



# 数据结构与算法（十一）

张铭 主讲

采用教材：张铭，王腾蛟，赵海燕 编写  
高等教育出版社，2008.6（“十一五”国家级规划教材）

<http://www.jpk.pku.edu.cn/pkujpk/course/sjjg>



## 主要内容

- 基本概念
- 11.1 线性索引
- 11.2 静态索引
- 11.3 倒排索引
- 11.4 动态索引
- 11.5 位索引技术
- 11.6 红黑树



## 11.3 倒排索引 ( Inverted Index )

- 11.3.1 基于属性的倒排
  - 要求检索结构中某个或若干个属性满足一定条件的结点 ( 不是按关键码的值检索 )
  - 按照属性建立索引
- 11.3.2 对正文文件的倒排
  - 以文中的词 ( word ) 为索引项建立的索引



## 11.3 倒排索引

## 教师数据库主表

| EMP#  | NAME | Department | Profession | Specialty | Address |
|-------|------|------------|------------|-----------|---------|
| 0155  | 李宇   | 数学         | 教授         | 代数        | C105    |
| 0421  | 刘阳   | 外语         | 助教         | 英语        | E310    |
| 0208  | 赵亮   | 物理         | 助教         | 力学        | C211    |
| 0211  | 张伟   | 物理         | 讲师         | 原子物理      | D508    |
| 0132  | 王亮   | 数学         | 助教         | 几何        | E220    |
| 0119  | 王卓   | 数学         | 讲师         | 代数        | B102    |
| 0330  | 孙丽   | 计算机        | 教授         | 软件        | A108    |
| 0455  | 刘珍   | 外语         | 讲师         | 法语        | A225    |
| 0310  | 周兵   | 计算机        | 讲师         | 英语        | B423    |
| 0341  | 何江   | 计算机        | 助教         | 计算机       | F406    |
| ..... |      |            |            |           |         |



## 11.3.1 基于属性的倒排

- 对某属性按属性值建立索引表，称倒排表
- “属性 - 指针”对 (attr, ptrList)
  - （属性值，具有该属性值的各记录指针）
  - 指针可以是关键码，或该记录的主文件地址
- 颠覆主文件的顺序，因而称为倒排索引
- 属性往往是离散型的
  - 对于连续型的索引，往往用B树
- 倒排文件：带有倒排索引的文件

## 11.3.1 基于属性的倒排

## 教师数据库倒排表

| Department list | EMP#                   |
|-----------------|------------------------|
| 数学              | 0155, 0132, 0119       |
| 物理              | 0208, 0211             |
| 计算机             | 0330, 0310, 0341       |
| 外语              | 0421, 0455             |
| Profession list | EMP#                   |
| 教授              | 0155, 0330             |
| 讲师              | 0211, 0119, 0455, 0310 |
| 助教              | 0421, 0208, 0132, 0341 |
| Specialty list  | EMP#                   |
| 代数              | 0155, 0119             |
| 几何              | 0132                   |
| 力学              | 0208                   |
| 原子物理            | 0211                   |
| 软件              | 0330, 0341             |
| 英语              | 0421, 0310             |
| 法语              | 0455                   |



## 优缺点

- 优点：
  - 能够对于基于属性的检索进行较高效率的处理
- 缺点：
  - 花费了保存倒排表的存储代价
  - 降低了更新运算的效率



## 11.3.2 对正文文件的倒排

- 正文索引 (Text Indexing ) 处理的就是“建立一个数据结构以提供对文本内容的快速检索”
- 方法
  - 词索引 ( word index )
  - 全文索引 ( full-text index )





## 词索引

- 基本思想：
  - 把正文看作由符号和词所组成的集合，从正文中抽取关键词，然后用这些关键词组成一些适合快速检索的数据结构。
- 适用于多种文本类型，特别是那些可以很容易就解析成一组词的集合的文本
  - 适用于英文
  - 中文等东方文字要经过“切词”处理



## 全文索引

- 基本思想：
  - 把正文看作一个长的字符串
  - 在数据结构中记录的是子字符串的开始位置
  - 查询就可以针对正文中的任何子字符串
- 可以对每一个字符建立索引，从而使查询词不再限于关键词
- 需要更大的空间



## 倒排文件使用最广泛的是词索引

- 词索引使用 **最广泛**
- 一个已经排过序的关键词的列表
  - 其中每个关键词指向一个倒排 (posting list)
    - 指向该关键词出现文档集合
    - 在文档中的位置



## 倒排索引建立示例

正文文件：由6个文档组成，每个文档都是长字符串

| 文档编号 | 文本内容                                      |
|------|---|
| 1    | Pease porridge hot, please porridge cold, |
| 2    | Pease porridge in the pot,                |
| 3    | Nine days old.                            |
| 4    | Some like it hot, some like it cold,      |
| 5    | Some like it in the pot,                  |
| 6    | Nine days old.                            |

| 文档编号 | 文本内容                                      |
|------|---|
| 1    | Pease porridge hot, please porridge cold, |
| 2    | Pease porridge in the pot,                |
| 3    | Nine days old.                            |
| 4    | Some like it hot, some like it cold,      |
| 5    | Some like it in the pot,                  |
| 6    | Nine days old.                            |

类似处理1中的其他词语

同理，处理2中的词语

依次处理所有文档

# 倒排索引

| 编号 | 词语       | ( 文档编号 , 位置 )       |
|----|----------|---------------------|
| 1  | cold     | (1,6)               |
| 2  | days     |                     |
| 3  | hot      | (1,3)               |
| 4  | in       |                     |
| 5  | it       |                     |
| 6  | like     |                     |
| 7  | nine     |                     |
| 8  | old      |                     |
| 9  | pease    | (1,1) (1,4) ( 2,1 ) |
| 10 | porridge | (1,2) (1,5)         |
| 11 | pot      |                     |
| 12 | some     |                     |
| 13 | the      |                     |



## 建立正文倒排文件

1. 对文档集中的所有文件都进行分割处理，把正文分成多条记录文档

切分正文记录取决于程序的需要

定长的块、段落、章节，甚至一组文档



## 建立正文倒排文件（续1）

### 2. 给每条记录赋一组关键词

以人工或者自动的方式从记录中抽取关键词

停用词( Stopword )

抽词干( Stemming )

切词 ( Segmentation )



## 中文切词 Chinese Segmentation

我知道你不知道我知道你不知道我知道你不知道

- 我知道，你不知道。我知道，你不知道我知道，你不知道
- 我知道你，不知道我。知道你不知道我，知道你不知道
- 我，知道你不知道我知道。你，不知道我知道你不知道





教育部部长周济说，今后若干年内，毕业生总量将会持续增加，每年都有数十万的增量，对毕业生就业工作无疑是巨大的挑战。2005年高校毕业生总量大、增幅高，各地和高校工作进展情况差异较大，并且还存在许多深层次的矛盾和问题。在全社会就业形势十分严峻的形势下，2005年高校毕业生就业工作压力十分突出。

操作选项

☒ 词语切分☐ 一级标注☐ 二级标注

输出格式

☒ 北大标准☐ 973标准☐ XML

运行

处理文件...

退出

关于...

1

结果

平滑参数:

0

当前结果评分:

-101.911148

处理用时:

10

ms

教育部 部长 周济 说 ， 今后 若干 年 内 ， 毕业生 总量 将 会 的 持续 增加 ， 每年 都 有 数十 万 的 增量 ， 对 毕业生 就业 工作 无疑 是 巨大 的 挑战 。 2005 年 高 校 毕业 生 总量 大 、 增幅 高 ， 各地 和 高校 工作 进展 情况 差异 较 大 ， 并 且 还 存在 许多 深 层次 的 矛盾 和 问题 。 在 全 社会 就业 形势 十分 严峻 的 形势 下 ， 2005 年 高 校 毕业 生 就业 工作 压力 十分 突出 。



## 建立正文倒排文件（续2）

3. 建立正文倒排表、倒排文件  
得到各个关键词的集合  
对于每一个关键词得到其倒排表  
然后把所有的倒排表存入文件

## 对关键词的检索

- 第一步，在倒排文件中检索关键词
- 第二步，如果找到了关键词，那么获取文件中的对应的倒排表，并获取倒排表中的记录
- 通常使用另一个索引结构（字典）进一步对关键词表进行 有效索引
  - Trie
  - 散列



## 倒排文件优劣

- 高效检索，用于文本数据库系统
- 支持的检索类型有限
  - 检索词有限（只能用索引文件中的关键词）
  - 倒排文件中的索引效率可能不高
  - 需要的空间代价往往很高



## 思考

- 怎样有效地组织属性倒排索引表？
- 一个关键词如果在同一个文本中多次出现，它在倒排文件中的索引项是否能进行合并？



# 数据结构与算法

谢谢聆听

国家精品课“数据结构与算法”

<http://www.jpk.pku.edu.cn/pkujpk/course/sjjg/>

张铭，王腾蛟，赵海燕

高等教育出版社，2008.6。“十一五”国家级规划教材