信息组织与检索

Information Organization and Retrieval

李国辉 汤大权 武德峰 编著

科学出版社
www.sciencep.com

信息组织与检索

李国辉　汤大权　武德峰　编著

国防科技大学学术专著专项经费资助出版

科学出版社

北　京

内 容 简 介

本书系统、全面地介绍和阐述了现代信息组织和检索的原理、方法以及最新发展。它把多媒体信息检索和常规的文本信息检索技术融合在一起，从计算机科学和信息技术的角度来看待信息组织和检索中的问题。本书由三大部分组成：信息及其组织、信息检索、应用。“信息及其组织”部分介绍数据和信息的概念、信息检索的数据模型、多媒体信息的内容描述、数据预处理和媒体结构化问题。“信息检索”部分介绍信息检索的方法和技术，包括用户查询接口、检索和索引算法、基于内容的多媒体信息检索方法和算法。“应用”部分介绍两种典型的信息检索应用：Web 检索引擎和数字图书馆。

本书可以作为计算机科学、管理科学与工程、图书馆科学、电子商务、信息管理与信息系统等专业的教材，也可以供从事 Web、Intranet、信息系统、数字图书馆、文档管理系统、专业媒体库系统和技术的研究、设计和开发的工程技术和管理人员参考。

图书在版编目(CIP)数据

信息组织与检索/李国辉，汤大权，武德峰编著. —北京：科学出版社，2003

ISBN 7-03-011037-4

Ⅰ. 信… Ⅱ. ①李… ②汤… ③武… Ⅲ. ①信息管理 ②情报检索 Ⅳ. ①G203 ②G252.7

中国版本图书馆 CIP 数据核字(2002)第 099035 号

责任编辑：鞠丽娜 韩洁 / 责任校对：钟 洋
责任印制：吕春珉 / 封面设计：陈 琳

科学出版社 出版
北京东黄城根北街16号
邮政编码：100717
http://www.sciencep.com

新蕾印刷厂 印刷

科学出版社发行 各地新华书店经销

*

2003 年 1 月第 一 版 开本：B5(720×1000)
2003 年 1 月第一次印刷 印张：20 3/4
印数：1—4 000 字数：403 000

定价：32.00 元

(如有印装质量问题，我社负责调换(路通))

前　言

在过去 20 年里，信息管理、计算机和图书馆科学方面的专家在信息检索领域开展了大量的包括建模、分类、系统体系结构、用户接口、数据可视化、过滤、查询语言等方面的研究。进入 20 世纪 90 年代以后，信息检索领域不断发展，超出了原来的文本索引和文本文档搜索的主要目标。这是因为互联网和数字化时代的到来，信息的形式是多种多样的，不仅仅是经典的文本，还有图像、视频、音频等其他形式的信息载体。Web 逐步成为人类知识和文化的环球库，允许前所未有的思想和信息的共享。传统的信息检索概念和方法在更新和发展，同时又产生了现代的信息检索技术。本书正是一本系统、全面介绍现代信息组织和检索的原理、方法以及最新发展的书籍。我们根据十多年从事信息系统方向的教学和研究工作的积累和经验，结合“八五”、“九五”国家预研和省部委基金项目在信息系统、信息管理和多媒体信息检索方面取得的科研成果的基础上撰写了这本书。

虽然信息检索方面出版了一些书籍，但是系统介绍信息检索领域技术方面的专业书籍和教材还不多，尤其是把多媒体信息检索和常规的文本信息检索技术融合在一起的教材和专业书籍很少。从传统的观念看，文本检索和多媒体检索是不相同的，但是在数字汇聚的趋势下，文本、图像、视频、音频等将会越来越多地组合在一起表达内容和信息。本书就是一本综合介绍文本和多媒体信息检索的书籍。这是本书的主要特点。

本书从计算机科学和信息技术的角度，为信息检索提供一个整体的和系统的视图，即这本书关注于信息组织和检索方面的计算机方法及其技术，系统地阐述信息的组织和检索原理、概念、模型、方法和应用。本书讲述的关键问题是：

- 如何描述信息资源或信息的承载对象，以便有效地利用这些资源？这就是信息的组织问题。
- 如何查找出用户需要的信息？这就是信息检索问题。

1. 本书的章节划分

本书由三大部分组成：信息及其组织，信息检索，应用。信息及其组织介绍数据和信息的概念、信息检索的数据模型、多媒体信息的内容描述、数据预处理和媒体结构化问题。信息检索介绍信息检索的方法和技术，包括用户查询接口、检索和索引算法、并行与分布信息检索技术、基于内容的多媒体信息检索方法和算法。应用部分介绍两种典型的信息检索应用：Web 检索引擎和数字图书馆，如表 A 所示。

表 A　本书的主题和章节

主题	章
信息及其组织	数据及其文档形式
	特征内容处理
	信息检索模型
	多媒体数据内容描述模型
信息检索	查询与用户接口
	索引和搜索
	并行与分布信息检索
	基于内容的多媒体信息检索
应用	Web 信息检索
	数字图书馆

本书由李国辉负责统稿，并撰写了第 1～3、5、6、8、9、11 章，汤大权撰写了第 4、7 章，武德峰撰写了第 10 章。

2. 本书的阅读

图 A 解释本书各章节的总体结构及其各章的关系，读者可以在阅读时注意它们之间的顺序和关系。读者可以顺序阅读，也可以挑选其中的章节进行阅读。在第 1 章，给出了信息组织和检索的基本概念和本书的概貌，是本书的起点。从第 2 章到第 5 章是信息及其组织部分。我们首先学习数据形式及其特性(第 2 章)，然后介绍如何把原始文档和媒体按照检索要求进行处理，提取逻辑特征和属性(第 3 章)，之后我们学习如何有效地组织和描述这些数据(第 4 章和第 5 章)。

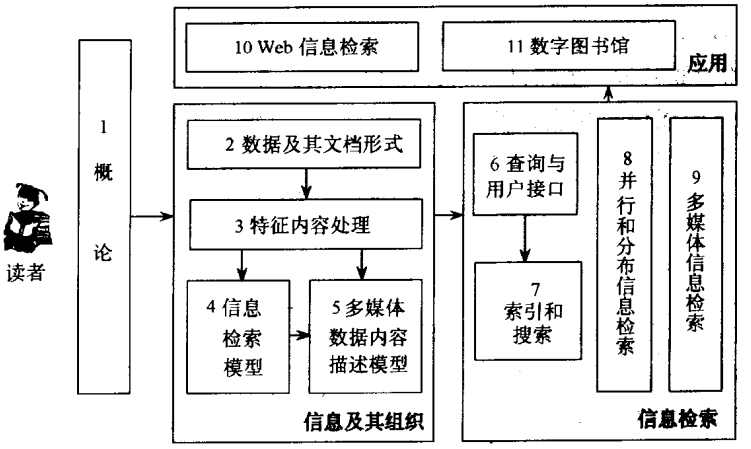

图 A　本书的章节概貌

从第 6 章到第 9 章是信息检索主题，这里主要涉及的两个问题是查询和搜索，它们的内容分别放在第 6 章（查询与用户接口）和第 7 章（索引和搜索）。并行和分布信息检索和多媒体信息检索的内容作为两个综合问题，分别放在第 8 章和第 9 章。

最后两章是信息检索技术的两个典型的应用，即 Web 信息检索（第 10 章）和现代数字图书馆（第 11 章）。这两个应用都是全新的信息检索应用，其概念和技术涉及到全书所有章节的知识，但是在具体的应用环境中，它们又具有特定的含义和方法。

信息检索的性能需要通过定性的方法来评价。每一种检索方法的检索性能的优劣，需要一种合适的评价度量方法。在本书的后续章节中（主要在“信息检索”主题的各章中），介绍的检索方法都需要一致的、客观的性能评价，因此把检索性能的评价作为检索系统和过程中总体的一部分，放在本书的第 1 章介绍。

对于多媒体数据，关键问题是如何对包含多媒体对象的文档和多媒体数据进行建模、索引和搜索。多媒体检索的模型和技术与经典的文本检索不同。但是，现在明显的现实和趋势是：图像和其他多媒体数据与文本经常是结合在一起的，各种文档和信息资源越来越多地包含各种数据类型，Web 就是一个明显的例子。因此，未来所有的数据，无论是文本，还是图像、视频、音频，这些数据将按照一致的和统一的方式处理。本书正是尝试朝这个方向迈进，把多媒体信息检索的技术结合到各章节之中，而不是把它们割裂开来编写。这些内容分别放入第 2 章（多媒体数据形式及特征）、第 4 章（多媒体内容描述和模型）、第 5 章（多媒体结构化和特征提取）、第 6 章（多媒体查询和可视化接口）、第 9 章（基于内容的多媒体信息检索）、第 10 章（Web 中的多媒体检索）、第 11 章（数字图书馆中的多媒体信息组织和检索）。

3. 本书的读者对象

虽然本书是多个作者共同编写的，但是本书的结构和内容是经过精心设计和编写的。它既可以作为大学的教科书，又可以作为该领域或相关领域的研究人员和信息管理人员的参考书。本书所有的章节都集成在一个统一的框架下面，以统一的风格和结构形式呈现给读者。本书在每一章的开始，概述并引出全章的内容；每一章都有一个小结部分，它总结全章的要点和各部分的关系，讨论研究的问题以及发展趋势。这些讨论对于研究生以及研究人员是非常有价值的。

本书可以作为教材，用于各相关学科和领域的本科或研究生教学，例如计算机科学、管理科学与工程、图书馆科学、电子商务、信息管理与信息系统等专业的教学。相应的课程可以是“信息检索”、“现代信息检索技术”、“多媒体信息检索”、“信息检索专题”、“Web 检索和信息存取”、“数字图书馆”等。

信息组织和检索是信息技术的核心问题。希望本书的出版能够为我国信息化

建设添砖加瓦，做出一点点贡献。

4. 致谢

在本书的写作期间，许多人曾给予我们帮助和支持。首先感谢国防科技大学2001年度学术专著出版基金的资助，感谢国防科技大学科研部、国防科技大学人文与管理学院训练部、管理科学与工程系的专家、学者、科研和教学管理人员、同事和学生的支持和帮助。特别感谢胡晓峰教授、沙基昌教授、谭跃进教授、张维明教授、常兆城副院长，科研和教学管理人员张海新和张勇等，同事张茂军、王晖、王炜、涂丹、甘亚莉、张军，学生曹莉华、柳伟、王辰、熊华、薛峰、李恒峰、吴玲琦、汤义、倪泞、林洪文、胡军涛、周祥东、姚作梁、李梦君、段晓娟等。我们在本书中引用了许多经典性的论述和前沿研究资料，特向这些资料的作者表示感谢。感谢本书的编辑鞠丽娜副编审的辛勤工作。此外，我们还要对我们的家人致以特别的感谢。

由于作者的水平所限，书中难免存在一些缺点和错误，敬请广大读者批评指正。

作　者

于国防科技大学

2002年9月

目　录

第1章 概 论

【本章提要】 每天，数据采集、计算机数字化、卫星遥感、生产和经济运行、办公和管理等系统产生大量的数据。尤其是随着信息技术在全球迅猛的发展，人们每天都在与世界上最庞大的资源库 Internet 打交道。我们被“淹没”在数据的海洋中，为此需要有效地从大量的数据集中检索出信息的方法和工具，这就是信息检索的任务。所谓信息检索，就是根据用户的信息需求，从文档集中检索出与用户信息需求相关的文档子集。信息检索系统由信息组织和信息检索两大部分组成，信息检索性能的提高需要良好的信息组织。本章根据这个划分，首先阐述信息组织和检索的概念，然后给出信息检索系统的组成和一般的信息检索过程，并介绍常用的检索性能的评价方法，最后给出本书的总体组织轮廓。

1.1 信息组织和检索的概念

在这一节，我们首先从一个 Web 信息站点设计的例子中初步了解信息组织和检索的含义，然后在详细讨论什么是信息之后，给出信息组织和检索的概念。

1.1.1 Web 信息组织和检索的例子

我们首先通过一个以企业或单位的 Web 网站的设计和使用的例子来说明信息组织和检索中的问题。

Web 网站的设计，首先涉及到各类资料的组织问题，它们包括内容设计、导航设计和表现设计。内容设计就是内容的结构和分类组织。设计人员要确定分类的准则，然后根据分类准则，划分信息内容。例如一个大学的网站，可能按照学校概况、管理机构、院系设置、招生信息、信息资源、图书馆、科学研究、教师队伍、学生活动等主题分类。导航设计涉及到信息单元的浏览、用户与 Web 内容及结构的交互。Web 的两个基本元素是页和链，链把页面关联起来，构成巨大的“蜘蛛网”Web。Web 设计者要仔细进行导航的设计，把相关内容通过链联系起来。表现设计就是对信息内容和导航的视觉表现进行设计，例如颜色、表现结构布局、表现顺序、表现方式等，目的是提供一个易于获取信息的 Web 环境。

Web 网站设计与数据库设计是不同的。Web 网站包含文本、表格、图像、视频、音频等类型的数据，而不仅仅是常规数据库管理的数值字符数据。更重要的是，它们的初始设计目的不同。Web 设计目的是用于资料和信息的交流（当然，随

着 Internet 的发展，Web 的应用已经远远超出了原来的科研交流用途，向着更广泛的应用发展，例如电子商务、娱乐、教育等），而常规数据库系统的设计一般用于企业的数据管理。另外，它们采用的模型不同，Web 用的是超文本（超媒体）模型[1,2]，而数据库目前所采用的主要是关系模型或对象模型[3]。因此对于 Web 设计来说，是对内容进行组织和分类，进行交互导航设计，而内容的表示和描述用于信息检索。而在数据库设计中，用二维表格表示实体及其属性，通过严格的代数关系和约束提供数据库的查询。

设计的 Web 网站仅仅提供分类目录和链的浏览是不够的，尤其是对大中型的 Web 网站来说，用户在浏览过程中容易迷航，或在信息查找中花费太多的时间，而又难以得到相关的信息内容，因此对 Web 的搜索就是必不可少的信息检索手段。

我们仍然把例子限制在一个单位的 Intranet 内，Intranet 搜索引擎提供对一个企业或单位内部信息内容的搜索。例如，大学的主页提供的搜索引擎，可以对全校各学院、系、教师、学生、实验室、研究小组的网页内容进行搜索。设计的检索系统能够为任何水平的用户，包括专业用户和一般用户提供有效的信息检索服务。

最普遍的查找信息的一种方法是使用目录，例如书的目录、电话目录、资料的目录等形式。大部分的 Web 搜索网站的主页面都向用户呈现目录界面，例如，YAHOO!。但是，当内容较多时，目录的层次会很大，这时通过目录查找内容就非常费时费力，而且目录不能容纳太多的内容。于是，另一种方法就是通过关键词对 Web 进行搜索。

为了对 Intranet 的内容进行搜索，查找出有用的信息，搜索系统首先要做的事就是定期地到 Intranet 内各网页上收集数据，通过特殊的算法，去掉冗余的链和结点，然后为这些网页构造出元数据，并建立索引。在这里，元数据是关于 Web 页的信息，例如标题、页的长度、内容链、外部链、从根到本页面的最短路径等。索引是文件、文档或资料的一种目录表，其中包括定位和查找这些内容的关键词和引用标志。也可以把索引看成是一种数据库对象，能够实现对数据行的快速、直接地存取而不必扫描整个数据库表。对于数据库表的被索引字段的每个值，索引中都有一个入口，该入口包含指向具有该值的数据行的指针。

对于网页文档中的文本，通过提取词和词干，建立倒排索引。根据用户输入的词，搜索引擎搜索倒排索引，查找出相似的一系列结果，以 HTML 形式向用户呈现。

对于网页文档中的图像、视频和音频，可以结合基于内容的信息检索技术（在第 9 章中将详细介绍），由此用户不仅可以通过关键字进行检索，还可以根据多媒体数据的视听、语义和时空逻辑特征等内容进行多媒体信息的检索[4~7]。

以上典型的 Intranet 环境下的 Web 设计例子说明了信息组织和检索的一个

基本轮廓：从内容的组织到向用户提供内容的使用（信息检索和浏览等）过程中的基本处理方式。也就是说，为了向用户提供信息存取服务，需要两大部分的工作：一是信息组织，另一个就是信息检索。

1.1.2 数据、信息和知识

在信息检索这个概念中，我们经常会遇到“数据”和“信息”这两个名词。“数据”和“信息”经常会交叉使用，容易混淆，它们的含义不同，但是又有某种程度的联系。

数据（data）可以是文本、图像、视频、音频等基本数据元素。数据是按某一规格化方式对事实和概念的一种表示，适于人或自动装置进行通信、解释或处理。它是任何有意义或可以赋予含义的表达形式，例如字符或数字。在计算领域，数据是输入给计算机程序或例行程序的内容，它们可以经过算术或逻辑运算的处理，求得所处理后的结果。根据这种定义，程序和编程指令不属于“数据”的范围，“数据”指的是程序和指令进行处理的那些数据项（内容）。然而从广义来讲，“数据”也包括程序和编程指令。在这里，与“信息”一词相比较而言，“数据”指的是源数据或原始数据，而“信息”则定义为通过对数据进行处理之后获得的数据。

“信息”一词对应于英文的“information”，它在不同的领域有不同的解释。其含义可以涉及到哲学、心理学、信号处理、物理学方面的解释。有人把它定义为“关于某事的消息或事实”。牛津英语字典中，“信息”一词的解释是：通知、告知、告知的事情、消息、一种知识等。从这里的解释看，“信息”又与“知识”联系在一起。“知识”（knowledge）在牛津英语字典里解释为“通过经验获得的认识，是个人的信息范畴，是理论或实践上的理解，是对已知事实的总结。”广义地说，知识有经验知识和理论知识之分，知识是人类认识的成果和结晶，是有助于解决问题的有格式可重用的信息。

图1.1中给出信息的层次观点。最下层是数据，是信息的来源和原始资料，用数据可以表达信息。信息是经过处理、组织和表现出来的数据。读、听、看、理

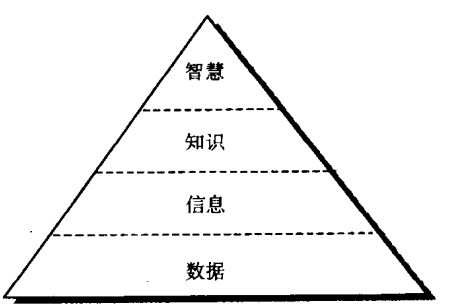

图1.1 信息的层次

解的信息，经过归纳和总结得出知识。最高层为智慧，是提炼和综合出来的知识和理解，它建立在知识之上。图 1.1 中用金字塔形状，下宽上窄，表示层次越高，抽象级别就越高；另一方面，随着层次的上升，要求的表达数据量就越少。因此，我们要注意在学习知识时不要忽略智慧的产生；在获取信息后，进一步提炼出知识；在数据的海洋中知道如何获取有价值的信息。

根据本书的主题，我们把重点放在“信息”的概念讨论上。为了更进一步理解什么是信息，我们可以从形式、内容和相关概念等方面去考虑。

从信息的特性看，信息可以通过广播和网络进行电子化交流，因此信息容易复制和实现共享。直观上看，信息与事物和事实有关，也许是一种物质、能量或抽象的概念；信息是新闻，因此重复以前接收到的消息不是信息；不正确的或反面的事实是错误的信息。这里强调消息的含义和使用。

从人的角度看，信息的定义更具主观性。信息的接收者是人，人的认知处理有多种级别，从感知到观察/关注，到最高级别的推理、形式推论和理解。人通过知识来判断信息的真实性。通过参考某些正面的观点，结合观察的事实和推理过程，产生归纳的结果。这些过程其实就是人们接受信息的过程。另外，并不是世界上产生的所有信息都是每个人关注的，不同的人关注不同的信息。例如，昆虫学家关心蚂蚁间的通信和信息交流，自然保护组织关心城市建设中树木被砍伐的情况等。其实，一个人不可能接受所有的信息，他/她只关心与自己有关的信息。

从信息的含义与形式看，含义和形式是信息的两个不同的层面。同一个信息含义可以用不同的形式表达。例如汽车，可以用文字表达，并表述它的型号和性能数据等；也可以用图片形象地呈现汽车的外形；或用一段视频播放汽车的行驶，信息表达更生动，当然还可以配有声音，全方位展示汽车性能。信息的含义需要媒体的表示，向信息接受者呈现（表现）信息的内容。

早在 20 世纪 40 年代，香农（Claude Shannon）就开始研究信息论[8]，提出了通信系统中的信息度量方法。在香农的信息论中，信息的度量是用一种对数度量单位，它用一组互斥事件的对数（以 2 为底）表示信源的信息量（熵）。

信息的种类繁多，数据量大。信息可以通过各种媒体承载，例如文本（包括书、期刊、Web、出版物、广告等）、视频、图像、广播、电视、电话、数值数据、表格等。

那么，我们日常打交道的包含信息的数据有多少呢？我们用兆字节（MB，10^6）来作单位是远远不够的。我们要用到吉字节（GB，10^9，又称千兆字节）、太字节（TB，10^{12}，又称兆兆字节，）、派字节（PB，10^{15}，又称千兆兆字节）、艾字节（EB，10^{18}）。

例如[9]，美国的国会图书馆有 20TB 的书籍数据量、13TB 的图片、200TB 的地图、500TB 的文件、2000TB 的录音资料，共计 3PB。1997 年的 Web 上大约有

2TB。在1989年，全球共生产4600多部影片，假设按MPEG-1的压缩标准记录，每部影片按1200MB算，一年的影片数据量达5.5PB。每年产生的图片大约有520亿幅，按每幅10KB，共520PB。NASA地球观测图像大约是11000TB。对于广播信息，那就更多了，假设十分之一是原始的电视资料，那么美国1600个电视台的电视视频数据大约每年有20PB，全球大约有200PB，而每年美国6900个广播电台的广播音频数据大约有1.7TB。每年还销售大量的CD和音乐磁带，意味着大约每年有15TB（美国）、60TB（全球）的音乐数据要产生，还有一个最普遍的消息来源是电话会话，在美国大约有4000PB的语音数据。

在1999年，Web上大约800兆的网页，Internet通信量大约每100天翻一倍。广播电台花费38年达到5千万听众，电视花13年，而网络只花了4年就达到了这个规模。据统计，数据业务的增长速度大大超过了话音业务[10, 11]。

与此相比，人的大脑记忆是有限的。把信息的接受、遗忘和日常工作所需的信息量考虑进去，人的大脑一般容纳约200MB。

这就导致了所谓的“信息过载”。今天最大的问题就是如何使人们忽略和拒绝不相关的和过多的信息。其实，并不是信息的过载，而是大量的、杂乱的、无关的信息和数据把人的视听给淹没了，人们得不到真正想要的信息，因此需要信息的组织和检索来满足用户对信息要求。

1.1.3 信息的周期

从产生到被利用，信息具有一个完整的生命周期，如图1.2所示。信息的生命周期有三个主要阶段：产生、检索、利用。

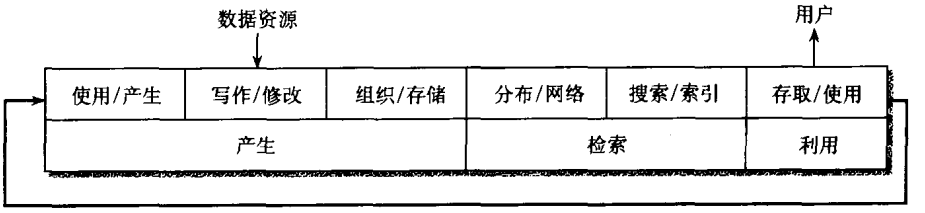

图1.2 信息的生命周期

1. 信息产生阶段

(1) 使用/产生

产生的信息供用户使用，使用过程中又可以产生新的信息。例如可以把信息转换为知识，知识也可以作为新的数据和新的信息。由此，信息的生命期开始了新的一轮周期。

(2) 写作/修改

信息的写作和修改是信息产生阶段的重要步骤。通过把数据、信息和知识进行一定的处理，产生出新的信息，或从观察和思考过程中产生信息。信息可以经过编辑并发布出去。

(3) 组织/存储

信息组织就是收集和综合信息，建立元数据库、原始资料库及其关系。元数据是描述数据的信息，支持信息的搜索。经过组织的数据以利于检索的各种形式存储在数字信息库中。

2. 信息检索阶段

(1) 分布/网络

一般来说，信息环境是一个分布的信息空间，用户通过网络在一个分布的信息环境中获取和利用信息。

(2) 搜索/索引

在分布的信息空间中进行信息的搜索，即如何从信息库中找到所需的相关信息。为了快速对信息库进行搜索，必须建立起索引。

3. 信息利用阶段

存取到信息之后，就可以使用获取的信息。存取可以是浏览、搜索或过滤等形式。过滤是通过在信息的输入和输出阶段建立的“配置”(即用户特定的参数配置）来为用户提供信息。用户可以选择要求的或相关的信息，实现个性化的信息服务。用户在信息使用过程中，可以产生新的信息，信息进入下一轮生命期。

1.1.4 信息组织和检索

在本书的信息组织和检索的概念叙述中，将大量用到文档这个词。我们用文档（document，又称为文献）来表示一个数据单元，文本是它的一种典型的形式，但是文档也可以包含其他的媒体，例如图像、视频和音频。文档可以是一个完整的逻辑单元，例如一篇研究论文、一本书或一本手册。它也可以是其中的一部分，例如一个自然段或多个自然段（或称为一节）、字典中的一个条目、一个汽车零件的描述等。

对于它的物理表示来说，一个文档可以是任何物理单元，例如一个文件、一个电子邮件或 Web 页。图像、视频和音频也可以作为单独的文件存在，多个图像、短视频和音频可以形成一个文件，长视频和音频文件一般单独为一个文件。这时的文档就是多媒体数据文件。大量的文档组成文档集 (collection)。这里的文档集是数字信息库的原始资料（数据）库。注意到，虽然文档这个词有个“文”字在内，但是它并不仅仅表示文本数据。随着信息技术和多媒体技术的发展，文档的

内容形式多样，可以包含图像等多媒体数据在内。因此在本书中，把文档看成是一个内容的载体或容器，它其中不仅仅包含文本，也可以包含多媒体数据。如果不特别指定，我们把文档广义地看成是包括普通文本文档、扩展的多媒体文档、多媒体数据在内的所有形式的数据单元。在信息组织和检索中，把文档看成是一个检索单元。

组织（organization）就是把数据按照一定的结构、顺序、排列方式组织起来；检索（retrieval）就是重新获得或恢复，是进行搜索、定位及读出数据的过程。从本节的例子和信息的含义来看，我们可以得出，信息组织就是按照信息检索的需要，对数据及其特性进行组织，而信息检索就是从大量的文档集中获取用户需要的相关信息。

与信息检索（IR，Information Retrieval）相近的一个概念是数据检索[12]。对于数据库系统来说，数据库的查询用到的是数据检索的概念。数据检索就是根据数据库的结构化属性来搜索，确定哪些文档的属性中包含用户查询的关键字。然而，这对于用户的信息需求来说，是非常不够的。因为赋予特定关键字的文档中可能包含更多的没有在关键词中反映出来的信息。事实上，用户更关心的是检索出有关某个主题的信息，而这些主题信息包含在文档的内容当中。在数据库系统中，其数据检索语言的目标就是检索出满足定义条件的所有对象，定义的条件可以用规范表达式和关系代数表达式来说明，而数据检索基于的表格式属性不能有效地表达文档的信息内容，因此不能有效地支持信息检索。

对于数据检索系统来说，如果在一千个检索出的对象中有一个差错的对象，就意味着检索的失败，因为它是一种精确匹配，例如，查询“run”，将只匹配 run，而不匹配 runs 或 running。但是对于信息检索来说，检索到的对象可以不太精确（部分匹配），允许有一些小的不明显的偏差。因此，数据检索和信息检索的主要区别是：信息检索涉及到用户的信息需求和提交的查询不总是结构化的，而且具有语义模糊性；而数据检索系统，例如关系数据库系统，涉及的数据具有完好定义的结构和语义。

数据检索为数据库系统的用户提供了一种查询的方案，但是它没有提供和解决信息检索中的问题。为了满足用户的信息需求，信息检索系统必须以某种方式“解释”文档库中数据单元的内容，并把检索的结果按照与用户查询的相关程度来排序。文档内容的“解释”涉及到从文档中提取语法和语义特性，并用这些特性去匹配用户的信息需求。

困难的是，不仅要知道如何提取这些特性，而且要知道如何利用它们来决定与查询的相关性。因此，相关性（relevance）是信息检索的核心。事实上，信息检索系统的主要目标是检索出所有与用户查询相关的文档，尽可能减少不相关的文档，因此，信息检索的一种规范定义为：从大量收集的数据或文档集 C 中，找到