普通高等教育“十一五”国家级规划教材
普通高等教育信息管理类专业规划教材

第2版

信息存储与检索

INFORMATION STORAGE AND RETRIEVAL

王知津　主编

机械工业出版社
CHINA MACHINE PRESS

普通高等教育“十一五”国家级规划教材
普通高等教育信息管理类专业规划教材

信息存储与检索

第2版

主　编　王知津
副主编　李　培　于晓燕
参　编　（按姓氏笔画排序）
陈芳芳　赵　洪　徐　芳
蒋伟伟　景　璟　樊振佳

机械工业出版社

本书供高等院校信息管理类专业学生学习信息检索专业课使用，有别于旨在向大学生普及信息检索方法的信息检索与利用类教材。本书内容涉及信息检索的原理、方法、技术、系统以及相关的网络知识等，共分9章：绪论、信息检索模型、文本信息存储与检索、多媒体信息存储与检索、Web信息存储与检索、并行与分布式信息检索、人工智能与自然语言检索、用户界面与可视化、信息检索评价与实验。

本书内容丰富，深入浅出，力图将计算机技术与信息检索紧密结合起来，具有信息检索专业性质，属于侧重“技术”的教材。本书不仅适合信息管理类专业学生使用，还可作为高等院校计算机类专业师生的参考书，对于从事信息检索系统、数据库以及网站开发、设计的实际工作者来说，也是一本较好的参考书。

图书在版编目(CIP)数据

信息存储与检索/王知津主编. —2版. —北京：机械工业出版社，2015.9
普通高等教育“十一五”国家级规划教材　普通高等教育信息管理类专业规划教材
ISBN 978-7-111-51235-6

Ⅰ.①信…　Ⅱ.①王…　Ⅲ.①信息存贮—高等学校—教材　②情报检索—高等学校—教材　Ⅳ.①TP333　②G252.7

中国版本图书馆CIP数据核字（2015）第189326号

机械工业出版社(北京市百万庄大街22号　邮政编码100037)
策划编辑：易　敏　责任编辑：易　敏　刘　静
责任校对：赵　蕊　封面设计：陈　沛
责任印制：李　洋
北京圣夫亚美印刷有限公司印刷
2016年1月第2版第1次印刷
184mm×260mm　·18.25印张·428千字
标准书号：ISBN 978-7-111-51235-6
定价：38.80元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
电话服务　　网络服务
服务咨询热线：010-88379833　机 工 官 网：www.cmpbook.com
读者购书热线：010-88379649　机 工 官 博：weibo.com/cmp1952
　　教育服务网：www.cmpedu.com
封面无防伪标均为盗版　　金　书　网：www.golden-book.com

第2版前言

本书第1版自2009年出版以来，受到广大读者的好评和欢迎，许多学校都采用本书作为教材或主要教学参考书，先后多次印刷。

随着计算机技术和网络技术的迅速发展，信息存储与检索的理论研究与实践活动也发生了许多变化。因为这些发展，本书有的内容需要修改，有的内容需要删除，有的内容需要增加。部分读者和用书教师也对本书的内容提出了建议和要求。在这种情况下，我们对第1版进行了修订。

此次修订的指导思想是：延续原书的写作风格和特色，保留原书的主要框架，只对个别结构进行微调；补充新内容，使其更加充实和饱满；删除个别陈旧过时的内容，突出基本理论、方法和技术；修改不十分恰当的内容及其文字表述，使之更加科学、完善和流畅。此外，我们还补充和更新了个别图、表、思考题以及参考文献。

本书各章节的修订者及具体分工如下：王知津（第1章）、赵洪（第2章）、陈芳芳（第3章第1~5节、第7节以及第8章）、景璟（第3章第6节）、徐芳（第4章）、蒋伟伟（第5章）、李培（第6、7章）、樊振佳（第9章）。

信息存储与检索是一个紧跟计算机技术及网络技术不断发展和变化的领域，新理论、新方法、新技术层出不穷，虽然我们尽了最大努力兼顾该领域的最新发展与学校教学的基本规律，并将其有机结合起来，但受学识、水平和能力的限制，缺点、疏漏在所难免，恳请各位专家、学者以及广大读者不吝赐教、指正，及时反馈意见和建议，以便将来再次修订时予以更正、补充和完善。

王知津

第1版前言

如果在30年前提起“信息检索”，恐怕没有多少人听说过，因为那个时候信息检索还远离广大最终用户，“信息检索”只是作为专业工作者的专用术语而存在。然而，这并不意味着广大最终用户不需要信息检索，恰恰相反，人们在学习、工作和生活的各个领域里，每时每刻都在需求信息和利用信息，只不过绝大多数的检索操作都不是用户亲自进行的，而是由专职人员代替完成的。然而，近几十年来，计算机技术、通信技术和网络技术飞速发展，特别是互联网延伸到世界的各个角落，成为一种大众工具，使信息检索也发生了翻天覆地的变化。今天的“信息检索”已经不是什么新鲜事，已变成了大多数人耳熟能详的常用术语。

信息检索是信息管理领域的核心部分。现代信息检索已经脱离了原来的人工操作方式，而与现代信息技术紧密结合起来，从而进入了一个崭新的历史发展阶段。自20世纪50年代初提出“信息检索”这个概念以来，历经半个多世纪的发展和建设，信息检索已成为一门新兴的交叉学科呈现在人们面前。信息检索已经逐渐形成了包括自身的理论、方法、技术和应用领域在内的完整的学科体系，尽管还存在一些没有解决或没有完全解决的课题，但这并不影响它沿着自己的既定方向继续前进。

目前，环顾国内外，关于信息检索的教材数量众多。仅就国内而言，绝大多数此类教材属于“方法”类，主要供在校大学生学习、掌握和运用检索方法，强化其利用信息的技能和技巧，带有普及性质。还有少数此类教材属于“技术”类，主要供高等学校信息管理类专业的学生使用，旨在使学生深入了解信息检索的原理、技术、系统以及相关的网络知识等，相较而言更专业。本书属于后者。

2005年，我们曾翻译出版了《现代信息检索》（机械工业出版社出版）一书。该书主要从计算机专业角度出发，将计算机技术与信息检索紧密结合起来进行介绍，但由于文化和教育背景不同，还不能完全适合我国学生。为此，出版社鼓励我们重新编写一本更加适合我国学生的信息检索教材，这成为我们编写本书的巨大动力。此后，本书被教育部列入普通高等教育“十一五”国家级规划教材，也得到了南开大学教材建设立项资助。

本书大体上分为4个部分共9章。第一部分是信息检索理论，包括第1、2章，主要介绍信息检索和信息检索系统的基本概念、原理、类型、结构及各种数学模型。第二部分是基本的信息存储与检索，包括第3~5章，重点介绍文本信息、多媒体信息和Web信息的存储与检索。第三部分是信息存储与检索的提高，包括第6~8章，着重介绍并行与分布式信息检索、智能信息检索、用户界面设计及信息检索可视化。第四部分的第9章是信息检索的评价，侧重介绍信息检索的相关性理论以及评价指标、方法与实验。

本书的编写思路和大纲由王知津提出，经集体反复讨论和修改后确定。各章节的编写者及具体分工如下：王知津（第1章）、赵洪（第2章）、陈芳芳（第3章第1~5节、第7节）、于晓燕（第4、8章）、江力波（第3章第6节、第5章第1~3节）、张收棉

（第5章第4节）、李培（第6、7章）、樊振佳（第9章）。全书的初审由于晓燕和李培完成，王知津负责终审和定稿。

在本书的编写过程中，我们参考和借鉴了大量的中外文书刊资料，由于篇幅所限，未能一一列出，在此对所有参考文献作者表示诚挚的谢意。正是这些参考文献作者的前期工作为本书的完成奠定了基础，并为我们提供了强大的写作动力和丰富的创新素材。本书得以顺利完成，与机械工业出版社易敏编辑所给予的大力支持、鼓励、指导、帮助和建议是分不开的，在此，我们一并表示感谢。

虽然我们尽了自己最大的努力编写好本书，但信息检索毕竟是一个快速发展和不断更新的领域，限于编者的学识、水平和能力，缺点、疏漏在所难免，恳请各位专家、学者和广大读者不吝赐教、指正，以便在本书修订时加以补充、更正和完善。

我们制作了与本书配套的PPT课件，使用本书作教材授课的教师可向出版社编辑索取（yimin9721@163.com）。

王知津

目　录

第1章 绪 论

【本章提示】 本章为信息存储与检索提供一个概貌，为后续各章的展开打下基础。本章主要阐述信息检索的概念、原理和类型等基本理论，介绍信息检索系统的概念、类型、物理结构和逻辑结构，讨论信息检索的研究内容、相关学科、产生和发展以及现状与未来趋势。要求重点掌握信息检索基本理论和信息检索系统两大部分，对于信息检索的研究现状与趋势可做一般了解。

1.1 信息检索基本理论

1.1.1 信息检索的概念

“信息检索”（Information Retrieval，IR，我国早期译为“情报检索”）一词最早出现于1952年，由美国学者穆尔斯（C. W. Mooers）提出，从1961年开始在学术界和实践领域中得到广泛的应用。信息检索这一概念首先假设包含相关信息的文献或记录已经按照某种有助于检索的顺序组织起来。信息检索就是对信息项进行表示、存储、组织和存取的全过程。对信息项的表示和组织应该能够为用户提供其感兴趣的信息的方便存取。遗憾的是，对用户信息需求进行全面而准确的描述不是一件轻而易举的事情。例如，在万维网（或者就是Web）环境中考察以下假设的用户信息需求：

找出包含能满足以下两个条件的有关某一学院网球队相关信息的所有网页（即文献）：①该网球队隶属于美国的一所大学；②该网球队参加过美国大学生体育协会（NCAA）举办的网球锦标赛。为了保证查找结果的相关性，检索到的网页必须包括该网球队在过去三年里在全国比赛中的名次及其教练的电子邮箱、地址或电话号码等信息。

显然，在目前的Web搜索引擎界面中，人们不可能直接采用这种对用户信息需求进行完整描述的方式来检索信息，用户必须首先将这些信息需求转换为搜索引擎（或IR系统）能够处理的查询式来查询。这种转换以其最普遍的形式生成一组关键词（或索引词），而这些关键词能够对用户信息需求的描述进行概括。

20世纪90年代以前，知道“信息检索”这个术语的人还不多。随着因特网的形成、发展和普及，信息检索才被越来越多的人所知、所用。就信息检索这个概念而言，不同的使用者对它有着不同的理解和解释，大体上可以分为两类：

第一类是广义的。对于专门从事信息检索及其系统的研究、开发和设计的少数人来说，“信息检索”的完整含义是“信息存储与检索”（Information Storage and Retrieval，ISR）。也就是说，把“信息检索”当作“信息存储与检索”的简称。这里所谓的信息检索，包括存储和检索两个过程。信息存储是指将有用信息按照一定的方式组织和存放起来；信息检索是指当用户需要这些信息时，再把它们从存放的地方查找和提取出来。因此，对于广义的信息检索来说，存储和检索缺一不可。本书采取信息检索的广义用

法，这就要求不仅要知道如何检索，也要知道如何存储，因为如何存储决定了如何检索。

第二类是狭义的。对于普通信息用户来说，在大多数情况下，“信息检索”可以用英文 Information Searching 来表达，其准确含义是“信息查询”或“信息搜索”。也就是说，所谓信息检索，是指按照一定的方式从现有的信息集合或数据库中，找出并提取所需要的信息。可见，狭义的信息检索仅指检索这一个过程，而不关心信息是如何存储的。

1.1.2 信息检索的原理

如上所述，广义的信息检索包括存储和检索两个过程，其基本原理可以用图 1-1 表示。

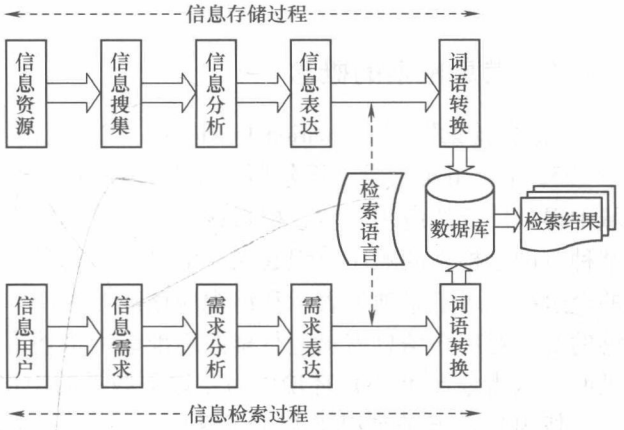

图 1-1 广义的信息检索的基本原理

在存储过程中，专门负责信息检索系统和数据库建立的人从各种各样的信息资源中，搜集有用信息，对有用信息进行主题内容的分析，找出能够全面、准确表达该信息主题内容的概念，借助于检索语言（通常是检索词表）把分析出来的概念转换成检索系统所采用的词语（在自然语言检索系统中，直接使用自然语言而不需要转换），再按照一定的规则和方式将这些有用信息组织成可供检索用的数据库，并存储在一定的介质上。这就是说，存储过程的实质是对信息进行标引，以形成信息特征标识，为检索过程提供入口和路径。

信息用户在工作、学习和生活中产生了信息需求，为了检索并获取自己所需要的信息，用户必须对自己的需求进行主题内容的分析，找出能够全面、准确表达该需求主题内容的概念，也要借助于检索语言（通常是检索词表）把分析出来的概念转换成检索系统所采用的词语，（在自然语言检索系统中，直接使用自然语言，而不需要转换。）再按照一定的检索规则和方式，制定检索策略，构造检索式，从数据库中查找并获取自己所需要的信息，最后，输出检索结果。当然，检索的全过程还应当包括对检索结果进行评价、反馈，或许还要重新制定检索策略，重新构造检索式，反复进行检索，直至检索出满意的结果为止。这就是说，检索过程的实质是对提问（从用户的信息需求中提炼出来）进行标引，以形成提问特征标识，然后按照存储过程所提供的入口和路径，从信息集合中查获与提问标识相符合的信息子集。可见，检索过程是存储过程的相反过程或逆过程。

在现实中，把用户的复杂信息需求与近乎无限的信息集合进行直接的比较和匹配是不现实的，取而代之的可行方式是对两者的简约代表进行比较和匹配，即间接比较和匹

配。因而，信息检索原理的实质就是提问特征标识与信息特征标识的比较和匹配，这种比较和匹配代表着信息需求与信息集合之间的比较和匹配。比较和匹配的结果，如果两者一致，则检索命中或检索成功；如果两者不一致，则检索未命中或检索未成功。

从用户的角度来看，信息检索原理的核心是用户所使用的检索词或者由检索词和运算符所组成的检索式与数据库中的检索词及其逻辑关系之间的比较和匹配机理。从集合论的观点来看，检索过程是对信息集合进行选择或划分的过程，选择或划分的依据就是一系列检索条件。由于存储过程和检索过程都不具有唯一性，所以，对于同一个信息需求或检索课题来说，检索方式也是多种多样的。

从图 1-1 可以看出，信息存储和信息检索有两个交汇处：一个是直接的，即表达信息主题内容的词语与表达需求主题内容的词语之间进行对比的交汇；另一个是间接的，即通过检索语言进行沟通，确保把存储用词和检索用词都统一到同一个检索语言体系中（对于自然语言检索系统来说，不存在存储与检索的间接交汇处）。

由此可见，信息存储和信息检索的直接交汇处是至关重要的，由此形成了信息检索的一致性匹配作用机理，如图 1-2 所示。

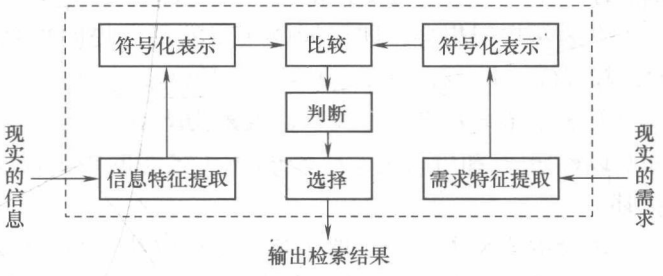

图 1-2 信息检索的一致性匹配作用机理

信息检索的一致性匹配作用机理包括五个机理。

（1）提取机理。从现实的信息和现实的需求中提取出能够揭示特定信息和特定需求的语法特征和语义特征。这些特征可以归纳成内容（内部）特征和形式（外部）特征，前者包括特定信息和特定需求的类别（如学科、专业）、主题等，后者包括信息和需求的名称（题名）、作者（责任者）、时间、编号等。

（2）表示机理。用适当的符号表示信息和需求的各种特征。符号是广义的，可以是文字、数字和符号，也可以是图形、图像、视频和音频。比如，用分类号表示信息和需求的类别，用关键词表示信息和需求的主题。

（3）比较机理。在检索项类型（如题名、作者、分类、关键词）相同的情况下，对代表特定信息的特征符号与代表特定需求的特征符号进行对比。比较的实质是相似性比较或一致性比较，即包括完全一致、部分一致和不一致，也包括等于、不等于、大于、小于。比如，对于两个词或词组来说，它们可以是完全一致、前方一致、后方一致、中间一致；对于两个编号来说，它们可以是相等、大于、小于。

（4）判断机理。在比较的基础上，对信息是否符合需求以及符合的程度加以判断。两者相符合的信息被检索出来（命中），不相符合的信息被拒绝（不命中）。从符合程度来看，可以是完全符合，也可以是部分符合。在部分符合中，还可以进一步细化。原则上，凡是符合特定检索所规定的比较条件和一致条件的信息，都应该是符合需求的，尽管它们符合的程度有所不同。

（5）选择机理。对于检索出来的结果，按照一定的标准加以选择，带有推荐首选或

着重使用的意义。选择的实质就是排序，排序有多种标准和方法，如相关度、权值和（加权检索）、时间（新颖性）、重要作者或单位等。

信息检索的一致性匹配作用机理的实质是简化现实的信息和现实的需求之间的匹配。把内容与形式都非常复杂的信息简化成信息特征的符号化表示，再把内容与形式都非常复杂的需求也简化成需求特征的符号化表示，将这两个非常简单的特征符号化表示进行比较、判断和选择，从而变复杂为简单，化模糊为清晰，大大提高了匹配效率。然而，这种简化也会带来一些弊病，造成误检和漏检。如何解决和避免这些问题，已经成为信息检索领域的重要研究课题。

按照信息检索原理，可以用如下代数结构来描述任何一个信息检索系统：

$$I = < T, D, Q, F, R >$$

式中，T 表示词语集合，$T = \{t_1, t_2, t_3, \cdots, t_n\}$，它代表某一检索系统或数据库的词语空间（或控制空间、属性空间、标引空间），是标引的结果和检索的依据，用于规范和控制标引和检索。

D 表示记录集合，如文献集合代表某一检索系统或数据库的文献空间，$D = \{d_1, d_2, d_3, \cdots, d_m\}$，$d_i \in D$，都有$(t_{i1}, t_{i2}, t_{i3}, \cdots, t_{in}) \in d_i$，且$(t_{i1}, t_{i2}, t_{i3}, \cdots, t_{in}) \subseteq T$。

Q 表示提问集合，即用词语表示的用户提问集合，$Q = \{q_1, q_2, q_3, \cdots, q_k\}$，$q_j \in Q$，都有$(t_{j1}, t_{j2}, t_{j3}, \cdots, t_{jn}) \in q_j$，且$(t_{j1}, t_{j2}, t_{j3}, \cdots, t_{jn}) \subseteq T$。

F 表示 $D \times T$ 的二元关系，表示为 $F = \{ < d, t, \mu(d, t) > \}$，$F$ 描述的是标引关系，以确定 d_i 和$(t_{i1}, t_{i2}, t_{i3}, \cdots, t_{in})$ 之间的相关程度，而 μ 值是这种相关程度的量化描述。

R 表示 $T \times Q \rightarrow \{d\}$ 的关系，表示为 $R = (t, q, d, \theta(t, d, q))$，$R$ 描述的是检索关系，与 F 是对称的；θ 表示包含$(t_{i1}, t_{i2}, t_{i3}, \cdots, t_{in})$ 的任意一个文献 d_i 与包含$(t_{j1}, t_{j2}, t_{j3}, \cdots, t_{jn})$ 的任意一个提问 q_j 之间相关程度的标准，即检索算法。检索算法可能是一个值（检索词），也可能是一个公式（检索式），它所产生的检索结果集是 $\{d\}$，它是 D 的子集，能够满足该检索算法 θ 所确定的、Q 所提出的以及 T 所规范的标准。该检索算法描述了标引方式 F 和检索方式 R 两个方面，F 和 R 是一个事物的两个方面，即什么样的 F 决定了什么样的 R，什么样的 R 来源于什么样的 F。

可以用数学空间的概念来描述标引和检索。对于标引来说，设 T 为词语空间，则 $T = \{t_1, t_2, t_3, \cdots, t_n\}$，这是一个 n 维的向量空间。设 D 为文献集合，则 $D = \{d_1, d_2, d_3, \cdots, d_m\}$，对于任意一个 d_i来说，在词语空间中都有一个确定的向量与之对应。也就是说，在 T 之上，用二元关系 F 对 d_i进行标引后，d_i就成为该 n 维 T 空间中的一个文献向量，而 m 个文献向量就构成了文献空间。类似地，设 Q 为提问集合，则 $Q = \{q_1, q_2, q_3, \cdots, q_k\}$，对于任意一个 q_j来说，在词语空间中都有一个确定的向量与之对应。也就是说，在 T 之上，用二元关系 F 对 q_j进行标引后，q_j就成为该 n 维 T 空间中的一个提问向量，而 k 个提问向量就构成了提问空间。对于检索来说，通过 R 关系确定：在任意一个 q_j向量的周围，或者在人为给定的范围内，蕴含多少个 d_i向量，即可被 q_j命中的文献。

1.1.3 信息检索的类型

信息检索的类型很多，可以从不同的角度进行分类，下面仅从信息检索的对象性质

和计算机检索技术两个方面阐述信息检索的类型。

1. 按照信息检索的对象性质划分

（1）文献检索（Document Retrieval）。文献检索的对象是文献，例如，检索有关“太阳能电池”方面的文献。这里所说的“文献”是指文献单元，即包含一个完整内容的单元，如一篇论文、一本图书、一份报告等，而忽略其物理载体（如纸介质、磁介质、光介质）、出版形式（如图书、期刊、报纸）、加工深度（如一次文献、二次文献、三次文献）等。进一步说，这里的“文献”可以是完整的原始文献，也可以是原始文献的替代品，如一条目录款目、一条文摘款目或一条索引款目。归根结底，文献检索的目标是检索出原始文献或原始文献的替代品。供文献检索使用的数据库是文献数据库，包括目录、文摘、索引、全文等数据库。

按照文献内容的完整性，文献检索又可以进一步分为书目检索（Bibliographic Retrieval）和全文检索（Full Text Retrieval）。

1）书目检索。所谓书目检索，是指检索对象为原始文献的替代品，即文献线索，而不是原始文献本身，要想阅读原始文献，还必须依据文献线索去进一步找到和获取原始文献。书目检索通常借助于文摘数据库、索引数据库、目录数据库来完成。书目检索的首要目标是检索出包含用户所需信息的书目记录，其数据库则由被存储文献的书目记录构成。

2）全文检索。所谓全文检索，是指检索对象为原始文献本身，主要是对全文中的字、词、句、段等进行检索，检索出来的结果就是原始文献，进而可以直接阅读和使用原始文献。全文检索通常借助于全文数据库来完成，通常可以提供报纸、手册、字典、百科全书、统计资料等的文摘或全文，其首要目标是找出能满足用户所需信息的某个实际文本。全文数据库包含文献的实际文本，最终的检索结果也是实际文本。应当指出，全文检索的完整含义不限于检索结果是全文，而是使用全文中的各种元素（如字、词、句、段等）进行检索。因此，如果只使用题名、作者、关键词、摘要等进行检索，而不能使用全文中的各种元素进行检索，即使检索结果同样是全文，也不是严格意义上的全文检索。

无论是书目检索还是全文检索，都假定存在一个有信息需求的目标用户群。当用户提出询问时，系统应能提供包含他们所需信息的书目记录或全文文本。

文献检索是最典型的信息检索，也是信息检索的早期类型。对于学术研究来说，文献检索仍然是目前使用最普遍的信息检索类型。在许多情况下，可以把文献检索直接理解为信息检索的同义语。

（2）数值检索（Numeric Retrieval）。数值检索有时也叫数据检索（Data Retrieval）。数值检索的对象是以数字形式表示的具体数值，如生产指标、统计数据、物价、股票及理化特性等，主要应用于科学研究、工程设计和经济统计等领域。数值的范围不限于数字本身，还包括图形、图表、数学公式、化学分子式及结构式等非数字型的数值。数值检索的目标是检索出能满足给定条件的、能够直接使用的数值，如钢铁产量、国内生产总值（GDP）、居民消费价格指数（CPI）、汽车的价格、黄金的密度、聚氯乙烯的分子结构、尼罗河的全长、喜马拉雅山的高度等。供数值检索使用的数据库是数值数据库。例如，物理数据库可以提供有关物质的密度、比热容、沸点、熔点、拉力和压力等参