具的应用并非一日之功，作为非信息专业的学生谈何容易？因而，信息检索是一门专业技术。

而基本技能论者认为，一方面，人们在生活中不可避免地存在信息需求，必须独自解决一些信息的搜索和获取问题，自身必须具备一定的检索能力；另一方面，信息检索工具，尤其是面向网络的信息检索工具越来越大众化，使普通用户掌握一些实用的信息检索工具成为可能。因而，信息检索是一项基本技能，或者称之为信息素质、信息素养(information literature)。

我们认为，信息检索技能不可孤立，它与你所处的信息环境和信息需求相关：如果你是专业科研工作者，或者是企业工程师，对课题或技术的相关信息必须精确获取，你可能必须接触专业的检索工具，而且你对检索结果需求严格，这时信息检索的技术成分高于素质成分；如果你是一个市场人员，需要了解你所推销区域的人口、政策、消费能力、购买欲望等，一来可能有些信息没有专门的信息资源；二来你可能必须在短时间内作出决策，你可能就需要大致搜集相关信息并得出结论，此时信息检索的技巧成分就高于技术成分了。可以简单地认为，以决策为目标的工作，以信息检索技巧技能为主；以创新为目标的工作，以信息检索技术掌握为主，也需要相应的检索技巧和策略。

同学们的大学学习，是人生中创新能力最丰富、也最需要创新能力的阶段，本阶段的信息检索学习应该既学习一定的检索技术，也学习相应的检索策略与技巧。所以在本书中，我们主要借助不同信息检索工具的特征以及不同类型的信息分布，向同学们介绍了文献信息资源、数据和事实信息资源、专利信息资源、计算机信息资源、网络信息资源等常用信息资源类型的检索工具和技巧。虽然信息资源多种多样，但是在信息的获取和利用过程中，有些思想是可以作为共同的参考、标准、准则、原理和常识。美国经济学家格里高利·曼昆认为，很多领域的专业知识都是可以用一些中心思想统一起来，构成简单的“公理系统”。而在我们信息检索领域，也有这样一些“中心思想”存在，就是我们在导论部分准备向大家介绍的“信息检索十大原理”。

可能刚开始的时候，你完全不理解，或者不认同这样的原理，你也不必担心这本书与你的期望相去甚远。在以后的各章中，我们将更加充分地展示这些原理的内涵以及对你的检索技能的需求，介绍这些原理只是希望你快速地了解信息检索技能的概况，寻找知识的启发点。总之，如果你关注创新和效率，本书就值得你关注。正如莫尔斯在描述信息检索工具时，提出了著名的“莫尔斯定律”：如果一个系统或工具的使用使得信息检索的效果

不如不使用这一系统和工具的话，这一系统或工具就不会被使用。同样，“如果你阅读和学习这本书，对你的检索能力没有任何帮助的话，请告诉其他检索者不要学习这本书了。”你可能对此不屑一顾，这确实是一句科学经典，因为有很多“科学”把简单的问题弄得非常复杂，而对效果没有改进；这是所有严肃科学研究的基本准则。同样，编者也是很严肃地承诺：希望通过信息检索技能的讲解能够对大家的检索过程和结果有所帮助，不要把大家认为的“自然存在”的检索技能复杂化而没有改进效果。

一、检索技能是现代社会必备的基础技能

也许还有很多同学认为，信息检索技能应该是图书情报或者信息科学专业的专业技能知识，作为一般专业的大学生，知道信息检索的基本工具和使用方法即可。

确实，“术业有专攻”，我们不可能把大家都培养成为“间谍”级别的信息获取高手，但是在当前的学习以及将来的工作中，大家也要意识到检索技能的重要。目前，大家应该意识到检索技能已经成为“信息素质”的重要组成部分，而信息素质则是大学生活中不可或缺的一个部分。美国教育技术 CEO 论坛 2001 年第 4 季度报告提出 21 世纪的能力素质包括的五个方面：基本学习技能（指读、写、算）、信息素质、创新思维能力、人际交往与合作精神。实践能力。而信息素质是其中的一个要素。

那么，什么是信息素质？1974 年，美国信息产业协会主席 Paul Zurkowski 给美国图书馆与信息科学委员会的报告中认为，信息素质是利用大量的信息工具及主要信息资源使问题得到解答的技能，在未来 10 年中信息素质将是国家发展的目标。围绕信息素质的讨论，人们对它的认识逐步深入，比较简明的阐述来自美国图书馆学会 ALA（American Library Association，1989），其内容包括：能够判断什么时候需要信息，并懂得如何去获取信息，如何去评价和有效利用所需要的信息。对大学生的信息素质要求，比较典型的来自美国高校和研究图书馆协会 CRAL 特别工作组，他们提出高等院校学生应具备的信息素质有六大指标：确定所需信息的范围、有效地获取所需的信息、鉴别信息及其来源、将检出的信息融入自己的知识基础、有效地利用信息去完成一个具体的任务。了解利用信息所涉及的经济、法律和社会问题，合理、合法地获取和利用信息。

所以，检索技能作为信息素质的一个组成部分，犹如学习能力一样，将与大家今后的学习生活朝夕相伴。一旦你忽视这一技能的存在，相对你的竞争对手而言，你在知识的获取环节就会落后，随之学习进度落后，进而竞争链条就会处处落后——所谓的个人综合素质的竞争，其实不仅是空间上多种素质的综合能力竞争，也包括时间上的竞争链条，如果起步阶段出错，会产生连锁反应。

而在所有的社会行业中，寻求的就是“发现和解决问题的能力”，我们现在所储备的专业技能、学习技能都是为将来发现和解决问题打基础。而这些问题的解决一方面是我们通过现在的学习能够将来独自解决的，但更多的问题却需要借助你所处的知识环境。向你所处的知识环境表达你的信息需求并获得结果，就是你所需要的检索技能，比如你可以向你周围的专家询问，可以通过专业文献的查找，可以通过网上的搜索引擎查找，可以通过网络的公共论坛（BBS）发布信息让热心的“网络知识志愿者”帮你解决，可以通过个人博客中具有共同兴趣的爱好者一同解决……总之，现在的信息资源分布相当广泛，专家个

人知识的揭示也具有一定的基础，如何获取和利用这些知识往往成为解决问题的关键。

概念：

文献：《文献情报术语国际标准（草案）》（ISO/DIS5127）认为，文献是“为了把人类知识传播开来和继承下去，人们用文字、图形、符号、声频、视频等手段将其记录下来，或写在纸上，或晒在蓝图上，或摄制在感光片上，或录到唱片上，或存储在磁盘上。这种附着在各种载体上的记录统称为文献。”

在以后的章节中，一些经验丰富的专家、老师会告诉大家，不同的问题所需要的支撑信息类型是不同的，而不同类型的信息的分布和获取存在较大差异；并不是像大家最初所认为的那样，“所有的信息检索都差不多”。因而，不管你是一个认为信息检索技能不屑一顾的“傲慢者”，还是一个认为信息检索都差不多的“偏见者”，希望都带着“检索技能是现代社会必备的基础技能”的“信息检索有用论”开始本书的学习。

二、检索技能可以创造财富

在众多的“数字英雄”中，杨致远和李彦宏的名字同学们可能并不陌生。分别作为google和baidu的创始人，他们创造了令人惊讶的财富规模，也创造了令人惊讶的创富速度！

那我们学习了检索技能就能遍身罗绮、家财万贯么？当然不是，检索技能创造的“财富”并不是这样的财富，这句话的真正内涵是：检索技能能够帮助你决策的速度和精度，速度意味节约时间，精度意味减少失败概率，这两者都是人生的“绝对财富”。举个简单例子，就业问题。将来大家就业的时候就是一个复杂的决策和选择问题，你面临无数企业的选择，哪个企业最适合你的发展？企业面临众多求职者的选择，哪个才是企业合适的选择？或者为什么你是企业的选择？这里面就涉及与你相关的两个“检索问题”了：第一，你要明确你的信息需求及其获取，行业、企业前景、职位、薪金范围、个人发展空间、地域限制等，然后怎样获取这些信息？第二，你要明确企业的“信息需求”，他们对求职者的能力、知识结构、健康、道德素质或者是职业素养如何定位，然后你在简历、面试环节给予企业明确的反馈。一份简历不能包打天下，正如一个“百度”不能解决你所有的信息需求，应该适当掌握一些检索工具和技巧，学会对专利、标准、法律法规、事实、科技文献等的检索和利用。

三、检索工具不是万能的

在信息检索领域，有两个相影相随的概念：查全率和查准率。查全率来形容检索工具或检索人员所检索到的相关信息与存在的全部相关信息的比值；而查准率是指检索工具或检索人员所检索到的相关信息与检索到的全部信息的比值。所以，在检索过程中，往往顾此失彼，查得全往往意味着不准，查得准却容易遗漏不全。信息检索过程就好比“挖花生”，挖出花生不可避免带出泥土，你拔出来可能全部是花生，泥土很少，但肯定有花生断在泥土里面了，比较浪费；如果你把从地表到地下50厘米的泥土全部铲起来，花生肯

定都在了，可是泥土太多，你再来掰泥土，也不适用。这也好比信息检索，很难二者兼顾。现在检索实验一般认为，查全率超过80%时，查准率不到10%；查全率在50%时，查准率大约25%；查准率50%时，查全率也只有25%左右。

你可能觉得，这个检索效率太“差劲”了！那岂不是要漏掉很多信息？所以说检索工具不是万能的，这是原因之一。因为现代检索理论认为，如果你不停地改变检索策略，改变检索入口词，改变检索系统，反复“覆盖”，是能够达到一定的检索效果的。当然，还有检索系统自身的问题，如果上面是在松软的沙地里拔花生，可能损失比较小；如果检索系统自身信息组织比较好，也能提高检索效果。

给大家介绍的这个“两难”问题，还只是众多检索问题中的一个，称之为“检索相关性问题”。还比如：

1. 信息的生命周期问题

信息是有生命的！知识毕竟是人们在一定时期、一定环境下对事物的认识，而信息可以认为是传递或准备传递的知识，它也具有这样的局限性，即信息的认知局限性。既然是认知过程的产物，那么认知可能正确，也可能错误；当正确的知识覆盖错误的知识时，错误知识的生命就“结束”了。著名科学哲学家波普的“知识进化模型”讲的就是这个道理，知识总是有发现提出、证伪、更新、再证伪……这样的循环上升过程。检索出来“错误”的信息，怎么办？显然，检索工具还不具备区分信息真伪的“超级智能”，它们仅仅能告知你信息的存在性，信息的有用性必须由检索者来区分！还有些信息本身具有实时性，控制指令、股票价格、新闻，虽然它们的有效时间不一样，“命不一样长”，但是都涉及到过时就无用的问题，可是很多检索工具没有办法过滤这些，需要用户识别！还有，信息总是在使用过程中显示其价值，于是一些信息学者用使用频次来表述信息的价值。美国学者普赖斯就曾经利用文献引用关系考察信息的时效性，提出普赖斯指数，即某一学科领域内，对发表年限不超过5年的文献的引用次数与总的引用次数之比值。统计结果表明，世界上93%～98%的科学杂志引用寿命为20年左右，也就是说1988年以前的文献在现在被引用率很低了。这说明什么问题呢？不是说这些信息资源不存在，而是没有人用了！大家想想，这部分信息是算在查全率里面的，但是对实际的应用并没有太大影响；反过来说，如果这部分信息纳入到查不全的那部分信息里面，如果适当改进（比如剔除这部分信息），是不是能够在保证查准率的同时，也提高查全率（分母变小了）！但可怕的是，生命周期它不是固定的，随着环境、技术的变化，它也发生变化；所以，现在很多检索专家想在这里“动刀子”，却不敢，只得“宁滥毋缺”！但是，检索工具没有办法的事情，用户可以轻易完成，比如检索工具设定一个时间范围，用户对所需信息资源的生命周期做到“心中有数”，大致的检索范围便一目了然，“多快好省”地获取信息！

2. 信息过载问题

信息过载，就是信息量超过人脑或系统的处理能力而出现信息臃肿、信息迷失，甚至是信息湮灭的问题。信息臃肿容易理解，你想知道“刘德华的父母”的相关情况，这条信息可能并不直接存在，需要从“刘德华”作为检索入口，点击，你会看到数以千万计的相关信息——哪一条才是你所需要的呢？God Knows！这也是信息迷失，因为数量多而迷失；还有一类是因为速度快而信息过载，其实也是单位时间内处理的信息量过大。比如一

个简单的公钥密码大概有 2^{100} 位，一秒钟之内从电脑读取显示，其速度肯定大大超过人脑的处理能力，即使是“过目不忘”的天才，他一秒也只能存取有限帧的信息，这样的信息照样“过目也忘”！信息湮灭，则是由于检索工具无法对信息来源的权威性考证，容易出现不一致的情况。比如，“珠穆朗玛峰究竟有多高?”有说 8848.13 米的，有说 8844.43 米的，回答问题，你反而不知道了，在有些是非回答问题上，“有信息反而不如没有信息”！这些都是检索工具无法解决的。

而且，目前新信息的增长是惊人的！据加州大学的一项研究成果反映，过去 20 年的信息总量已经大大超越了人类之前所有信息量的总和。检索工具总是难以覆盖最新、最全的信息，即使是公认最好的 google 搜索引擎所涵盖的信息总量也仅占全球网络信息总量的 5%左右。所以，检索工具不是万能的。所谓工具，无论多么复杂，只是辅助人们实现某些过程，替代部分劳动的产物。它既不是完整的解决方案，也不是无所不能的万能钥匙，检索需要用户的技巧和常识！但是，没有检索工具，有些检索也无法完成，比如访问用户从一个网站数以万计的网页中找到感兴趣的网页，没有检索工具，恐怕没有用户会一页页去浏览。除非，你本身就知道这张网页在哪个目录下，这就是我们要说的第四条：最佳的检索方式是记忆！

四、最佳的检索方式是记忆

所谓最佳的检索方式是记忆，同学们可以从两个角度看。第一种是不需要借助所处的知识环境，利用大脑储备的知识就能解决问题，所需要的知识直接从大脑存取——存取记忆，这无疑是一种“最佳检索”。第二种，当你知识透支，或者需要获取新知识时，你要借助所处的知识环境。但是如果你知道这些网址、书名、书号、专利号、作者或者其他线索，直接进入，而不必从漫漫知识海洋寻觅你所需的那一叶扁舟，岂不快哉！

但是，“人生也有涯，而知也无涯，以有涯求无涯，岂不殆乎”？因而，很多检索工具“投其所好”，为用户设计了很好的辅助记忆模式。比如 IE 浏览器，大家一定熟知有收藏夹的功能，就是为帮你把冲浪浏览过程中感兴趣的、有价值的网址保存下来，帮你记忆；还有它设有一个缓冲文件，可以保留你过去一段时间浏览或最近浏览的若干网页；还有地址栏也具有 cookies 功能，能够自动记录你最近浏览的网站 URL，只需下拉地址栏就能方便进入。这些都是帮你记忆的模式。现在 web2.0 环境下的辅助机制更多了，你可以用关键词订阅、网络书签、博客超级链接等多种形式记录你感兴趣的信息。

为什么有这些功能？因为网络服务商也知道，最佳的检索方式就是记忆。

五、检索工具走向傻瓜式

信息检索是一门技术，因为它要面向众多的信息检索工具，所以很多同学可能望而却步，是不是有可能学不会？虽然检索理论越来越复杂，检索的对象越来越多样，检索结果要求越来越精确，但是没有一个人认为检索技术是学不会的，因为检索操作越来越容易，那些检索复杂，过程繁琐的检索工具逐渐被大众的、傻瓜式的检索工具所代替。

在 20 世纪 60 年代，要完成对数据库的检索不仅需要复杂的技术流程，而且还需要足够的耐心。首先必须自己编写检索程序，其次程序必须转换成 ASCII 码，全部是“0”和

“1”组成的字符串，接着再用纸带打孔录入，然后再等计算机经过一整夜“无差错”地运算得到检索结果。到1969年，《化学文摘》电子版问世的时候，出版商已经编好了检索界面，通过目录浏览，或者字符串匹配，能够完成基本的检索过程。1972年，DIALOG系统投入商业运营，这套联机信息检索系统已经具有完备的检索算法，运用字段控制，可以实现存取、剔除、转录、输出等功能，基本实现了追溯检索（RS）、定题检索（SDI）、查新服务、文件传递和联机定购等服务。随着20世纪80年代视窗系统的问世和90年代网络的民用化发展，网络浏览器和搜索引擎相继出现，极大地丰富了我们的信息获取方式。简便、迅速、实用是现代检索工具的典型特征，所见即所得。高级检索也不必编写复杂的检索式，利用窗口的提示即可方便完成大部分的检索需求；对专业检索者保留了索引号、书号、报告编号、专利号等唯一标识码。对于网上信息泛滥的情况，也出台了唯一数字标识符（DOI）保证信息的真实性和权威性。比如在科技信息文献领域，每一篇上网的文献都给一个数字标识符DOI，不论这篇文章如何被引用、转贴，DOI都随文章一起，让用户知道这篇文章是权威的，数据和结论是正确的。目前，中国科学院数字图书馆项目正在致力中国数字信息资源的唯一标识符制定，不久的将来，大家也会看到中国科学数字信息贴上“合格证”、“质检证”。

总之，检索系统开发商本身也是以用户为中心，设计可用、实用、适用的检索工具，大部分检索工具会“老少咸宜”。既然如此，那么检索有何难？又何必当作一门课程专门学习呢？那是因为下面要介绍的第六条：确定检索范围有时比检索过程本身要复杂得多。而且，如果你需要非常精确的检索结果，这样的傻瓜工具可能还难当重任，你需要高级检索，甚至是自己编写检索程序去筛选网上信息，也就是第七条告诉大家的如何去掌握高级检索。

六、确定检索范围有时比检索过程本身要复杂得多

信息检索是要发现并提取所需信息的过程，从原理上讲要经过信息加工人员按照一套规则加工，然后存储入库，最后由检索者发现并获取信息的过程。信息检索，简单地说就是信息的有序化识别和查找。广义的信息检索包括信息的汇集、存储与查找，而狭义的信息检索仅指有序化知识信息的检索查找。通常人们所说的信息检索是指后一过程，即信息查找过程，也就是狭义的信息检索（Information Search）。但是，不论是狭义的信息检索，还是广义的信息检索，作为用户，要从数据库、信息库中获得自己所需要的资源，首先必须完成两个工作：第一，你如何描述你所需要的信息资源？第二，如何让系统知道你所需的信息资源范围？即使再精妙的检索工具，如果检索范围，或者更确切地说是检索策略和检索入口选取不当，也难以获得理想的检索结果。

那么，究竟如何去确定比较合适的检索范围呢？用一个词概括：知己知彼。所谓“知己”就是在检索实施之前，要分析自己的检索需求，它的概念范畴、类型、可能的形式、时间特征、作者、出版商信息、专门信息（版本号、ISBN号、分类号、标准号等）等等。现在，有很多专家就是研究检索需求表达的，根据需求我们把信息资源加工好的过程叫信息组织过程。那么，“知彼”就是要明确检索工具或咨询专家的领域范围或者知识匹配格式。比如，很简单的一个问题，如果你要找“文化大革命期间的中国经济学研究”，

在CNKI、维普的数据库里的信息含量都不是很大，至少第一手的资料会很少。因为这些工具收录的最早的文献也只有1979年。还有，利用Baidu找期刊论文或科技论文（也许很多同学这样做过），就不是最好的方法，Baidu在社会信息方面具有优势，更新速度快，但在科技文献里面不如一些专业的工具，比如Google的学术搜索（scholar. google. com）。同样，检索专家也是这样，很难想象你能从一个物理学家那里获得关于“中国上古历史”的准确而系统的信息和知识。举这两个例子，就是为了说明检索范围不仅仅是在一个检索工具前“冥思苦想”用什么词可以事半功倍，而且需要首先考虑用什么检索工具。

当然，确定检索范围是经验性很强的工作，经济学叫体验产品，就是你经历越多，越熟练，掌握得越好。如果把所有的检索范围经验都罗列出来，都可以编写《检索百科手册》了，而本书不能这样教大家，也不会这么做。我们的专家和老师教给大家的是怎样去积累检索经验，怎样避免最基础的检索误区，不至于“事事百度，次次句子”。因为在方法上，很多同学每次都是“句子”检索，有很多同学不知道如何获取检索需求，比如要找“亚洲金融危机对中国的影响”，他的检索词就是“亚洲金融危机对中国的影响”，一来检索范围受局限，会漏检很多重要信息；二来在网络检索中会检索到很多相关性不强的信息，很多搜索引擎会以“影响”作为检索词匹配。

由此而见，学精信息检索，恐怕不是非常简单。武汉大学信息管理学院的信息素养学会与湖北移动曾经组织过一次面向整个校园的有意思的比赛：信息搜索大赛。比赛结果表明：检索能力最强的不是信息管理学院的学生，但是在20强中，信息管理学院占有绝对优势（每个学院的参赛人数相当）。同学们，对此有何感想呢？在第二章，我们还有更多关于信息检索基本概念的知识与大家分享，也许对大家理解检索有所帮助。

七、高级检索往往事半功倍

美国的Amanda Spink和Bernard J. Jansen在最近的专著《网络搜索：网络公共检索》一书中综合研究了网络行为方面的进展。通过Excite，AlltheWeb. com，Alta Vista和Ask Jeeves1997年到2003年的查询数据，用户每一次检索行为会选用2～3个检索词；在开放查询服务中，用户对任务的描述主要通过关键词，而不是完整的句子描述；大约2/3的用户习惯一次检索到位，超过6/7的用户不会使用2次以上的查询检索，平均检索查询次数是1.6次；有大约8%的用户接受或者利用模糊检索来获取信息；虽然使用布尔代数和专业查询的用户呈现增加趋势（增加了28%），但仍然只占总用户数量的1/18，而且检索语言错误非常多；只有大约10%的搜索引擎提供布尔运算功能。对于检索结果而言，大多数用户只愿意阅读排名前几位的检索结果，平均浏览2.35页检索结果，超过70%的用户集中于检索结果的前10项。

那么，什么是高级检索？为什么高级检索的使用率不高？甚至本身搜索引擎都不提供高级检索呢？所谓高级检索，主要是指检索人员在检索入口采用比较复杂的检索式，利用多条件限制的方式获取信息的检索模式。相对而言，同学们用得最多的是主题检索，输入一个词，看检索结果中是否具有相匹配的内容。其实，我们也知道，揭示信息内涵的特征还是比较多的。传统的检索环境下，我们称之为标引词或叙词，用来标引和描述文献的词语。现在网络计算环境下，要描述信息的数据我们称之为“数据的数据”——元数据。元

数据有很多类型，不同的资源类型会有不同的元数据指标，也有网络通用的元数据体系，在后面搜索引擎介绍中会告诉大家这些元数据的作用，因为通过元数据的匹配比信息内容本身的匹配要相对容易，所以针对元数据的检索比针对内容本身的检索甚至更有效率！网络环境下的高级检索大部分都是针对元数据设计的。

既然高级检索有这么多优点，为什么大家不用？理由很简单，主题检索几乎不需要学习成本，而高级检索必须要学习，学习元数据的构成、检索字段、算法以及检索式的编写。因为学会这些的用户不多，所以使用率也就不高了。没有需求，搜索引擎和网站也就不会刻意编制高级检索体系了。

其实，高级检索中有很多是非常简单，而检索效果非常有效的工具，对于提高检索速度、扩大检索范围或者精炼检索结果相当有效，同学们不妨学习之。

八、检索知识专家比检索知识本身更有效率

检索过程是用户发现和获取所需信息的过程，目标是获得有用的信息，而过程是可以多样化的：你可以亲自去浏览和发现，也可以借助相关的检索工具，还可以直接询问。询问也是检索方法的一种，而且是最高效的一种，即使你对所需知识和信息一无所知，在合适的专家那里，你也能得到几乎完美的回答。但是，有效的询问除了沟通过程的高效，更重要的是你要询问正确的人。很多企业在解决他们的信息需求时，就借助这样的专家体系，尤其是医生、律师、经理和知识工作者，怎样快速而有效的找到解决方案往往决定其工作成败，所以现在检索专家很关注专家系统在企业的应用：如何发现和标引专家？

当检索对象从知识本身，跨越到知识的载体时，我们讲检索已经是一种社会化概念了，而且知识载体，也就是专家像是知识网络中的一个知识结点，构成了一种社会化的知识网络。在这样的环境下，发现和获取信息和知识就是社会网络检索、专家检索。

作为检索技能的一种，我们希望同学们对“专家集会”的若干场合有所了解，知道在什么样的社会环境中去获取所需信息。

九、找到信息并不是检索的结束：检索结果的分析、精炼和组织

检索过程是信息获取利用的一个环节，检索到信息只有与我们的需求相结合，才能被利用，才能发挥“知识的力量”，所以找到信息并不是检索的结束，要根据利用效果随时调整自己的检索过程，信息检索和信息利用相辅相成，检索过程再好利用不好也是枉然。因而，最后强调信息的分析、评价和利用，争取用对信息、用好信息。

相关性分析是最常用的评价手段，但是相关性的概念本身带有一定模糊性，没有定量的评定标准，不同用户对同一资源检索结果可能有不同的评价。因而，相关性评价方法一般限于一定范围之内，如手工检索的相关性判断依据是检索课题本身与检出信息在内容主题上相符合，其结论主要由用户完成；机检条件下相关性判断依据是检索提问标识与系统标识的相符性（检索策略），其结论由计算机完成；信息适用性的判断，即检出信息在相关前提之下的适用价值，其结论由用户确定。

另外，国外一些学者提出用“有效性”、“实用性”的概念取代“相关性”来评价检索效果，将检出信息的时效性、可获取性和信息吸收程度综合考虑，来评定检索结果的适用

程度。但是，这些概念仍然围绕用户检索需求的相关性展开。

通过对检索结果的分析和评价，能够保障检索结果的利用率。同时，要利用好检索结果，还要学会精炼和组织。所谓信息精炼主要是对检索结果在时间、主题上作一定的归纳和整理，提炼核心观点，分析并总结新观点的过程。其实也是教育学里强调的基于资源的学习过程。一些教育专家认为，“基于资源的学习是根据一定的学习目标，以培养学生综合素质为核心，让学生通过搜集相关资源，判断资源的真与假、是与非，并以此为基础进行演绎、推理、概括等思维过程后，用合理的形式呈现结论，使之越来越完善、合理的学习过程”。

这就意味着，信息检索应该是有的放矢，有学习目标，制定相关性的评价，然后才是对检索结果的精炼和利用。

十、尊重隐私、版权和伦理，会有更丰厚的检索回报

最后强调信息伦理。信息伦理是在虚拟环境下，大家因该遵循的公共秩序和道德体系。因为信息检索是信息发现和获取的过程，所以有些人就会利用信息检索的技术获取一些不该获取的信息，打破了正常的信息秩序，我们称之为违反信息伦理，而严重的可能是信息犯罪。那么，有哪些信息伦理是不应该被打破的呢？一句话概括是，强调信息的公共获取，所获取信息一定要具有公共性或公开性。

如果大家对信息公共性不容易理解，那么我们可以简单举几个例子。

第一，不应该挖掘获取他人的个人隐私。很多国家都有人口数据库，都有人事记录档案，在没有合法的手续之前，这些信息都不应该被获取。

第二，不要随意侵害别人版权。比如对音乐、电影和科技论文，作者都具有相应的信息版权，除非你获得合法的使用权，随意下载、转贴和利用都视为侵权。当然，网络版权是比较复杂的问题，因为一方面网络环境下公开信息意味着信息公共性，可被获取利用；另一方面版权所有者的作者权和相关延展权利依然有效，目前存在很多争论。但是，作为利用者，应该尽量回避类似纠纷，尤其在大量应用或正式利用过程中。

第三，伦理问题。不要恶意去攻击、篡改和扭曲他人信息。比如黑客行为、给他人的私人网络设置木马程序等行为，将给他人信息获取带来极大不便，因而我们称为有违道德。

第四，泄漏机密信息。有些信息涉及国家利益或企业的经济利益，不宜公开传播，如果通过检索等相关信息获取手段获取、公开和传播这些信息是不道德，甚至是违法的。

知识卡片：

所谓信息伦理，是指涉及信息开发、信息传播、信息的管理和利用等方面的伦理要求、伦理准则、伦理规约，以及在此基础上形成的新型的伦理关系。信息伦理是信息技术的价值制导，它为信息技术的运用设定善的价值坐标。信息技术本身是价值中性的，而人的行为则具有明确的价值向性。在信息伦理的指引下，通过人们运用信息技术的行为，价值中性的信息技术，就可以导致善的价值的生成。

大家也许会认为，上面这些规定缩小了你可利用的“信息范围”，损害了你的“信息检索权”。事实上，正是信息秩序的保障，才让更多的信息贡献者愿意共享自己的知识或信息，更多企业和机构愿意开发信息资源，反而丰富了你所能获取的信息内容。两相比较，孰优孰劣，一目了然。

当然，信息检索涉及的方法和对象都是多样化的，要完成从熟悉到精通并非一日之功，在开始学习信息检索这门课程之前，再让我们来回顾一下本章告诉我们的几个基本原理：

第一，信息检索技能是重要的：检索技能是现代社会必备的基础技能，检索技能可以创造财富；

第二，信息检索如何实现：检索工具不是万能的，检索工具越来越傻瓜化，最佳的检索方式是记忆，确定检索范围有时比检索过程本身要复杂得多，学会高级检索往往事半功倍，检索知识专家比检索知识本身更有效率；

第三，如何利用检索结果：找到信息并不是检索的结束，检索结果的分析、精炼和组织，尊重隐私、版权和伦理，会有更丰厚的检索回报。