数据结构2019



hwdong hw-dong

为什么Why

信息社会、数据为王

- 数据是一切信息技术:数据分析、人工智能、商业应用、科学研究、政治经济决策的基础。"巧妇难为无米之炊"
- 谁拥有了数据, 谁就拥有了未来。人工智能不在于算法而在于数据。

程序=数据结构+算法

- 数据结构研究如何合理组织、存储和处理数据。
- 良好的数据结构规定了数据处理的效率,高效的算法才能高效地处理数据。

数据结构决定了数据处理的效率

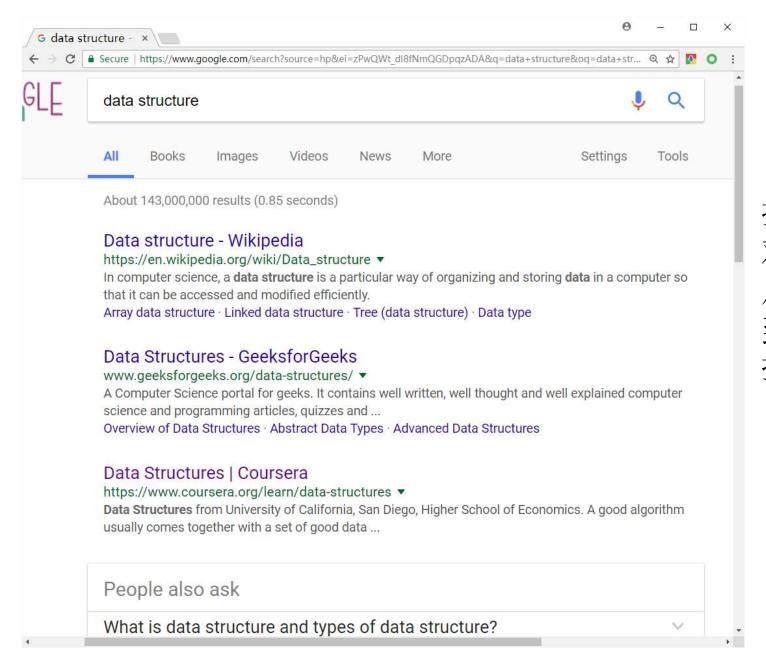
•一伙人去买东西,不能乱哄哄的,可以排成一队。

• 衣橱里的衣服、书柜里的书,不能杂乱无章,应该有条理的摆放,

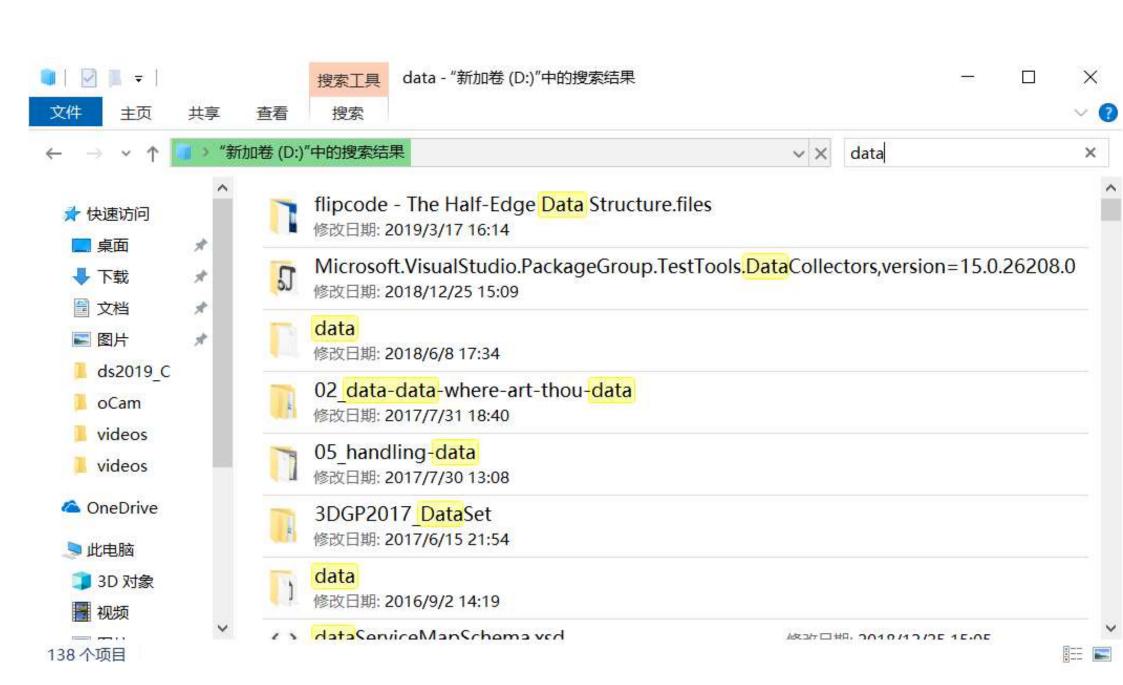
才能方便你寻找。







搜索引擎如何时刻面 对每秒几亿次点击, 从大量数据中快速找 到您要搜索的信息并 排序?









一新宁易则





¥ **1650.00 包邮** 6233人付款 分期免息Apple/苹果 iPhone 7 苹果7plus美

数据结构: 计算机学科的王者

- 数据结构就是研究以何种方式存储、组织数据,以便高效率地处理这些数据。处理数据的步骤(过程)-算法。
- 熟悉精通数据结构才能设计和开发高效率的软件系统,是软件工程师的基本内功。所有软件系统都离不开数据结构。不懂数据结构的称为"码农",懂数据结构的才是程序员、计算专家。如同普通泥瓦工和建筑设计师。
- 硅谷、国内BAT等企业面试软件编程人员的面试题基本都是数据 结构方面的题目。 刷题

只要学会一门编程语言和数据结构就足够了,其他的操作系统、计算机网络网上看一下大致知道一下基本原理就行了。比如操作系统就是进程、内存、设备的管理,计算机网络就是通信协议TCP/IP了解一下。软件工程没啥用途,编译原理基本用不上。

@有个梨UGlee •

学计算机的捷径是有的,就是传统上说的计算机专业的几门课,算法和数据结构结合实际语言看个简单的就行,操作系统和计算机网络都看Tannenbaum的。编译原理实战用不上,但龙书前半本是需要找时间看的,不然永远无法真正理解什么是计算机语言(当然也不是看了这本就懂了,但起码比没看提高很多。图形学和软件工程看自己兴趣吧。这些看完了就是前面的帖子说的如何通过Linux全面学习计算机系统了。

具体学一门编程语言,其实是最不重要的。很多写了十年Java和C++的,碰到流和并发还写不对,他要是读过一些内核代码可能早就迈过这个坎了。几乎没人会在编程里第一个遇到某种问题。所以找到别人之前怎么解决的比自己悟天悟地悟空气重要得多。

研究什么What

• 数据元素之间的关系、存储表示、处理算法。

各种信息管理系统、数据库,数据以表(员工表、货物表、学生 名单、股票行情、聊天记录)的形式表示。

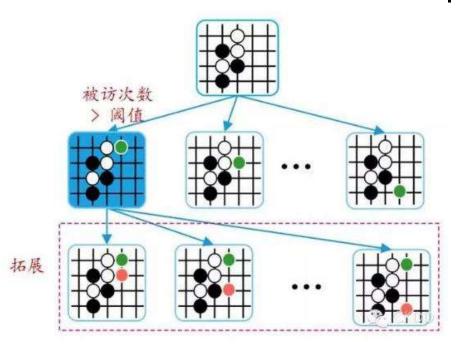


各种信息管理系统、数据库,数据以表(员工表、货物表、学生 名单、股票行情、聊天记录)的形式表示。

| 学号 | 姓名 | 性别 | 籍贯 | 年龄 |
|-------|-------|----|----|----|
| 98131 | 张三 | 男 | 北京 | 20 |
| 98164 | 李斯 | 女 | 上海 | 21 |
| 98165 | 王武 | 男 | 广州 | 19 |
| 98182 | 赵柳 | 女 | 香港 | 22 |
| 98224 | • • • | | | |

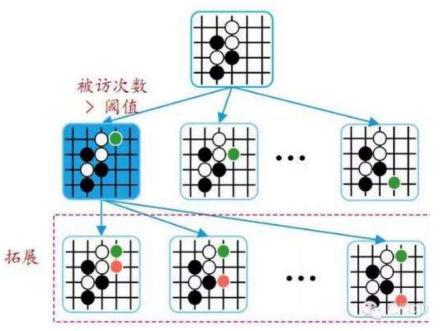
线性表:数据元素之间是一对一的关系。 $(a_1, a_2, \dots a_n)$

• 树型结构:数据元素之间是一对多的层次结构。如组织结构、家族族谱、基因图谱、棋局(AlphaGo)。



AlphaGo如何对棋局的大量状态 空间进行快速搜索?

树型结构:数据元素之间是一对多的层次结构。如组织结构、家族族谱、基因图谱、棋局(AlphaGo)。



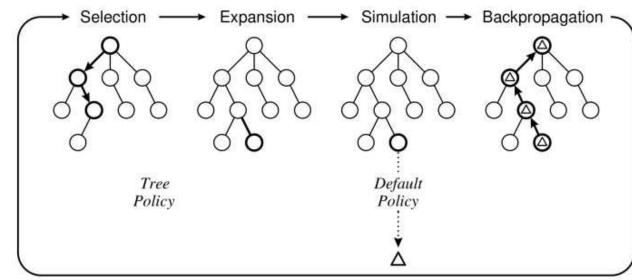
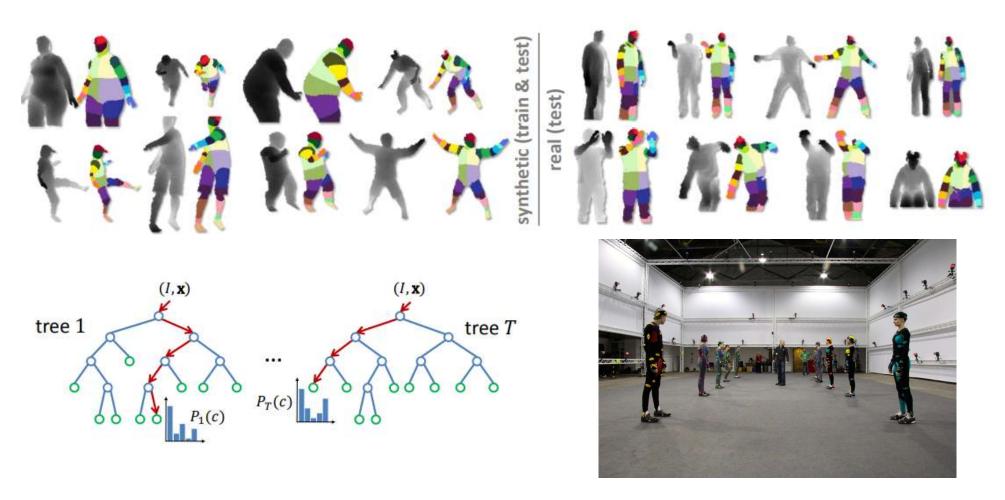


Fig. 2. One iteration of the general MCTS approach.

蒙特卡洛树搜索



随机森林: Kinect实时从深度图像中检查人体骨骼

Data-driven Fluid Simulations using Regression Forests

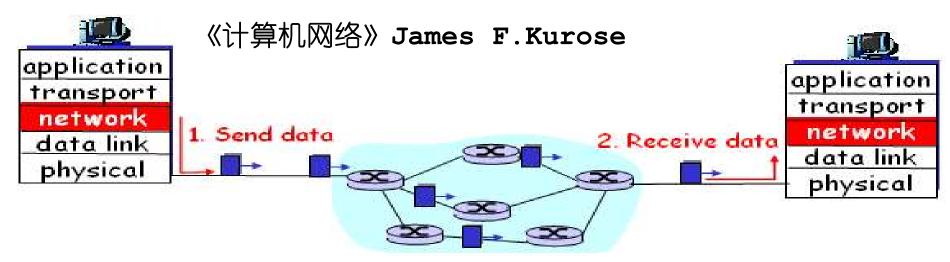
L'ubor Ladický*¹, SoHyeon Jeong*¹, Barbara Solenthaler ¹, Marc Pollefeys ¹ and Markus Gross ^{1,2} (*Joint First Author)

¹ETH Zurich, ²Disney Research Zurich

• 图形结构:数据元素之间是多对多的层次结构。计算机网络、社交网络、web网页、交通图、...



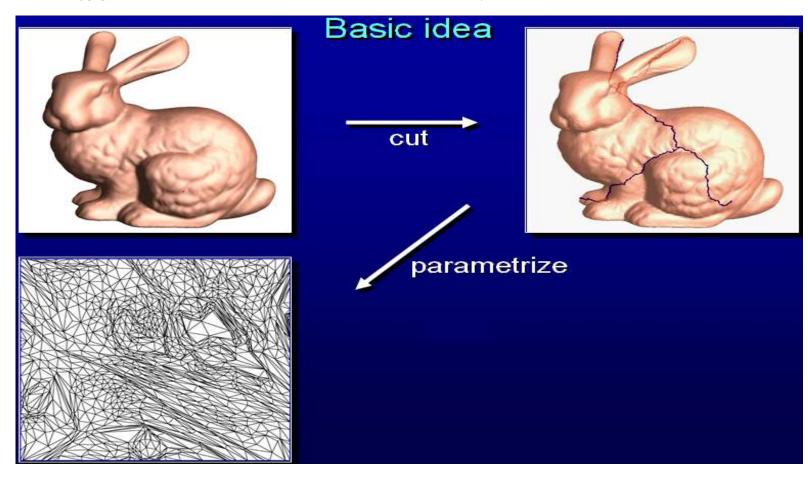
最短路径问题: 路由算法 导航 网络路由 (Routing Protocal):数据报从源ip地址以最低代价 传送到目的ip地址?



Global:

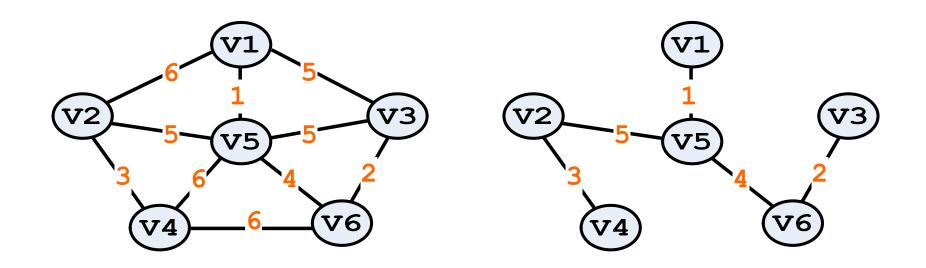
Link-State Routing Algorithm - Dijkstra Decentralized: Distance Vector Routing Algorithm

3.3 D网格模型的测地线-两点间的最短距离

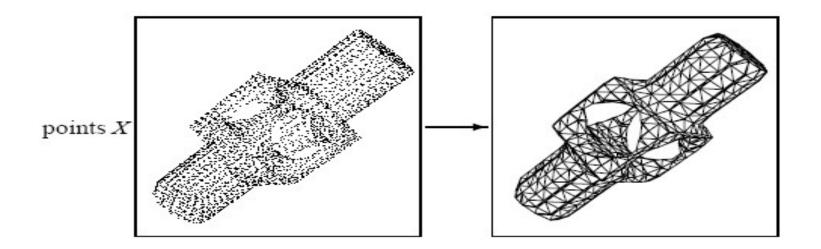


最小生成树

• 图形结构:数据元素之间是多对多的层次结构。计算机网络、社交网络、web网页、交通图、...



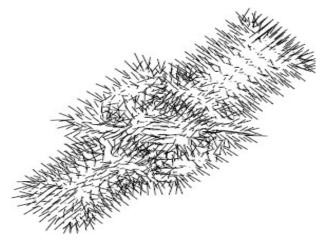
• Hooppe曲面重建算法(Surface reconstruction from unorganized points.H. <u>Hoppe</u>, T. <u>DeRose</u>, T. <u>Duchamp</u>, J. <u>McDonald</u>, W. <u>Stuetzle</u>. *ACM SIGGRAPH 1992*, 71-78.)



• 用最小生成树一致化所有点的法矢量。然后用MarchingCube抽取等值面



Riemannian图



一致的法矢量

课程内容

线性表、栈和队列、串和数组、树和二叉树、优先队列和堆、图、 查找、排序等

如何How

课程特点

- 化难为易:使复杂困难的算法变得简单易懂。动画、实例说明原理,代码执行动画演示
- 内容丰富:内容全面(涵盖考研的所有数据结构内容)、介绍数据结构的应用案例、并有习题和刷题。
- 间接明快: 无废话和抖机灵, 都是干货。
- 原理+代码实现:大学教学中经常只讲原理,没有具体代码实现。

关注

https://a.hwdong.com

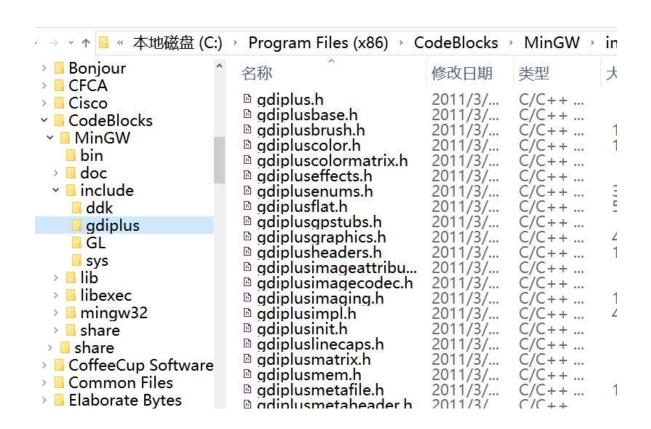
B站或微博: hw-dong

网易云课堂: hwdong

腾讯课堂: hwdong.ke.qq.com

QQ群: 101132160

电脑里的文件可以放在不同文件夹,文件夹里还包含子文件夹, 查找时可以按照名字或时间排序来查找。



数据结构概述

- 什么是数据结构
- •数据元素、数据项
- 逻辑结构和物理结构
- 算法性能分析

C+版数据结构



hwdong

B站 和微博: hw-dong

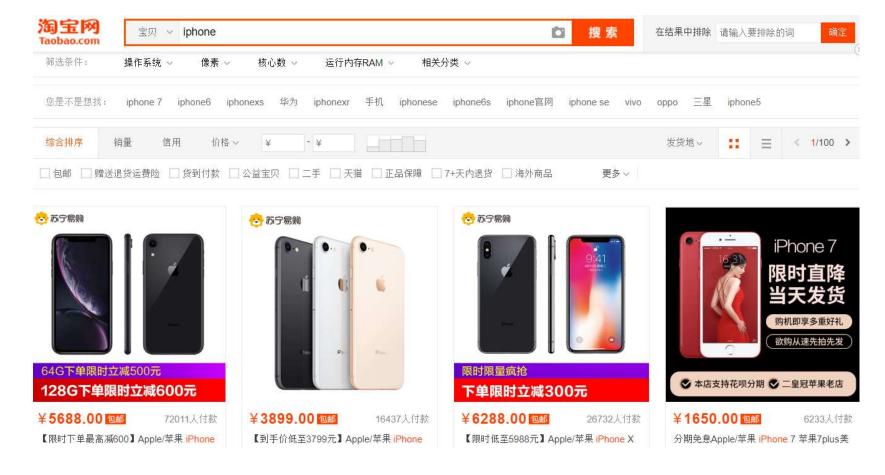
什么是数据结构

工资单、货物清单、学生名单等放在一张表,并按照姓名或编号排序。那么查找效率就会很高。

| 学号 | 姓名 | 性别 | 籍贯 | 年龄 |
|-------|-------|----|----|----|
| 98131 | 张三 | 男 | 北京 | 20 |
| 98164 | 李斯 | 女 | 上海 | 21 |
| 98165 | 王武 | 男 | 广州 | 19 |
| 98182 | 赵柳 | 女 | 香港 | 22 |
| 98224 | • • • | | | |

什么是数据结构

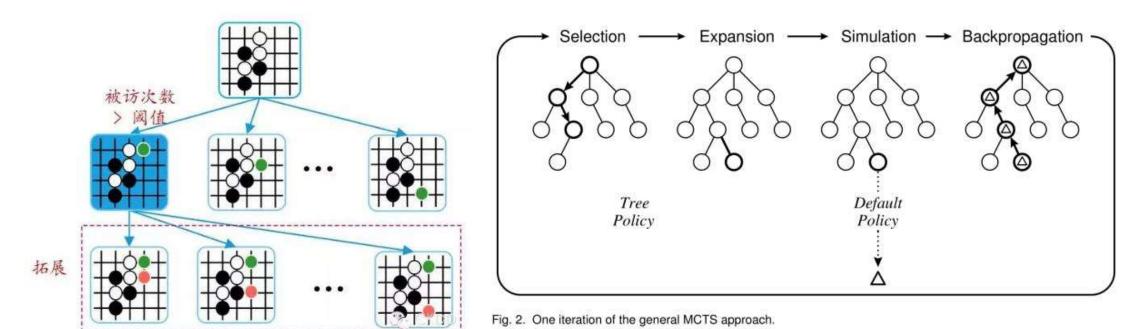
• 电子商务购物



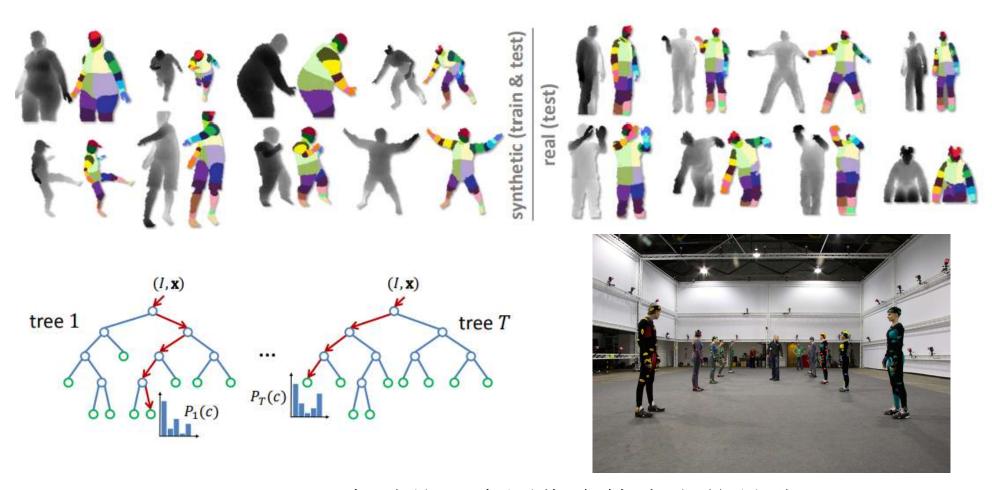
什么是数据结构

• 地图, 地点之间通过道路相连, 构成一个几何图, 可以用图的算法求解交通查询如最短路线等。





AlphaGo如何对棋局的大量状态 空间进行快速搜索?



Kinect实时从深度图像中检查人体骨骼

什么是数据结构

- 计算机处理的数据类型各种各样,有文本、图像、视频等。
- 大数据时代数据越来越多,如果没有一个合理的结构组织这些数据,软件系统将不能有效处理这些数据。
- 数据结构研究对象:由很多同类型数据元素构成的数据集(大数据)
- 研究数据元素之间的关系及其存储表示,如何高效地对这些数据进行处理(操作)。

课程内容

- 线性表
- 栈和队列
- 字符串、多维数组
- 树和二叉树
- 图
- 查找
- 优先队列和堆
- 排序

面向面试或考研的刷题

从动画例子到原理,最后实战编程

数据元素和数据项

- 本课程主要研究同类型数据元素构成的数据集。数据元素描述一个具体对象的所有属性(属性也称为"数据项"),也称为"记录"。
- 数据元素是数据的基本单位,在程序中常作为一个整体考虑和处理。 理。

 登录号
 书名
 借阅者编号

 001
 理论力学
 9002

 (一条记录)
 高等数学
 9001

 ...
 ...
 ...

数据元素和数据项

•数据项:数据的不可分割的最小单位

•一个数据元素可由一或多个数据项构成

3个数据项

| 登录号 | 书名 | 借阅者编号 |
|-------|-------|-------|
| 001 | 理论力学 | 9002 |
| 002 | 高等数学 | 9001 |
| • • • | • • • | • • • |

逻辑结构和物理结构

数据结构主要从2个方面研究如何有效地存储组织和处理数据。

•逻辑结构:从抽象/高层角度来描述数据的特征和操作的抽象含义。

物理结构:从具体实现的角度讨论数据及其逻辑操作在计算机的表示和实现。

- 数据的数学/逻辑模型: 从抽象角度刻画数据的特点和操作
- •比如"整数"

1, 2, ...

运算:加、减、乘、除等

不涉及是1个苹果,还是 2个人,2斤面粉,....

- 数据的数学/逻辑模型: 从抽象角度刻画数据的特点和操作
- •比如抽象的看"手机"是一个什么东东?

一个电子设备,有外壳、触 摸屏、开关按键、触摸屏、 摄像头 操作:

开关机 拨打电话 收发短信 拍照 使用APP



手机是一个抽 象概念,不涉 及厂家品牌的 具体手机。

- 数据的数学/逻辑模型: 从抽象角度刻画数据的特点和操作
- 数据元素(比如各种表格中的记录)之间在抽象/逻辑上看具有 什么样的关系?

线性(一对一:一个挨着一个) 棋局(一对多)

• 对数据有哪些操作: 这些操作的抽象/逻辑含义

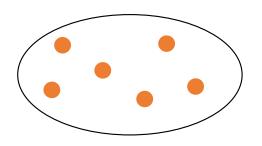
插入 删除 读取 修改 查询

而不关心比如删除一个人还是一条记录

• 从抽象角度刻画数据元素之间的关系和对数据进行的操作.

•数据元素之间的关系。有4种:集合、线性、树型、图型或网状。

•集合关系:数据元素属于一个集合,它们之间无特定关系

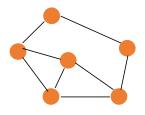


•数据元素之间的关系。有4种:集合、线性、树型、图型或网状。

•线性关系:一个接一个。比如学生花名册中的记录之间

•**树型**:一对多。比如资源管理器中的文件夹、组织结构、家族族谱、决策树、状态树

• 图型: 多对多。比如交通图、计算机网络、社交关系



- 从抽象角度刻画数据元素之间的关系和对数据进行的操作
- 从抽象角度描述的数据结构,称为"抽象数据类型"(Abstract data type),简称为ADT。
- •比如一个ADT叫做"线性表或列表" (List)。

线性关系: (a₁, a₂, a₃, ...)

- 存储(store) 一系列数据元素
- 读(read) 某个位置的数据元素
- 修改(modify) 某个位置的数据元素

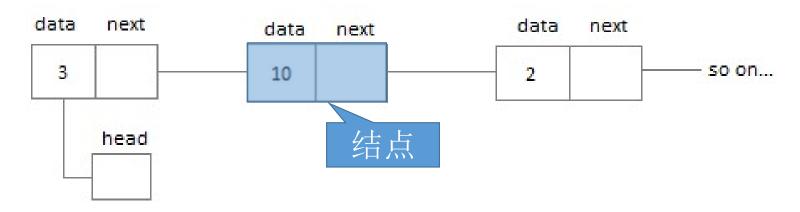
物理结构

- 数据及其操作在计算机内部的表示和实现
- ·如C语言的数组(array)就是ADT类型的"线性表"的具体实现

```
int array[10];
array[2] = 30; /*通过下标位置修改*/
printf("%d",array[1]);
```

物理结构

- 数据及其操作在计算机内部的表示和实现
- ADT类型的"线性表"也可以用链表实现

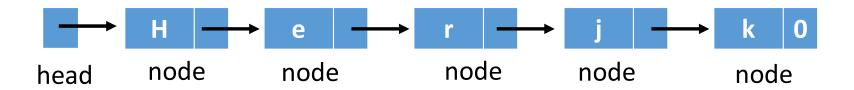


同一个ADT可以有不同的物理实现: 数组、链表

物理结构

k

- 存储结构数据元素及其关系在计算机中存储
- 顺序映象(顺序存储结构),如数组 H e U以存储地址的相邻性表示数据元素间的逻辑关系
- 非顺序映象(链式存储结构),如链表通过指示信息表示数据元素间的逻辑关系



算法性能分析

- 算法是对数据处理过程的步骤/指令的描述, 是指令的有限序列
- 算法的特性:

有穷性:有穷步,每步有穷时间内完成

确定性:每个步骤都有确切的含义,相同的输入具有相同的执

行路径和结果

可行性: 每个步骤或指令应该是可行的

输入输出:有输入数据,产生或输出执行结果

算法性能分析

• 好的算法应当满足:

正确性:算法应能满足具体问题的需求

可读性: 算法应易于阅读和理解

健壮性: 输入数据非法时, 算法也能适当作出反应或进行处理

高效性: 算法执行时间短, 占用存储空间少

•程序:算法在计算机上的实现

•程序的执行时间取决于如下因素:

算法本身

问题的规模

编程语言

编译程序

硬件性能

同一个算法程序在不同的语言、编译程序和硬件的条件下,执行时间是不同的.

比如:算法A在硬件A上执行时间为1秒,算法B在硬件B上执行时间为2秒,并不能因此就认为算法A的效率更高

•程序:算法在计算机上的实现

•程序的执行时间取决于如下因素:

算法本身

问题的规模

编程语言

编译程序

硬件性能

同一个算法在不同的语言、编译程序和硬件的条件下,执行时间是不同的.

所以评价算法的性能应当排除这三者 的影响

只需要考虑算法本身和问题的规模

算法时间效率的量度

• 事后统计法: 测量一个算法执行所需要的时间

缺点:

需要编写测试程序 测量结果依赖于具体的软、硬件

• 事前分析估算法

• 选择一个"基本操作",分析基本操作执行的次数。

```
for(i=1; i<=n; ++i)
  for(j=1; j<=n; ++j) {
     c[i][j]=0;
     for(k=1; k<=n; ++k)
        c[i][j]+=a[i][k]*b[k][j];
}</pre>
```

• 执行次数 F = n³

- 选择一个"基本操作",分析基本操作执行的次数。
- 用该基本操作的重复次数表示算法的执行时间,一般为问题规模 n 的函数 f(n), 简称"频度"。
- 算法的时间复杂度为:和f(n)同阶的简化无穷大量
 T(n)=O(f(n))

• 例1

```
for (i=2; i<=n; ++i)
  for (j=2; j<=i-1; ++j) {
      ++x;
      a[i][j] = x;
}</pre>
```

$$f(n) = \frac{(n-1)(n-2)}{2} = \frac{n^2 - 3n + 2}{2}$$

$$T(n) = O(f(n)) = O(n^2)$$

• 例2

```
for(i = 1; i <= n; i ++) {</pre>
                                          执行次数=n,
                                          T(n) = O(n)
     for(j = i; j \le n; j ++)
                                             执行次数都是
         if (data[j] < data[m])</pre>
                                             n(n+1)/2,
                m = j;
                                             T(n) = O(n^2)
           if (m != i)
                temp = data[m];
                data[m] = data[i];
                data[i] = temp;
```

$$T(n)=O(n^2+n)=O(n^2)$$

矩阵运算的时间复杂度

• 两个n维向量a=(a₁,a₂,...a_n)^T, b=(b₁,b₂,...b_n) ^T的点积

$$(a_1 \ a_2 \ \dots \ a_n) \begin{pmatrix} b_1 \\ b_2 \\ \dots \\ b_n \end{pmatrix} = a_1 b_1 + a_2 b_2 + \dots + a_n b_n$$

2n-1(flops)=n(次乘法)+(n-1)(次加法)

矩阵运算的时间复杂度

设A是mxn矩阵,x是n维向量,则Ax是一个m维向量,每个元素由 A的一行向量和x点乘得到

$$\begin{pmatrix}
- & - & \dots & - \\
a_1 & a_2 & \dots & a_n \\
- & - & \dots & -
\end{pmatrix}
\begin{pmatrix}
x_1 \\
x_2 \\
\dots \\
x_n
\end{pmatrix}$$

m(2n-1) (flops)

矩阵运算的时间复杂度

设A是mxn矩阵, B是nxp矩阵,则AB是mxp矩阵,每个元素由A的一行和B的一列两个n维向量点乘得到

$$\begin{pmatrix} - & - & \dots & - \\ & & & \\ \alpha_{i1} & \alpha_{i2} & \dots & \alpha_{in} \\ - & - & \dots & - \end{pmatrix} \begin{pmatrix} - \dots & b_{1j} & \dots & - \\ - \dots & b_{2j} & \dots & - \\ - \dots & b_{nj} & \dots & - \end{pmatrix}$$

mp(2n-1) (flops)

(AB)x ? A(Bx)

• 那种方式更快?

A是mxn矩阵,B是nxp矩阵,则AB是mxp矩阵,x是px1矩阵.

AB: mp(2n-1)

(AB)x: mp(2n-1) + m(2p-1)

Bx: n(2p-1)

A(BX): n(2p-1) + m(2n-1)

```
A = randn(n,n); B = randn(n,n); x = randn(n,1);

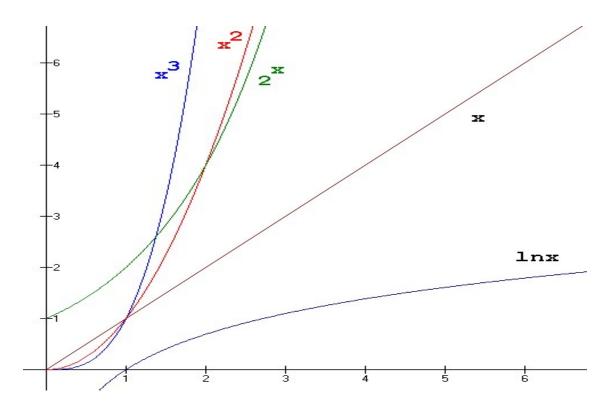
t1 = cputime; y = (A*B)*x; t1 = cputime - t1;

t2 = cputime; y = A*(B*x); t2 = cputime - t2;
```

The table shows the results on a 2.8GHz machine for eight values of n.

| n | time (sec.) | time (sec.) |
|------|-------------|-------------|
| | method 1 | method 2 |
| 500 | 0.10 | 0.004 |
| 1000 | 0.63 | 0.02 |
| 1500 | 1.99 | 0.06 |
| 2000 | 4.56 | 0.11 |
| 2500 | 8.72 | 0.17 |
| 3000 | 14.7 | 0.25 |
| 4000 | 34.2 | 0.46 |
| 5000 | 65.3 | 0.68 |

• 各种常见的渐进复杂度, 如: aⁿ>n^b>log_cn



- 算法执行时所需存储空间的量度,记作S(n), 其中 n 为问题的规模.
- 一般不考虑存放数据本身占用的空间,只考虑执行算法所需辅助空间,除非数据所占空间与算法本身有关.

•排序: $(a_1, a_2, ..., a_n)$

冒泡排序: 1个数据元素的辅助空间

归并排序: n个数据元素的辅助空间

时间复杂度

 $O(n^2)$

 $O(n\log n)$

算法性能分析

- 有些算法的复杂度与输入数据有关
- •如冒泡排序,当输入数据基本有序时,其时间复杂度为O(n),基本无序时,为O(n²),平均为O(n²)
- 此时就应该分最好情况、最差情况、平均情况来讨论

C+语言

- 变量即用即定义
- 引用变量
- 指针和指针变量: p++, p[i], *(p+i)
- typedef
- struct
- 内存分配与释放: malloc, realloc, free
- 函数指针

关注

https://a.hwdong.com

B站或微博: hw-dong

网易云课堂: hwdong

腾讯课堂: hwdong.ke.qq.com

QQ群: 101132160