第 8 讲 怎样研究算法—排序算法研究示例

1、快速浏览---本讲视频都讲了什么?

【视频 8.1 为什么要研究排序算法--结构化数据表查找问题】

排序算法是计算学科最基础的、用途最广泛的算法，也是构造其他高效算法的基础算法。为什么这么说呢，本段视频以结构化数据表查找问题为例，揭示了“已排序数据的处理”和“非排序数据的处理”在效率方面有什么差异---进而揭示了排序算法的重要性。请看视频是如何理解“数据表查找”中的排序问题的… …。

【视频 8.2 为什么要研究排序算法--非结构化的数据文档查找问题】

本段视频与视频 8.1 一样，也是从问题理解角度阐述排序算法的重要性，但更为有特色。这里举了一个非结构化数据文档的查找问题，比如论文、网页等全文数据库的检索问题，如何快速的检索呢，这里面的排序问题又是怎样的呢？怎样构建索引与排序索引？怎样确定一篇文章/论文的关键字？它们和排序有什么关系呢？请看视频是如何解说的……。

【视频 8.3 基本排序算法-内排序】

在理解了排序问题后，本段视频给出了三种基本的内排序算法的讲解，分别是：插入法排序、选择法排序和冒泡法排序，分别从算法思想、算法描述和模拟执行排序算法的执行过程与结果进行了讲解……

【视频 8.4 受限资源约束下的算法-内排序与外排序问题】

排序算法需要消耗资源，即内存，而内存容量是有限的，待排序的数据可能是无限的，无限的数据怎样利用有限的资源进行排序呢，这时就要用到存储体系中的外存，利用外存辅助排序，由此出现了内排序问题和外排序问题，二者思考的出发点是不同的，有哪些不同呢？请看视频…。

【视频 8.5 基本排序算法-外排序】

本段视频给大家介绍了一种外排序算法—多路归并排序算法。如何基于磁盘，并充分利用内存排序大规模数据的算法。在此算法中展现了如何充分利用内存资源的思想，通过一步步模拟算法执行的过程展现算法的基本思想，并进行了各种情况下的讨论，…

【视频 8.6 PageRank 网页排序算法】

本段视频从另外的角度介绍了排序问题的求解方法：即网页的排序问题，其排序算法 PageRank 是一个著名的算法。该算法展现了如何从问题语义挖掘求解思想，如何将求解思想表达成数学模型，如何通过迭代计算进行求解的基本思想。最难能可贵的是，该算法通过迭代进行求

解和数学上利用特征方程进行求解的一致性，说明了算法研究中数学的重要性。

2、学习要点指南

2.1 要点一：理解怎样研究算法，理解由问题到算法，由问题到环境再到算法设计

本讲内容是算法思维三部分内容之一，目的是使学生准确理解“问题➔算法”以及“问题➔资源➔算法”，即理解问题与算法的关系，强化对问题的理解，由问题深入理解而发现算法，而不是为算法而算法，理解问题、算法与资源环境的关系，强化学生对受限资源约束下的算法构造而不是无限资源约束下的算法构造。本讲以排序问题为例阐述上述内容。

算法研究首先要理解问题、理解问题求解的需求。我们从结构化数据查找与统计和非结构化数据查找与搜索两个方面来看，排序是解决这两个问题的重要手段：一方面，相比于建立在非排序数据基础上的算法而言，建立在排序数据基础上的算法可能效率更高；另一方面，对许多复杂的问题，排序可能是问题求解必需的先行手段。

此外对于算法研究，不仅仅是找到该问题的一个算法，而且还要寻找更快速的算法。但更重要的是要分析算法所适应的环境，例如当待排序对象能够完全装入内存时如何排序—**内排序**，和当待排序对象数据量特别大不能完全装入内存时如何排序--**外排序**。本讲尤其以外排序为例，展现了怎样才能发现更好的算法。

进一步，本章以网页排序算法 PageRank 为例，介绍了如何从不同视角挖掘问题求解思想与算法，如何由问题到数学：网页排序--网页重要度计算 ➔ 网页链接：正向链接和反向链接 ➔ 从问题语义挖掘求解思想：反向链接数越多越重要，反向链接有权重，一个网页反向链接的权重由指向该网页的网页重要度和其正向链接来确定 ➔ 表达成数学：0,1-矩阵、权值矩阵(转移概率矩阵) ➔ 如何求解：求稳定性 ➔ 数学解法：特征方程及其含义。通过由网页重要度的计算到数学的特征方程含义的探讨，给出了由问题到数学的一个典型案例。

2.2 要点二：内排序与外排序算法的关注点上的差异

本讲介绍了三个内排序算法和一个外排序算法。内排序算法通常是内存上应用的算法，由于其可能被反复使用，因此其关注点强调尽可能地减少算法的步骤。虽然三个内排序算法复杂性都是 $O(n^2)$，但其执行时间上还是有细微差别的，不仅仅是算法思想的不同，能否细致地区分三个算法很关键：以递增排序为例，选择法每轮次仅比较，没有交换，直至找到最小值(或最大值)后做一次交换，而冒泡法每一轮次是通过依次比较相邻两个元素的方法来找最小值(或最大值)，如果前一元素比后一元素大(或小)，则交换前后两个元素，交换可能频繁发生。则选择法可能比冒泡法速度快。

然而外排序由于涉及到磁盘读写问题，一次磁盘读写可能即相当于完整地执行一遍或多遍的内存数据排序，因此其关注点在如何尽可能的降低磁盘读写的次数。围绕尽可能少的读写磁盘，

即最大可能的利用内存来进行算法的设计。通过外排序算法的讨论，大家可以体验到算法如何尽可能的利用受限的资源的。

2.3 要点三：怎样理解 PageRank 算法

PageRank 算法的理解可以分几个层面：

(1)问题求解思想层面

一般，在这一层面大家都能理解，只是想到想不到的问题，有些人可能认为太简单，也有些人可能不太相信，但其实算法就应是很简单的。这一层面即算法是如何通过问题语义的理解逐渐挖掘算法的求解思想的：按反向链接的个数确定重要度 ➔ 按反向链接的加权和确定重要度 ➔ 权值如何确定呢，由正向链接确定，正向链接数越多，权值越小。

这一层面关键是要注意区分概念：正向链接和反向链接。

(2)算法的表示及计算层面

这一层面是要理解用迭代计算的方法计算网页重要度。首先将网页之间的链接关系表达成矩阵的形式 ➔ 注意区分水平方向/垂直方向是正向链接/反向链接 还是反向链接/正向链接，后者与矩阵乘法的规则相一致 ➔ 将矩阵转换成权值形式 ➔ 观察矩阵乘法得到的结果正好是反向链接的加权和 ➔ 设定网页重要度的初值，按矩阵乘法计算反向链接的加权和得到下一次的网页重要度，如此迭代计算下去 ➔ 初值如何设定呢，其实可任设初值，因为不断的迭代直到第 n 次迭代结果与第 n-1 次迭代结果相同为止，可计算出网页重要度。

(3)算法的数学计算求解层面

解释算法求解与数学特征方程求解之间的关系。

3、常见问题

略。