第 1 章

信息素养与信息检索

在信息时代，如同读书识字对我们的意义一样，信息素养（information literacy）已经是国民必备的基本生存能力之一。美国图书馆学会（American Library Association，ALA）认为，“具有信息素养的人，能够判断什么时候需要信息，并懂得如何去获取信息，以及如何去评价和有效利用所需的信息”。然而，可以很轻易地确认，现在是迄今为止信息最多的时代，同时也是信息最为混乱的时代。我们比以往更多地接触信息，获得信息，同时也更容易在信息中迷失。培养国民信息素养已经成为图书馆学和信息科学在 21 世纪最为主要的挑战之一。

在所有培养信息素养的途径中，为社会公众和专业人士提供相应的信息素养教育课程已经成为全球最为通行的做法，尤其是在高等教育领域，几乎每所高校都提供了类似课程。自 1984 年我国教育部印发《关于在高等学校开设“文献检索与利用”课的意见》以来，历经 30 余年，信息检索课教学已全面纳入高校，特别是本科阶段的通识教育体系，在高校信息素养教育中居于核心地位。这些课程或以“信息素养”冠之，或以“文献（信息）检索”冠之，等等，其目的都在培养高校在校生的信息素养，以适应节奏越来越快的社会生活。

1.1 信 息 素 养

1.1.1 信息素养的概念

信息素养的概念最早由美国信息产业协会主席保罗·泽考斯基（Paul Zurkowski）于 1974 年提出，他将信息素养定义为“利用大量信息工具及主要信息源使问题得到解答的技能”。此后，信息素养一直被定义为一种内在的能力。2003 年和 2005 年，联合国教科文组织曾分别召开两次专题性的世界大会，并发布《布拉格宣言》和《亚历山大宣言》，强调“信息素养是人们在信息社会和信息时代生存的前提条件，是终身学习的重要因素，能够帮助个体和组织实现其生存和发展的各类目标。它能够确定、查找、评估、组织和有效生产、使用和交流信息来解决问题”。

信息素养的培养必须界定其标准，得益于得天独厚的优势，这一标准在高等教育领域被最先提出。2000 年，ALA 与美国大学与研究图书馆协会（Association of College and Research Libraries，ACRL）共同提出《高等教育信息素养能力标准》（*Information Literacy Competency Standards for Higher Education*）。《高等教育信息素养能力标准》包含 5 项能

力指标、22 项表现指标和 87 项成果指标，比较全面地概括了高等院校学生信息素养的标准。这 5 项能力指标包括如下内容。

（1）有能力决定所需信息的性质和范围。

（2）可以有效地获得需要的信息。

（3）能够评估信息及其出处，并能将挑选的信息融合到他们的知识库和价值体系。

（4）不管是个人还是作为团体的成员，都能够有效地利用信息来实现特定的目的。

（5）熟悉许多与信息使用有关的经济、法律和社会问题，并能合理、合法地获取信息。

2015 年，ACRL 在《标准》的基础上又提出《高等教育信息素养框架》（*Framework for Information Literacy for Higher Education*）。《高等教育信息素养框架》改变了原有《高等教育信息素养能力标准》的立足点，强调对学生元素养的培养，将“权威的建构性和情境性”“信息创建的过程性”“信息的价值属性”“探究式研究”“对话式研究”“战略探索式检索”6 个阈概念及其涵盖的知识技能和行为方式设定为学习成果，用以指导信息素养课程及教学计划的重新设计。

1.1.2 信息素养教育与信息检索课程

从 20 世纪 80 年代开始，由于新技术手段在西方发达国家的快速发展，世界各国都意识到信息素养教育的重要性，认为公民的信息素养水平会直接对国家的竞争力产生重要影响，加强全体公民的信息素养教育被提上新的议事日程。在这一方面，美国、英国、日本等信息化程度较高的国家走在了世界前列。

1. 美国的信息素养教育

美国是信息素养及其教育的发源地。自泽考斯基首倡信息素养概念之后，美国教育界一直对其高度重视。1989 年，美国信息素养主席委员会发布《信息素养主席委员会总结报告》，对信息素养的概念界定、要素构成，以及对个体学习、生活所起的隐形作用做了系统阐述，对美国未来信息素养教育的发展提供了建议。1996 年，美国信息素养论坛确定了《信息素养教育在普通教育计划中的作用框架》。2016 年以来，美国政府先后推出“CS For All”（全民计算机科学）、“AI For All”（全民人工智能）教育计划，进一步推动信息技术创新教育。

在高等教育阶段对信息素养的研究上，美国专门的组织机构和协会对学校信息素养培养发挥了重要作用，特别是 ALA 与 ACRL 更是居功甚伟。2000 年，ALA 与 ACRL 共同发布《高等教育信息素养能力标准》，为高校学生的信息素养提供了一系列指导性评测标准。2015 年，美国大学与研究图书馆协会开启审视信息素养的全新视野，推出了新的研究成果，发布了《高等教育信息素养框架》。依托这些组织发布的一系列关于信息素养的标准、报告和教育政策，美国高等教育领域信息素养教育实现了快速稳定发展，同时也为其他国家根据不同国情制定自身信息素养标准提供了参考和基准。

2. 英国的信息素养教育

英国是全球公认的高等教育强国，在信息素养教育方面，积累了丰富的经验，形成

了独特的优势和特色。英国开展信息素养研究和教育的主要机构有英国国家大学图书馆协会（Society of College，National and University Libraries，SCONUL）、英国图书馆协会（Chartered Institute of Library and Information Professionals，CILIP）、联合信息系统委员会（Joint Information System Committee，JISC）及各个高校的图书馆（信息服务）机构。依托这些组织和机构，建立了一套层层细化的高校学生信息素养标准，在全国范围层面，出台了一系列通用标准。

2003 年，CILIP 提出了信息素养定义，并定义 7 项信息技能，即信息需求、可获取的资源、如何检索信息、评估信息的需求、如何使用或拓展结果、信息使用的伦理道德与责任，以及如何交流、分享、管理成果。1999 年，SCONUL 发布《高等教育信息技能——7 顶指标》，2011 年又将 7 项指标模型更新成为核心模型，包括信息需求识别、信息需求研究、检索策略计划、信息获取、信息评价、知识管理和知识展示与创新。这 7 项指标模型按照科研学习流程对信息素养框架进行整理规划，使其能够更加自然、合理地融入其中。目前，该模型已成为多个区域性协会、机构及高校制定相关政策的主要依据。JISC 则通过支持各种研究与服务项目，来激励英国高校数字技术的创新应用，从而服务于信息素养教育。

此外，英国的信息素养教育在地区层面和学校层面也可圈可点。比如，在地区层面的有《威尔士信息素养框架》《苏格兰信息素养框架》等，在学校层面有剑桥大学的《为了信息素养的课程》等。

3. 日本的信息素养教育

日本是高度重视信息素养教育的国家之一。1986 年，日本《关于教育改革的第二次报告》中就公开提出“信息利用能力”一词。1998 年，日本图书馆协会出版《图书馆利用教育准则》。同年，京都大学开始在全校开设“信息探索入门”的基础课程，此后，信息素养教育逐渐发展到了日本全国的各所大学。2000 年以后，包括“信息素养”等课程在内的信息相关课程已成为日本所有大学的必修课程。

2014 年，在参照 ACRL 发布的《高等教育信息素养能力标准》的基础上，日本结合自身实际制定了《高等教育信息素养标准》。在该标准中对信息素养作出了如下定义：所谓信息素养，就是当需要信息的时候，能够认识信息，有计划地收集、评价、整理、管理信息，以及利用信息并且有效地发布信息的能力。该标准适用学生、教师、大学经营者、图书馆员等对象，它将信息探索过程分为 6 个阶段，并在每个阶段制定了由初级到高级的能力要求。

4. 中国的信息素养教育

在中国，系统化的信息素养教育起步于高校。1981 年，《人民日报》先后刊载刘毅夫、潘树广关于《建议在高校开设文献检索课》的文章。1983 年 10 月，全国召开“全国高校《文献检索与利用》课专题讨论会”。1984 年教育部印发《关于在高等学校开设“文献检索与利用”课的意见》（〔84〕教高一字 004 号）文件。1985 年 9 月颁发《关于改进和发展文献课教学的几点意见》，提出文献检索课“要逐步实现分层次连续教育”的教学指导思想。这两个文件的颁发，为文献检索课的教学奠定了基础。1992 年 5 月

原国家教委印发《文献检索课教学基本要求》，对文献检索课的课程性质、教学目的、要求等做了细致而全面的规定，从而使文献检索课从形式到内容更加规范化、系统化。此后历经 30 余年，文献检索课程不断适应时代变迁，或沿用传统名称，或易名为信息检索和信息素养，在教学内容与形式上也不断更新优化，全面纳入高校，特别是本科阶段的通识教育体系当中。

目前，我国高校信息检索课程多由图书馆负责，少数则由专业教师承担。2015 年 12 月，教育部印发《普通高等学校图书馆规程》，指出高等学校图书馆的主要职能是教育职能和信息服务职能。高校图书馆应全面参与学校人才培养工作，充分发挥第二课堂的作用，采取多种形式提高学生的综合素质。图书馆应重视开展信息素养教育，用现代教育技术加强信息素养课程体系建设，完善和创新新生培训、专题讲座的形式和内容。随着社会的发展、时代的进步，以信息检索为主的信息素养教育被越来越多的学校重视，发挥着越来越重要的作用。

1.2 信息与信息检索

1.2.1 信息及其相关概念

有关“信息”一词，读者已经通过书本、网络等各种渠道了解过，对其并不陌生。此外，在学习和学术研究中，还经常接触“文献”这个词。甚至，在某些场合，文献和信息又常常被一并提起。那么，究竟什么是文献？什么又是信息？这是首先要了解的问题。

1. 文献

“文献”一词有着悠久的历史，最早见于《论语》：“夏礼吾能言之，杞不足征也；殷礼吾能言之，宋不足征也。文献不足故也，足则吾能征之矣。”宋代朱熹认为这里的“文”是指典籍。而“献”是指熟知史实的贤人。

根据中国国家标准《信息与文献 参考文献著录规则》（GB/T 7714—2015），文献是记录有知识的一切载体。国际标准《文献情报术语国际标准（草案）》（ISO/DIS 5127）则将文献定义为：“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。”综合这两个标准中的定义可以认为，文献包含两个基本要素，一是知识，二是相应的载体。例如，将知识记录在纸张上，该纸张便可以称为文献；将知识以电子文件形式存储在磁盘上，该磁盘也可以称为文献。由此也可以看出，文献的核心是知识，没有知识，纸张便不能称为文献。当然，载体也同样重要，不同的载体对于文献的保存与传播有着非常明显的影响。

2. 信息

到目前为止，在学术界似乎对信息的概念还没有达成一个广泛的共识。其中，流传最为广泛的是信息论的创立者申农（Shannon）给出的定义，即“信息是不确定性的消

除”。我国著名的信息学专家钟义信教授认为，“信息是事物存在方式或运动状态，以这种方式或状态直接或间接的表述”。美国信息资源管理专家霍顿（Horton）给信息下的定义是：“信息是为了满足用户决策的需要而经过加工处理的数据。”他认为，信息是经过加工的数据，或者说，信息是数据处理的结果。

不管信息如何定义，人们对信息的理解最初哪怕是到现在也是借助于信息的载体而进行。一般认为，信息依附于某一特定载体就形成了文献。文献是传递信息的介质，是固化的信息。文献的本质是知识信息，所以很多时候又合称为文献信息，在很多场合也使用“文献”这一词语作为信息的替代术语。在本书中，我们无意严格区分信息和文献这两个概念，甚至在某些场合我们会使用文献信息这一复合词组。

3. 数据、知识、智慧等其他概念

与信息常被一同提及的概念还有数据（data）、知识（knowledge）和智慧（wisdom）。传统上，人们认为数据、信息、知识、智慧构成了一个金字塔形状（图 1.2-1）。

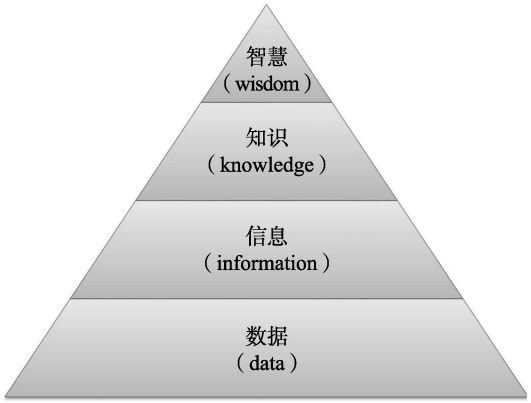

图 1.2-1　DIKW 模型

“数据”一词来源于拉丁语“dare”（意为给予），具有假设、事实、评估等意义。数据无处不在。过去，人们习惯把数字的组合称为数据，如 3.1415。但在今天，这样的理解显然不够全面，一般认为数据是可以被记录和识别的一组有意义的符号，它可以通过原始的观察或度量得到，如人们的姓名数据、学生的成绩数据、天气温度数据、语音数据等。数据具有客观性的特点，即只要记录下来，数据就是客观存在的。

信息由数据加工得来。数据是信息的原始类型，而信息是经过加工的数据。比如，当人们在研究学生的成绩单时，上面记录的分数仅仅是一些数据。要读懂这些数据，就必须了解数据背后要表达的含义。一旦对数据做出解释，我们就能得到成绩单上所蕴含的信息。信息具有客观性，即从数据中加工得出的信息是稳定的。比如，通过分析不同班级的成绩可以得出 A 班的平均成绩比 B 班高。同时信息也具有主观性，对其解读会因每个人的主观认识不同而不同。比如，同样是 60 分，有些人解读为刚好及格不够理想，有些人则会认为与上一次比较有很大的提高。

数据和信息都是客观存在的，而知识则是由人类大脑筛选、组织和理解的信息。马克卢普（Marchlup）认为，“信息意味着传输，可以通过被告知而获得，但知识是一种

状态，必须借由思考而获得”。

智慧是生命体所具有的基于生理和心理器官的一种高级创造思维能力。拥有知识并不意味着具有智慧。在日常生活中，智慧体现为更好地解决问题的能力。

1.2.2 信息的分类

根据不同的划分标准，文献信息可以划分成多种类型。常用的分类标准主要有出版类型、载体的形式、加工层次、内容的公开程度等。

1. 按出版类型划分

按照出版形式，文献信息可以划分为图书、连续性出版物及特种文献。其中，连续性出版物包括期刊、报纸，特种文献包括学位论文、会议文献、专利文献、标准文献、研究报告、政府出版物、档案资料等。此种划分方式是目前通用的方式，后续的讨论均基于此划分方式。不同类型文献信息的介绍详见本书第 3 章。

2. 按载体的形式划分

为了有效地存储、传播知识信息，人类先后发明了各种各样的物质材料来记录信息。目前，文献信息按载体划分主要有纸张型、缩微型、音像型、电子型等 4 种。

（1）纸张型文献。它是以印刷等为手段，将信息记载在纸张上形成的文献。它是传统的文献形式，不需要借助其他工具便可阅读，但存储密度小、体积大，不便于管理和长期保存。

（2）缩微型文献。它是利用光学技术以缩微照相为记录手段，将信息记载在感光材料上形成的文献，如缩微胶卷、缩微平片。其特点是存储密度大、体积小，便于保存和传递，但必须借助专门的设备才能阅读。世界上许多文献信息服务机构都将长期收藏的文献制成缩微品加以保存。

（3）音像型文献。它是采用录音、录像、摄影、摄像等手段，将声音、图像等多媒体信息记录在光学材料、磁性材料上形成的文献，也称视听型文献，如音像磁带、唱片、幻灯片、激光视盘等。其特点是形象、直观，尤其适于记录用文字、符号难以描述的复杂信息和自然现象，但其制作、阅读需要利用专门设备。

（4）电子型文献。它是指以数字代码方式将图、文、声、像等信息存储到磁、光、电介质上，通过计算机或类似设备阅读使用的文献，也称机读型文献。电子文献种类多、数量大、内容丰富，如各种电子图书、电子期刊、网络数据库等。其特点是信息存储量大，存取速度快，传递信息迅速，易更新，可以融文本、图像、声音等多媒体信息于一体，信息共享性好、易复制，但必须利用计算机硬件和特定软件才能阅读。

3. 按加工层次划分

按照信息的加工处理方式，信息可以划分成零次信息、一次信息、二次信息和三次信息。

（1）零次信息。它主要指尚未经过系统整理的零散信息，如未正式发表的手稿、讨论稿、实验原始数据、人们在某些专业会议上口头交流的经验或某些论点、审稿意见

等。零次信息内容新颖，往往包含作者最真实的情感、观念及瞬间产生的思想灵感，不失为一类重要的信息资源。以往，由于零次信息或未经过系统整理，或存储条件限制等其他因素，未能得到充分交流与利用，但现在零次信息已越来越得到重视。例如，已有许多期刊要求在投稿时须提交相关的实验原始数据并予以公开，审稿意见一并附录等。

（2）一次信息。它主要指作者以本人的研究成果为基本素材而创造或撰写的文献，如图书专著、期刊论文、专利说明书等。一次信息包含的内容往往比较具体、详尽、系统。

（3）二次信息。它主要是指信息工作者对一次文献信息进行加工、提炼和压缩之后得到的产物，是为了便于管理、检索和利用一次文献而编辑、出版和累积起来的工具性文献。一般包括目录、题录、文摘、索引等。

（4）三次信息。它主要是指对有关的一次文献和二次文献进行广泛深入的分析研究之后概括而成的产物，具体包括述评、综述、文献指南等。很多时候，三次文献与一次文献在体例上具有相似性，比如，同一期刊中，研究论文（research article）或原创性论文（original article）为一次信息，而综述性论文（review）为三次信息。通常，三次信息包含的信息量非常大，是读者应该重点关注的类型。

4. 按内容的公开程度划分

文献信息按照内容的公开程度，可划分为白色文献、黑色文献和灰色文献。

（1）白色文献。它一般是指正式出版或在社会上公开流通和传播的文献，如图书、期刊、报纸、专利、标准等。

（2）黑色文献。它一般是指尚未被公开，仍处于保密状态的文献信息，如保密的科技报告、技术资料等。

（3）灰色文献。1997 年在法国卢森堡举行的第三次国际灰色文献会议中对灰色文献做出如下界定：“不受商业出版商控制，而由各级政府、学术单位、工商业界所产制的各类印刷与电子形式的资料。”一般认为，灰色文献是介于白色文献与黑色文献之间，不属于保密级别，但也不公开出版发行或传播的文献。灰色文献包括：非公开出版的政府文献、学位论文，不公开发行的会议文献、科技报告、技术档案，不对外发行的企业文件、企业产品资料、贸易文件（包括产品说明书、相关机构印发的动态信息资料）和工作文件，内部刊物、交换资料、赠阅资料等。灰色文献所涉及的信息广泛、形态多样、内容丰富、时效性强、社会价值巨大，受到当今各国政府及情报机构的广泛重视。但由于其流通渠道特殊，读者获取需要掌握一定的技巧。

有关灰色文献的更多资料，读者可以关注 GreyNet International 网站（http://www.greynet.org/）。该网站于 1993 年 12 月创建，目标在于促进灰色文献领域人员和组织之间的对话、研究和交流，并进一步寻求在网络环境中识别和分发关于灰色文献的信息。读者通过该网站可以了解包括灰色文献国际会议、灰色资源的创建和维护、邮件列表（listserv）、灰色文献杂志等信息。

1.2.3 信息检索

信息检索（information retrieval）是一种有意识地、主动地获取信息的过程和行为。作为一种实践活动，信息检索由来已久，但作为一个比较规范化的学术术语，最早由莫尔斯（Mooers）于 20 世纪 50 年代首次提出。莫尔斯也因此获得 1978 年的美国信息科学协会荣誉奖。随着信息资源的急剧增长，以及各种网络搜索引擎的出现与普及应用，信息检索这一术语逐渐由学术界向社会公众传播开来。

信息检索的概念有广义和狭义之分。广义地说，是指将信息按一定的方式组织和存储起来，并根据用户的需要找出相关信息的过程。所以，它的全称又叫信息存储与检索，即包括信息的“存”和“取”两个环节。作为一种有目的的和组织化的信息存取活动，信息检索中的存和取之间存在密不可分的关系。良好的信息组织与存储是保证检索的质量和效率的前提，为此，许多检索系统都应用了精密的分类与主题检索语言。当然，过于复杂的组织与存储对检索系统的性能要求也越高，因此在实际的检索系统中，往往要在存和取之间做出兼顾与平衡。

狭义的信息检索则仅指该过程的后半部分，即从信息集合中找出所需信息的过程，相当于“信息查询”或“信息查找”。曼宁（Manning）从技术的角度认为，信息检索是从文档集合中查找满足某种信息需求的具有非结构化性质的资料。我们将这一定义适度扩大化，将狭义的信息检索定义为从信息源的集合中，找出与用户信息需求有关的信息资源的过程。比如，从图书馆的大量藏书中查找某一种图书，从论文数据库中找出包含有特定主题的论文，等等。值得注意的是，对非信息检索领域从业人员的读者而言，一般只关注取的过程，很显然读者使用百度并不会去详细研究百度的内部检索机制。当然，适当了解存的过程与机制对于提高取的效率是有益的。

1.3 信息检索的出现背景

在人类漫长的历史中，信息检索是短暂的，它的出现是人类社会活动趋于复杂化，信息量不断增长的必然结果。考察信息检索出现的背景对于理解信息检索的概念及后期学习具有十分重要的意义，同时亦十分有趣。

1.3.1 信息的增长

在古代，信息是较为匮乏的。在文字出现前，信息的传播均依赖口语。口语传播的弊端是传播的信息和知识可能会在每次复述中被遗忘或修改。随着文字的出现，信息的传播逐渐由口语传播转向文字传播。文字具有持久性、复杂性、可移动性，文字传播成为文化传播和历史传承的重要工具，客观上成为信息增长的基石。总体上看，随着时代的发展，信息的数量呈现不断增长趋势，其中技术的推动力量是主要的。

在我国古代，国家藏书以皇家藏书为主，因此我们可用历朝皇家藏书的数量情况作为整个社会图书数量的代表来进行考察信息的增长过程。在汉代，皇家藏书有记载的数

字是 677 种，共 13 269 卷。读者所熟知的北宋史学家司马光在编撰《资治通鉴》时曾形容自己读过的书“浩如烟海”。在今天看来，这一规模并不算大。

纸张发明以前我国使用竹简作为书写的主要材料。竹简作为书写材料非常笨重，虽然也有使用较为轻便的绢帛，但是成本非常昂贵，也不适于书写，客观上限制了知识的制作与传播，也提出了发展更为方便的书写材料的要求。

西汉时期（公元前 206 年），中国已经出现纸张，但此时的造纸术较为粗糙。东汉元兴元年（105 年），宦官蔡伦改进造纸术后，我国才开始大规模使用纸张作为书写材料。蔡伦使用树皮、麻头及敝布、渔网等原料，经过挫、捣、炒、烘等 4 个工艺步骤制造出纸。这种纸，原料容易找到，又很便宜，质量也提高了，因此逐渐普及使用。为纪念蔡伦的功绩，后人把这种纸叫作“蔡侯纸”，由此造纸术也成为我国古代“四大发明”之一。造纸术的出现使书籍的生产变得容易，书籍的数量也因此大大增加。到了 721 年，唐朝皇家的藏书有 5000～6000 种（约 8.9 万卷）。

北宋庆历年间（1041—1048 年），毕昇发明了活字印刷术。活字印刷术发端于唐朝的雕版印刷术，是一种古代印刷方法，是中国古代劳动人民经过长期实践和研究发明的，也是印刷史上一次伟大的技术革命。活字印刷术先制成单字的阳文反文字模，然后按照稿件把单字挑选出来，排列在字盘内，涂墨印刷，印完后再将字模拆出，留待下次排印时再次使用。2010 年 11 月 15 日，活字印刷术被联合国教科文组织保护非物质文化遗产政府间委员会第五次会议审议通过，列入“急需保护的非物质文化遗产名录”。由于活字印刷术的发明，书籍的批量生产比以往更为容易。

在与我国明朝同时期的欧洲（15 世纪），德国人古登堡（Gutenberg）发明了活字印刷机。西班牙历史学家和传教士门多萨（Mendoza）在《中华大帝国史》（1584 年出版，该书是西方世界第一部详细介绍中国历史文化的巨著）一书中提出，古登堡发明活字印刷机受到了中国活字印刷技术的影响，其途径可能通过两个：一是经俄罗斯传入；二是通过阿拉伯商人传入。古登堡的印刷机实现了书籍和小册子的大量快速复制。在印刷机出现之前，一个修道士手抄完成一本《圣经》大约需要一年的时间，当时牛津大学的藏书量仅为 122 本，每本书的价值相当于一个农场或葡萄园，而古登堡印刷机在投入生产的第一年就印制了 180 本《圣经》。印刷机的发明有力地推动了欧洲的转型、文艺复兴和宗教改革。到 1501 年，即印刷机发明 50 年后，在欧洲有 2.7 万～3.5 万本书被复制印刷，总量超过 1000 万份。

到了近代，文献信息已不仅仅局限于图书，其他类型的文献，如期刊、专利、报纸等陆续出现，信息的数量也呈现快速增长态势。普莱斯（Price）在《巴比伦以来的科学》（*Science Since Babylon*）一书中曾以期刊为例，验证了期刊的数量每隔 50 年增加 10 倍，这被称为“普莱斯指数增长规律”。联合国机构世界知识产权组织（World Intellectual Property Organization，WIPO）发布年度报告显示，2021 年世界各地的创新者通过《专利合作条约》（*Patent Cooperation Treaty*，PCT）提交了 277 500 件国际专利申请，这是有史以来的最高数量。其中，中国专利申请人通过 PCT 提交的国际专利申请达到 69 540 件。这些专利文献，对于任何一个人，穷其一生恐怕也难以阅读完。

1.3.2 信息的爆炸

随着计算机技术（1946 年，ENIAC 问世）和网络技术（1969 年 Internet 的前身 ARPANET 问世、1989 年 Internet 问世）的出现，信息呈现爆炸增长的趋势。归纳起来，大概有两个因素促使了信息爆炸的出现。

一是信息技术的发展。正如在前文所提及的，技术的发展有可能是信息载体的变化，也有可能是印刷技术的进步，在信息的增长中处于决定性的地位。信息技术和计算机技术的发展使得信息的存储和复制，网络技术的发展则使得信息的传输和传播变得比以往任何时候都更加便利。在 21 世纪初，就有人预言，21 世纪头三年所产生的新信息会超过过去此前人类历史所积累的信息总和。很显然，这一预言已然成为事实。

二是信息生产方式的变化。正如我们所熟知的，在以往印刷时代，甚至在 Web 1.0 时代，信息往往只被控制在少数权威手里，多数大众只能被动地接收这些信息，无论是图书还是报纸，甚至是门户网站均是如此。然而，随着网络技术的发展，信息生产方式也在发生变化。当前，依赖于群体智慧的用户产生内容（user generated content，UGC）成为信息爆炸的新起点，在博客、论坛、微博、微信等各式各样的社会网络服务（social network services，SNS）上，人人皆可在网络上进行信息的生产，人人皆是信息源，信息由此呈现大爆炸式的增长。

1.3.3 信息矛盾的转变

物以稀为贵。在古代，由于数量稀少，图书是非常珍贵的，通常普通百姓要想读书，较为困难。一般而言，我国图书主要收藏于皇家藏书机构和宗教藏书场所。显然，普通人是无法进入皇家藏书机构读书的，宗教藏书场所往往成为贫寒学子读书的重要场所。我们经常在影视剧中看到书生在寺庙中借读大致就是这个原因（当然还有其他原因，如寺庙大多比较清静等）。此外，一些富贵人家也可能建有私人藏书楼，比如大家所熟知的宁波范钦的“天一阁”就是非常著名的私人藏书楼。这些私人藏书楼规矩甚多，不太可能会让无关人士进入读书。甚至在封建礼教的约束下，女性也曾不被允许进入藏书楼。

2300 多年前，亚里士多德（Aristotle）和弟子们在林荫道上散步。走到一棵树下，亚里士多德若有所思地发了一声感叹：“无书可读啊！”弟子们便问道：“难道图书馆里的书您都读完了？”亚里士多德回答说：“早在自己收门徒之前，图书馆里的书就已经全部读完了，连可以搜集到的外国书籍也都已经倒背如流。”在发了一通感叹之后，这位“古代最博学的人”一边做老师，一边动手写下了《物理学》《诗学》《尼各马可伦理学》《政治学》《形而上学》《工具论》等世界名著。不管此故事是否属实，大体可认为，当时的主要矛盾主要是信息远远无法满足大众的信息需求。

信息的快速增长和爆炸，毫无疑问，大大满足了人们的信息需求。今天，信息所面临的矛盾显然已经发生根本性的变化。信息已经是真正的“浩如烟海”，人们穷其一生，也无法读完一个小型图书馆的藏书，更不用说互联网上如此庞杂的信息。

此外，由于信息生产方式的变化，大量的虚假信息充斥网络，也成为当代面临的重要社会问题。来自社交媒体上模糊的甚至匿名的信息，只负责“聚合”而不负责筛选的门户网站和视频网站上充斥着未被加工过的碎片化的信息，广播上和电视、网络上专家根据不确定的消息发表的推测和个人观点等，比比皆是。信息论的提出者申农曾把信息定义为“不确定性的消除”，但是在这个特定情境下，恰恰是信息增加了不确定性。

今天，要想获得自己所要的有用的信息正变得愈发艰难。正如约翰·奈比斯特（John Naisbitt）在《大趋势》一书中感慨的：“我们淹没在信息中，但是却渴求知识。”如何从真正浩如烟海的信息中找出所需要的信息正成为人们最为关注的问题，而这也是信息检索要解决的主要问题。德国柏林图书馆门前有这样一句话：“这里是知识的宝库，你若掌握了它的钥匙，这里的全部知识都属于你。”显然，信息检索就是这把钥匙。

1.4　信息检索活动

1.4.1　我国古代的信息检索活动

信息检索作为一种实践活动由来已久。在我国古代，对图书进行分类以便系统性地查找和阅读是较为通用的做法。早在西汉时期，学者刘向、刘歆父子便受命主持了我国历史上第一次大规模整理群书的工作。在每一部书整理完毕时，刘向便撰写一篇叙录，记述这部书的作者、内容、学术价值及校雠过程。这些叙录后来汇集成了一部书，即我国第一部图书目录书《别录》。刘向死后，其子刘歆继续整理群书，并把《别录》各叙录的内容加以简化，把著录的书分为六略，即六艺略、诸子略、诗赋略、兵书略、术数略、方技略，再在前面加上一个总论性质的“辑略”，编成了我国第一部分类目录书《七略》。

《七略》以学术性质作为分类标准，首次展示了我国古代的图书分类方法。同时，在著录上也确立了较为完整的著录方法，除编有内容提要外，还利用了“互见法”和“分析法”。《七略》创立出的分类法和著录法对我国图书馆目录的发展产生了深远影响。该书早已亡佚，但它的基本内容都被保存在班固的《汉书·艺文志》中，因此，《汉书·艺文志》成为今存最早的古籍分类目录。

汉代以后，各种官修，私撰的古籍分类目录不断涌现，分类方法也不断改进。西晋荀勖的《晋中经簿》将六略改为四部，即甲部录经书（相当于六艺）、乙部录子书（包括诸子、兵书、数术、方技）、丙部录史书、丁部为诗赋等，奠定了四部分类的基础。东晋李充所编的《晋元帝书目》根据当时古籍的实际情况，将史书改入乙部，子书改入丙部，这样，经、史、子、集四部分类已略具雏形。四部体制的最终确立，体现在《隋书·经籍志》中，这部实际上由唐初名臣魏征所编的目录，正式标注经、史、子、集四部的名称，并进一步细分为 40 个类目。

四部体制的分类只能大致地对图书进行分类和简单的检索，还不能完全深入图书的内容。为了满足对于图书内容快速的查找和阅读，古人又创造性地发明了“类书”这一

检索工具。类书之名，最早见于后晋刘昫撰写的《旧唐书》，到了北宋，欧阳修等撰《新唐书》时改“类事”为“类书”，从而定下“类书”之名。

简而言之，类书就是一种分门别类地汇辑资料，并按门类、字韵等编排以备查检的工具书。由于类书收录资料的全备性，许多人又将类书称为古代的百科全书。但严格意义上的类书具备两个特点：一是“分门别类”，即对搜集来的资料分类编排，把同类资料排列在一起；二是“录而不作”，即纂辑罗列现有的资料，而非编者自己的论述或考辨，编者最多只是在某些资料前后加几句简单的按语，这一点与现今的百科全书有着非常大的区别。

类书按采集资料范围来划分，有综合性和专门性两大类。

（1）综合性类书：收录范围广，覆盖面宽。大型官修类书多是综合性的，如《太平御览》《古今图书集成》等。

（2）专门性类书：仅采集某一方面的资料，但就某个特定领域来说，其资料甚为全面丰富，如《通典》《册府元龟》《格致镜原》《太平广记》等。

以下为《太平御览》“学部·卷一”中对“叙学”这一词条的辑录（节选）。

○ 叙学

《易·文言》曰：学以聚之，问以辨之。

《白虎通》曰：学以言觉也，觉悟所不知也。

《论语·为政》云：子曰：“学而不思则罔，思而不学则殆。”

又曰：卫灵公曰：“君子谋道，不谋食。耕也，馁在其中；学也，禄在其中。”

又曰：生而知之者，上也；学而知之者，次也；困而学之，又其次也；困而不学，民斯为下矣。

《礼记·学记》曰：君子之于学也；藏焉，修焉，息焉，游焉。

又曰：善问者如攻坚木，先其易者，后其节目。

又曰：学，不学操缦，不能安弦；不学博依，不能安诗。

又曰：善学者，师逸而功倍，又从而庸之；不善学者，师勤而功半，又从而怨之。

又曰：凡学之道，严师为难。师严然后道尊，道尊然后民知敬学。

又曰：善待问者如撞钟，叩之以小者则小鸣，叩之以大者则大鸣。

又曰：玉不琢不成器，人不学不知道，是故古之王者，教学为先也。

《国语》曰：文公问元帅于赵衰，曰：“郗穀可，行年五十矣，守学弥惇。夫学，先王之法，义之府也。”

又曰：范献子聘于鲁，问具、敖山，鲁人以乡对。献子曰：“不为具、敖乎？”对曰：“先君献武之讳也。”献子归，遍戒所知曰：“人不可以不学。吾适鲁而名其二讳，为笑焉。惟不学也。人之有学，犹木之有枝叶，犹庇荫人，而况君子乎？”

《家语》曰：子路见孔子。孔子问曰：“何好？”曰：“好长剑。”子曰：“以子之能加之以学，岂可及乎！”子路曰：“学岂有益哉？”子曰：“狂马不释策，操弓不反檠。木受绳则正，人受谏则圣。受学重问，孰不顺成？”子路曰：“南山有竹，不揉自直，斩而用之，达于犀革，何学之为？”孔子曰：“括而羽之，镞而砺之，其入不益深乎？”子路拜曰：“敬受命。”

读者按照目录找到“叙学”词条，便可以查阅到相关书籍中对此的论述。

总的来看，我国古代已出现了信息检索的雏形，但信息检索活动仍没有作为一种专门技能。随着时代的发展和信息的不断增加，在近代，信息检索逐渐发展成为一种专门的技能。

1.4.2　近代手工信息检索活动

信息检索作为一种专门技能，其历史可以追溯到图书目录、索引和文摘等检索工具产生的时代。

目录是读者最为熟悉的，它是按次序编排以供查考的图书或篇章的名目，现代图书一般都提供相应的目录。

所谓索引，旧称通检、备检或引得（index），最早出现于西方，主要是中世纪欧洲宗教著作的索引。18 世纪以后，西方开始有主题索引。中国的索引出现较晚，一般认为，明末傅山所编的《两汉书姓名韵》是现存最早的人名索引。索引组成的基本单位是索引款目。款目一般包括索引词、说明或注释语、出处 3 项内容，一般按照某一顺序进行有序化编排。索引的主要功能是为方便人们准确、迅速地获得文献资料提供线索性指引。常见的索引主要有报刊论文资料索引、文集篇目索引、语词索引、文句索引、关键词索引、专名索引、主题索引等。图 1.4-1 为《隋书人名索引》（邓经元编，中华书局出版，1979 年）的索引款目。


0021₁—0022₃ 鹿廘竟盧彥齊

0021₁ 鹿

71鹿愿
64/1521
65/1529
67/1577
80/1802

竟陵郡公　見楊璨
竟陵郡公　見楊坦

0021₇ 盧

74盧陵王　見蕭績

0022₂ 彥

22/626
22/628
22/638
23/667
24/676
25/704
35/1094

图 1.4-1　人名索引

文摘的历史也同样悠久。早在 1665 年，法国就出版了西方世界第一本非严格意义上的学术期刊《学者报》（*Le journal des Scavans*）。该期刊主要目的是对当时在欧洲印刷的主要书籍进行编目和简要描述，并提供当前学术著作的可读性和批判性说明，实际上是一种摘要性的期刊。《学者报》的出版也奠定了期刊这种独特的出版形式，仅 3 个月后，第一本严格意义上的期刊《伦敦皇家学会哲学论坛》（*Philosophical Transactions of the Royal Society of London*）就付梓了，此后西方陆续创办了数量可观的非摘要性期刊。到了 19 世纪初叶，随着期刊数量越来越多，纯粹的以摘录文献信息的文摘型期刊开始出现。文摘型期刊系统收集文献的摘要信息，为读者提供集中阅读和多角度的查询。

1830 年，柏林科学院在柏林和莱比锡出版了著名的文摘刊物《药学总览》（*Pharmaceutisches Central-blatt*）。《药学总览》的问世，标志着专供检索使用的文摘刊物从一般刊物中分离出来，被单独编辑出版。这一标志性事件常常被认为是手工信息检

索工作的正式开端，从那时起，信息检索经历了巨大的变化与发展。伴随着独立的文摘性刊物的出版和使用，索引工作也得到了很大的发展，并且逐渐转向为文摘刊物服务，与文摘刊物紧密结合在一起。索引与文摘的结合，使各种手工检索工具的查询功能得到不断提高和完善。

1946 年，世界上第一台电子计算机问世，到 20 世纪 70 年代初期，基于计算机技术的联机信息检索开始步入商业应用。在这段时间里，手工检索仍处于主流地位并达到其发展的高潮。进入 20 世纪 70 年代以后，在信息检索最为发达的英美等国家，手工检索便逐渐退出主流地位，取而代之的是现代计算机检索。

1.4.3 现代计算机检索阶段

在纸质环境下，索引检索工具尽管操作简单，但显然效率很低。随着文献信息的不断增长，传统的利用印刷型索引工具进行手工检索的方式已不能适应。1945 年夏天，美国科学家布什（Bush）在《大西洋月刊》（*Atlantic Monthly*）上发表《诚若所思》（*As we may think*）这一经典论文，在其中他设想了一种叫作“Memex”的机器，“想象一个未来的设备……人们可以在其中存储所有的书籍、记录和通信信息，并且可以以极高的速度和灵活性与这种机械设备进行互动咨询。这种设备是对人本身记忆的直接扩大和补充”。布什的这一设想开创了现代信息检索的先河。随着 1946 年第一台计算机问世，人们开始探索将计算机技术应用于信息检索。计算机信息检索主要经历了早期的脱机批处理检索、联机实时检索、光盘检索等发展阶段。

1. 脱机批处理检索时期

早在 20 世纪 50 年代初期，美国麻省理工学院的研究人员就开始利用计算机进行代码化文摘检索的可行性研究试验，将输入计算机的信息（文献的题录、文摘等）全部存储在磁带上，检索提问主要采用“穿孔卡片”和“穿孔纸带”作为存储媒介。

1954 年，美国海军兵器中心实验室利用 IBM-701 型计算机（电子管计算机），将文摘号和少量标引词存储在计算机中，进行相关性比较后输出检索结果文献号，由此诞生了世界上第一个计算机文献信息自动化检索系统。1958 年，美国通用电器公司将其加以改进，输出结果增加了题名、作者和文献摘要等项目。此后，美国化学文摘服务社于 1964 年建立了文献处理自动化系统。同年，美国国立医学图书馆建立了医学文献分析与检索系统。

这一阶段通常被称为脱机批处理检索，其特点是不对一个检索提问立即作出回答，而是集中大批提问后进行处理，且进行处理的时间较长，人机不能对话，因此检索效率往往不够理想。但是，脱机批处理检索中的定题服务对于科技人员却非常有用，定题服务能根据用户的要求，先把用户的提问登记入档，存入计算机中形成一个提问档，每当新的数据进入数据库时，就对这批数据进行处理，将符合用户提问的最新文献提交给用户，可使用户随时了解课题的进展情况。

2. 联机实时检索与光盘检索时期

进入 20 世纪 60 年代后期，随着计算机技术的不断进步，第三代计算机集成电路计

算机开始出现，与此同时，高密度海量随机存储器、硬磁盘及磁盘机问世并投入使用，信息检索进入人机对话式的联机实时检索时期，表现为：用户可以通过检索终端设备与检索系统中心计算机进行人机对话，从而实现对远距离之外的数据库进行检索的目的，即实现了联机信息检索。

联机实时检索时期取得了一系列更加具有突破性的发展成就，其中最突出的表现是一批联机检索服务系统的创建和应用。例如，1965 年美国系统发展公司（SDC）建立的 ORBIT 系统（书目情报分析联机检索系统）、1966 年美国洛克希德（Lockheed）公司研制开发的 DIALOG 系统、1973 年建成的 ESA-IRS 系统（欧洲航天局信息检索系统）等。

脱机批处理系统主要采用顺排文档检索技术，其批处理方式虽然比手工检索便利了很多，但用户还是不能与系统进行实时对话，也不能对检索策略进行及时调整。为克服脱机系统的这些缺陷，上述联机系统广泛使用了倒排文档检索方式，另外还对布尔逻辑检索、截词检索、位置检索等检索技术加以试验和运用。遗憾的是，联机检索虽然能够实现信息的实时更新，但由于当时的网络通信技术还不是非常成熟，国际联机检索费用过高，导致这一检索过程只能为少数专业机构和用户所采用，其普及率并不高。

在联机检索快速发展的同时，随着 20 世纪 80 年代光盘技术的诞生，基于光盘的信息检索技术也日趋成熟并得到光盘应用。1985 年，第一张正式的 CD-ROM 数据库产品 BIBLIOFILE（美国国会图书馆机读目录）问世。1987 年，DIALOG 系统也开始推出其光盘检索服务（Ondisc），由此，光盘数据库产品及其服务在信息检索领域逐步开展并逐渐流行起来。作为一种新型的存储设备，光盘与计算机（尤其是个人计算机）的结合，为人们提供了一种崭新的检索环境和服务模式。操作方便、不受通信线路的影响与制约等特点，使光盘检索深受专业用户青睐，并逐渐与联机实时检索形成了互相补充、互相竞争的发展格局。

1.4.4　网络化信息检索活动

进入 20 世纪 90 年代，起始于 1969 年的 Internet 技术在经历了早期军事、科技与教育领域里的试验与应用推广之后，开始步入社会化的商业应用。信息检索的主流平台也迅速转移到以万维网为核心的网络应用环境中，信息检索开始步入网络化检索时期。网络化信息检索是联机实时检索的延伸，在保持信息更新速度的同时，也克服了联机实时检索费用高昂的弊端，网络信息检索的用户逐渐由信息检索从业人士向普通研究人员和大众发展。特别是搜索引擎的出现，更是将网络信息检索的概念普及。

在这一时期内，信息检索的网络化主要体现在两个方面。一是各类数据库检索系统和联机检索系统逐渐将自己的服务转移到具有分布式网络结构特性的 Web 平台上，如 DIALOG 系统、INSPEC 系统、化学文摘等。目前，我们所用的检索工具几乎都是基于 Web 平台。二是基于 Web 的搜索引擎系统层出不穷，如谷歌（Google）、百度、Bing（必应）等，技术发展迅速，应用日趋广泛。