

第 1 章 绪论

1.1 什么是数据结构

1.2 基本概念和术语

1.3 抽象数据类型

1.4 算法和算法分析

1.1 什么是数据结构

- 简单说，数据结构是以某种方式联系在一起的数据元素的集合。程序中的**数据结构**反映了程序员在程序中**表示信息的方法**。**算法**反映了如何**处理信息的方法**。
- 数据结构研究的是数据元素之间抽象化的**相互关系及这种关系**在计算机中的**存储表示**。对每种结构定义各自的运算，设计出相应的算法，并用某种语言实现该算法。
- 数据结构的地位：数学、硬件、软件之间。核心专业基础课。
- 是编译程序、操作系统、数据库系统以及其他系统程序、大型应用程序的重要基础。

1.1 什么是数据结构

对于一个课题，在计算机领域一般遵循下面的解决程序：

需求分析——总体设计——模块分割——建立数学模型
——解数学模型的算法——程序编制——调试——结果

1.1 什么是数据结构

思考：

1. 计算机不能做什么？
2. 用枪打树上鸟的问题。

1.1 什么是数据结构

思考：1. 计算机不能做什么？



模型、算法与学习之间的关系：

模型是样本、算法与学习三者的综合体现，有时可用准确率、召回率等一系列性能指标来衡量其优劣。

在样本给定的情况下，算法之于具体任务的好坏则直接决定了模型性能的优劣，但本身并不直接用这些指标来衡量。学习则是由样本通过算法而生成模型的过程。

准确率、召回率等评价指标是针对模型，而不是算法。

样本（数据）是基础，而对样本（数据）的描述需要用[数据结构](#)。



1.2 基本概念和术语

1. 基本术语

(1) **数据**：描述客观事物的数字、字符以及所有能输入到计算机中并被计算机程序处理的符号的集合。

(数字、字符、声音、图形、图像等等)

(2) **数据元素**：数据的基本单位，在计算机程序中常常作为一个整体进行考虑和处理，如记录 / 结构。

(3) **数据项**：数据的不可分割的最小单位，如结构中的域。

(4) **数据对象**：性质相同的数据元素的集合，是数据的一个子集。

1.2 基本概念和术语

1. 基本术语

(5) 逻辑结构：数据与数据之间的联系

(6) 存储结构：数据结构在计算机中的存储表示（或称映像），又称物理结构。它包括数据元素的表示和关系的表示。

1.2 基本概念和术语

2. 数据结构

(1) 定义：是相互之间存在一种或多种特定关系的数据元素的集合。

数据之间不是相互独立的，他们之间有某种特定的关系，这种数据元素之间的关系，称为“结构”。

结构 = 实体 + 关系

- 另一种定义：按照逻辑关系组织起来的一批数据，按一定的存储方法把它存储在计算机中，并在这些数据上定义了一个运算的集合。
- 形式定义：二元组 (D, S) 其中 D 是数据元素的有限集， S 是 D 上关系的有限集。

1.2 基本概念和术语

2. 数据结构

(2) 四种基本结构 (逻辑结构) p5

集合：元素仅属于同一个集体，没有其他关系。

线性结构：存在一对一关系，序列相邻，次序关系。

树形结构：存在一对多关系，层次关系。

图状结构：(网状结构) 存在多对多关系，任意性。

- 存储器模型：一个存储器 M 是一系列固定大小的存储单元，每个单元 U 有一个唯一的地址 $A(U)$ ，该地址被连续地编码。每个单元 U 有一个唯一的后继单元 $U' = \text{succ}(U)$
- 物理结构就是逻辑结构到存储器的一个映射。
- 四种存储结构：顺序存储、链接存储、索引存储、散列存储

(3) 实例：

表：计算机系人事表

工号	姓名	性别	职务	教研室	工作时间	发表论文
01			系主任	软件	2000.1	A,B
02			教研室主任	软件	2000.1	B,C,E,F
03			教师	软件	2015.8	C,D
04			教师	应用	2012.8	A,G
05			教师	应用	2000.9	E,I
06			教师	应用	2007.2	F,J
07			教师	软件	2003.8	D,L
08			教研室主任	应用	2001.7	G,H
09			教师	应用	2006.8	H,I,J,K
10			教师	软件	2016.2	L,K

- 线性结构示例 $R = \{ \langle 05, 01 \rangle, \langle 01, 07 \rangle, \langle 07, 02 \rangle, \langle 02, 08 \rangle, \langle 08, 04 \rangle, \langle 04, 10 \rangle, \langle 10, 03 \rangle, \langle 03, 06 \rangle, \langle 06, 09 \rangle \}$
- 树型结构示例
 $R = \{ \langle 01, 02 \rangle, \langle 01, 08 \rangle, \langle 02, 03 \rangle, \langle 02, 10 \rangle, \langle 02, 07 \rangle, \langle 08, 04 \rangle, \langle 08, 05 \rangle, \langle 08, 06 \rangle, \langle 08, 09 \rangle \}$
- 图状结构示例 $R = \{ (01, 02), (01, 04), (02, 05), (02, 06), (02, 03), (03, 07), (04, 08), (05, 09), (06, 09), (07, 10), (08, 09), (09, 10) \}$

1.2 基本概念和术语

3. 数据结构的划分

(1) 按数据结构的性质划分

- 数据的逻辑结构——数据元素之间的逻辑关系（设计算法 -- 数学模型）
- 数据的物理结构——数据结构在计算机中的映像。（存储结构，算法的实现）

(2) 按数据结构在计算机内的存储方式来划分

- 顺序存储结构——借助元素在存储器的相对位置来表示数据元素之间的逻辑关系。
- 链式存储结构——借助指示元素存储地址的指针表示数据元素之间的逻辑关系。
- 索引存储方法——在存储结点的同时，还建立附加的索引表，索引表中的每一项称为索引项，形式为：**关键字，地址**。
- 散列存储方法——根据结点的关键字直接**计算**出该结点的存储地址。

四种存储方法可结合起来对数据结构进行存储映像。

1.2 基本概念和术语

3. 数据结构的划分

(3) 按数据结构的操作来划分

- 静态结构——经过操作后，数据的结构特征保持不变（如数组）。
- 半静态结构——经过操作后，数据的结构特性只允许很小变迁（如栈、队列）。
- 动态结构——经过操作后，数据的结构特性变化比较灵活，可随机地重新组织结构（如指针）。

1.3 抽象数据类型—— ADT

- 定义：是指基于一个逻辑类型的**数据模型**以及定义在该模型上的一组**操作**。每一个操作由它的输入和输出定义。
- 一个含抽象数据类型的软件模块，通常应包含：**定义、表示、实现** 3 个部分。
- 形式：三元组 (D, S, P) D 是**数据对象**， S 是 D 上的**关系集**， P 是对 D 的**基本操作集**。 P8
- 说明：一个 ADT 的定义并不涉及它的实现细节。这些实现细节对于 ADT 的用户是隐藏的。隐藏实现细节的过程称为封装。数据结构是 ADT 的物理实现。ADT 的每一个操作均由一个或多个子程序来实现。

1.3 抽象数据类型—— ADT

- 抽象数据类型和**数据类型**实质上是一个概念。数据类型是一个值的集合和定义在这个值集上的一组操作的总称。分为原子类型和结构类型
- 数据抽象与过程抽象，实现信息隐蔽和局部化以一个严格定义的过程接口的方式，在数据结构上提供一个抽象。数据的实现和处理细节被隐蔽了

1.4 算法和算法分析

- 1. 算法的定义 p13

指一系列**确定的**而且是在**有限步骤内**能完成的操作。

1.4 算法和算法分析

2. 算法的特性

- (1) 有穷性（能执行结束）
- (2) 确定性（对于相同的输入执行相同的路径）
- (3) 有效性（可行性）
 - 用于描述算法的操作都是足够基本的
- (4) 0 至多个输入
- (5) 1 至多个输出

问题：程序是不是算法？

如操作系统程序，只要系统不遭破坏，它就永远不会停止，即使没有作业要处理，仍处于一个等待循环中，等待新作业的进入。因此，操作系统程序不是一个算法。

1.4 算法和算法分析

3. 算法与数据结构的关系

- 计算机科学家沃斯（ N.Wirth ）提出的“**算法 + 数据结构 = 程序**”揭示了程序设计的本质：对实际问题选择一种好的数据结构，加上设计一个好的算法，而好的算法很大程度上取决于描述实际问题的数据结构。算法与数据结构是互相依赖、互相联系的。
。
- 一个算法总是建立在一定数据结构上的；反之，算法不确定，就无法决定如何构造数据。
—— 请注意 p6 图 1.6 上面 2 行的内容

例 1 ：编写程序查询某城市某人的电话号码

建立一张登记表，存放 2 个数据项：姓名 + Tel。
好的算法取决于这张表的结构及存储方式。

（ 1 ）将表中结点按照姓名顺序地存储在计算机中，依次查找，可能遍历整个表都找不到。

（ 2 ）再建立一张姓氏索引表，姓 + 表中的起始地址。不需查找其他姓氏。查找效率提高。

例 2: 设计一个考试日程安排表，使在尽可能短的时间内安排完考试。

要求同一个学生选修的几门课程不能安排在同一个时间内。

姓名	选修 1	选修 2	选修 3
	A	B	E
	C	D	
	C	E	F
	D	F	A
	B	F	

- 解决这个问题，首先选择一个合适的数据结构。用无向图表示，图中的顶点表示课程，不能同时考试的课程之间连上一条边。则该问题就抽象成对该无向图进行“着色”操作，即用尽可能少的颜色去给图中每个顶点着色，使得任意两个相邻的顶点着不同的颜色。同一种颜色表示一个考试时间。
- 答案：1：A,C 2:B,D 3:E 4:F
- 解决问题的关键步骤是先选取合适的数据结构表示问题，才能写出有效的算法。

1.4 算法和算法分析

4. 算法设计的要求 p13 ~ p14

(1) 正确性

四层含义 p14 a,b,c,d

(2) 可读性

首先是给人读，然后才是机器执行

(3) 健壮性 容错性

(4) 效率与低存储量需求

1.4 算法和算法分析

- 定义：一个算法如果能在所要求的资源限制内将问题解决好，则称这个算法是有效率的。
- 例如，一个资源限制是：可用来存储数据的全部空间、运行时间
 - 可能是分离的内存空间限制和磁盘空间限制
 - 允许执行每一个子任务所需要的时间

1.4 算法和算法分析

5. 算法的分析 —— 算法性能的评价

- 评价标准：

- 1) 算法所需的计算时间
- 2) 算法所需的存储空间
- 3) 算法的简单性

- 度量算法执行时间的两种方法 p14

- 1) 事后统计法

 此方法有两个缺陷 p14

- 2) 事前分析估算法

 此方法取决于多个因素： p14

1.4 算法和算法分析

6. 时间复杂度

(1) 定义 :

一般情况下, 算法中基本操作重复执行的时间是问题规模 n 的某个函数 $f(n)$, 算法的时间量度记作

$$T(n) = O(f(n))$$

它表示随问题规模 n 的增大, 算法执行时间的增长率和 $f(n)$ 的增长率相同, 称作算法的渐进时间复杂度, 简称时间复杂度。

- 语句的频率: 该语句重复执行的次数。

(2) 引例

(a) `{ ++x; s=0 } ++x` 的频率为 1

(b) `for (i=1; i<=n; ++i) { ++x; s+=x; } ++x` 的频率为 n

(c) `for (j=1; j<=n; ++j)`

`for (k=1; k<=n; ++k) { ++x; s+=x; } ++x` 的频率为 n^2

- 理解: a) $n*n+4*n=O(n*n)$;

b) n 的平方或 n 的 2000 次方总没有 2 的 n 次方增长快

大 O 运算规则

规则 1 : $kf(n)=O(f(n))$ 大 O 忽略常数因子

规则 2 : if $f(n) = O(g(n))$ and
 $g(n)=O(h(n))$ then $f(n)=O(h(n))$ 传递性

规则 3 : $f(n) + g(n)=O(\max\{f(n),g(n)\})$

规则 4 : if $f1(n)=O(g1(n))$ and
 $f2(n)=O(g2(n))$ then
 $f1(n)*f2(n)=O(g1(n)*g2(n))$

1.4 算法和算法分析

- 最佳、最差、平均时间复杂度
 - 无法精确计算基本操作的执行次数（频度）时，只须求出其关于 n 的增长率
 - 与输入数据集有关，如冒泡程序，可考虑平均时间复杂度和最坏情况下的时间复杂度

1.4 算法和算法分析

表：时间复杂度和算法运行时间的关系（1986年代）

$T(n)$ n	$O(\log n)$	$O(n)$	$O(n \log n)$	$O(n^2)$	$O(n^3)$	$O(n^5)$	$O(2^n)$	$O(n!)$
20	4.3us	20us	86.4us	400us	8ms	3.2s	1.05s	771 世纪
40	5.3	40	213	1600	64ms	1.7min	12.7 天	2.59* 1032 世纪
60	5.9	60	354	3600	216ms	13min	366 世纪	2.64* 1066 世纪

1.4 算法和算法分析

结论

- (1) 当 $f(n)$ 为对数函数、幂函数、或它们的乘积时，算法的运行时间是可以接受的，称这些算法是**有效算法**；当 $f(n)$ 为指数函数或阶乘函数时，算法的运行时间是不可接受的，称这些算法是**无效的算法**。
- (2) 随着 n 值的增大，增长速度各不相同， n 足够大时，存在下列关系：

对数函数 < 幂函数 < 指数函数

1.4 算法和算法分析

$O(1)$ 常量阶，与 n 无关

$O(\log n)$ $\log n$ 阶

$O(n)$ n 阶

$O(n \log n)$ $n \log n$ 阶

$O(n^2)$ 平方阶

$O(n^3)$ 立方阶

$O(2^n)$ 指数阶

增长率由慢到快 图示见 p16 图 1-7

尽量少用指数阶的算法

- 说明：教材 p17 第 2 行：除特别指明外，均指最坏情况下
- 的时间复杂度

1.4 算法和算法分析

空间复杂度

- (1) 存储算法本身所占用的空间
- (2) 算法的输入 / 输出数据占用的空间
- (3) 算法在运行过程中临时占用的辅助空间
- 原地工作：若辅助空间相对于输入数据量是常数，则称此算法是原地工作。
- 若所占空间量依赖于特定的输入，按最坏情况分析。

思考及要求

1. 给出数据结构的概念，并对其进行深入的理解。
2. 通常所说的四类基本的数据结构是指什么？请分别给出一个具体的例子。
3. 给出数据结构通常的分类。
4. 关于算法
 - (1) 给出算法的定义，并对其五个重要特性进行理解；
 - (2) 给出一个“好”的算法应该达到的目标；
 - (3) 对算法“正确”的含义进行理解。
5. 请对以下概念进行理解：
 - (1) 时间复杂度
 - (2) 空间复杂度
 - (3) 抽象数据类型（的含义）
6. 请思考：
 - (1) 逻辑结构和物理结构的关系；
 - (2) 数据结构、逻辑结构、物理结构之间的关系；
 - (3) 对算法效率进行度量的方法。
7. 对 C 语言进行比较全面的复习。