

信息检索与数据挖掘

第5章 向量模型及检索系统
——第二讲 检索系统

课程内容

- 第1章 绪论
- 第2章 布尔检索及倒排索引
- 第3章 词典查找及扩展的倒排索引
- 第4章 索引构建和索引压缩
- 第5章 向量模型及检索系统
 - 向量模型
 - 检索系统的评分计算
- 第6章 检索的评价
- 第7章 相关反馈和查询扩展
- 第8章 概率模型
- 第9章 基于语言建模的检索模型
- 第10章 文本分类
- 第11章 文本聚类
- 第12章 Web搜索
- 第13章 多媒体信息检索
- 第14章 其他应用简介

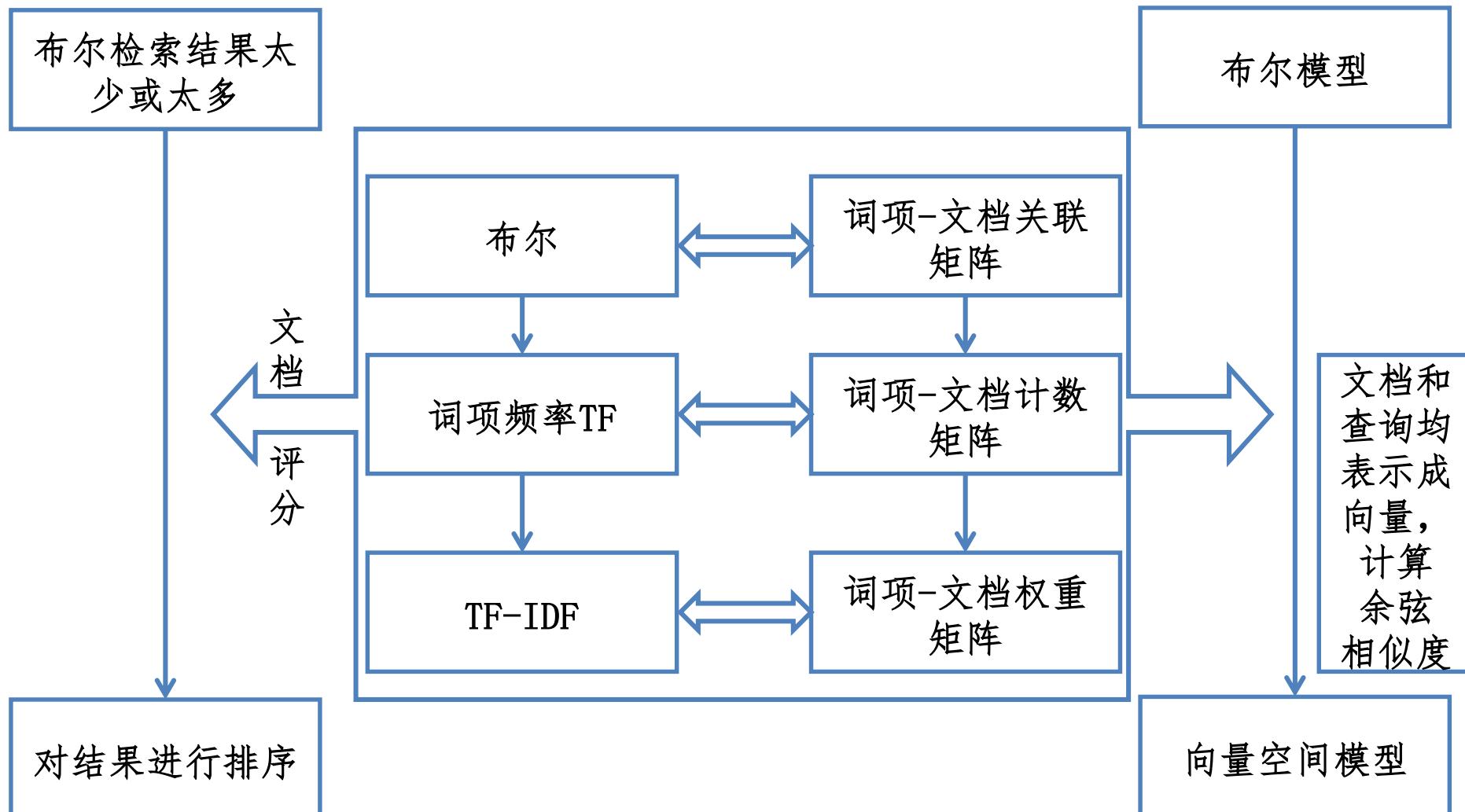
本讲提纲

- ① 上一讲回顾
- ② 结果排序的重要性
- ③ 结果排序的实现
- ④ 完整的搜索系统

提纲

- ① 上一讲回顾
- ② 结果排序的重要性
- ③ 结果排序的实现
- ④ 完整的搜索系统

回顾：从布尔模型到向量空间模型



回顾：词项频率tf

- t 在 d 中的对数词频权重定义如下：

$$w_{t,d} = \begin{cases} 1 + \log_{10} tf_{t,d} & \text{if } tf_{t,d} > 0 \\ 0 & \text{otherwise} \end{cases}$$

- 文档-词项的匹配得分 $\sum_{t \in q \cap d} (1 + \log tf_{t,d})$

回顾： idf权重

- df_t 是出现词项 t 的文档数目
- df_t 是和词项 t 的信息量成反比的一个值
- 于是可以定义词项 t 的 idf 权重：

$$\text{idf}_t = \log_{10} \frac{N}{\text{df}_t}$$

(其中 N 是文档集中文档的数目)

- idf_t 是反映词项 t 的信息量的一个指标

回顾：tf-idf权重

- tf-idf权重

$$w_{t,d} = (1 + \log tf_{t,d}) \times \log_{10}(N / df_t)$$

- tf-idf 是信息检索中最著名的权重计算方法
- tf-idf值随着词项在**单个文档中出现次数**增加而增大
- tf-idf值随着词项**在文档集中数目**减少而增加

某一特定文件内的高词语频率，以及该词语在整个文件集合中的低文件频率，可以产生出高权重的TF-IDF。因此，TF-IDF倾向于过滤掉常见的词语，保留重要的词语。

回顾：tf-idf 权重机制变形

Term frequency	Document frequency	Normalization
n (natural) $tf_{t,d}$	n (no) 1	n (none) 1
I (logarithm) $1 + \log(tf_{t,d})$	t (idf) $\log \frac{N}{df_t}$	c (cosine) $\frac{1}{\sqrt{w_1^2 + w_2^2 + \dots + w_M^2}}$
a (augmented) $0.5 + \frac{0.5 \times tf_{t,d}}{\max_t(tf_{t,d})}$	p (prob idf) $\max\{0, \log \frac{N - df_t}{df_t}\}$	u (pivoted unique) $1/u$
b (boolean) $\begin{cases} 1 & \text{if } tf_{t,d} > 0 \\ 0 & \text{otherwise} \end{cases}$		b (byte size) $1/CharLength^\alpha, \alpha < 1$
L (log ave) $\frac{1 + \log(tf_{t,d})}{1 + \log(\text{ave}_{t \in d}(tf_{t,d}))}$		

回顾：Queries表示成向量

- 关键思路1：对于查询做同样的处理，即将查询表示成同一高维空间的向量
- 关键思路2：在**向量空间**内根据**queries**与**文档向量**间的距离来排序

回顾：查询和文档之间的余弦相似度计算

$$\cos(\vec{q}, \vec{d}) = \frac{\vec{q} \bullet \vec{d}}{\|\vec{q}\| \|\vec{d}\|} = \frac{\vec{q}}{\|\vec{q}\|} \bullet \frac{\vec{d}}{\|\vec{d}\|} = \frac{\sum_{i=1}^{|V|} q_i d_i}{\sqrt{\sum_{i=1}^{|V|} q_i^2} \sqrt{\sum_{i=1}^{|V|} d_i^2}}$$

- q_i 是词项*i*在query中的tf-idf 权值
- d_i 是词项*i*在文档中的tf-idf 权值
- $\cos(\vec{q}, \vec{d})$ \vec{q} 与 \vec{d} 的余弦相关性
- 等价于向量 \vec{q} 与 \vec{d} 夹角的余弦值

从示例看TF、IDF

https://www.baidu.com/s?wd=%E4%B8%AD%f 中科大 新校长_百度搜索 中科大 新校长 - 必应

Baidu 百度 中科大 新校长 中科大 新校长 百度一下 添加百度到桌面，搜索更便捷！ 百度首页 消息 设置 ▾

网页 新闻 贴吧 知道 音乐 图片 视频 地图 文库 更多»

百度为您找到相关结果约962,000个

为您推荐：中科大下任校长 倪文龙 中科大校长 中科大新校长是谁

中科大 新校长的最新相关信息

万立骏任中科大校长 曾有二十余项发明专利(图)

 在就职讲话中，**新校长**万立骏不仅“敬佩各位先生从容执着，严谨治学，兢兢业业，教书育人的大师风范”，“赞颂**科大**学子为科学发展学习，自己不断成长，也把...
新浪新闻 1天前

中科大新校长万立骏:尽心尽力当好服务员 环球网
万立骏出任中科大新校长:不负组织所托和... 凤凰网
万立骏出任中科大校长 官网已更新 网易财经
万立骏任中科大校长 系中央候补委员(图) 凤凰网

1天前
2天前
2天前
2天前

万立骏出任中国科大学新校长 新闻 腾讯网
2天前 - 万立骏出任**中国科大**新校长 万立骏讲话 喻琼摄影 喻云林在讲话中表示，万立骏同志担任**中国科大**校长，是党中央、国务院从中管高领领导班子建设全局和中...
news.qq.com/a/20150327... - 百度快照 - 81%好评

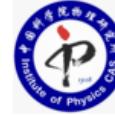
中科大新校长万立骏到任 称自己和科大早就结缘(图) 网易新闻中心
1天前 - 在就职讲话中，**新校长**万立骏不仅“敬佩各位先生从容执着，严谨治学，兢兢业业，教书育人的大师风范”，“赞颂**科大**学子为科学发展学习，自己不断成长，也把...
news.163.com/15/0328/0... - 百度快照 - 85%好评

中科大新校长万立骏走马上任 系中央候补委员-搜狐教育
 2天前 - 据人民网安徽频道消息，3月27日上午11时，**中国科学技术大学**召开全校教授干部大会。会上，中组部有关负责人宣布，万立骏任**中科大**

相关学者

 **万立骏** 物理化学家
 **潘建伟** 中科院院士
 **王恩哥** 科学院院士
 **舒红兵** 十二届全国政协委员

其他人还搜

 **方璐**
 **中国科学技术大学**
 **中国科学院化学研究所**
 **中国科学院物理研究所**
隶属于中国科学院
多学科综合性研究机构

相关人物

 **林建华**
 **宋永华**
 **侯建国**
 **朱清时**

从示例看TF、IDF

https://www.baidu.com/s?wd=%E4%B8%AD%f 中科大 万立骏_百度搜索 中科大的新校长 - 必应 中科大新校长 - 搜狗搜索 中科大新校长 - Google 搜索

中科大 万立骏 百度一下 添加百度到桌面，搜索更便捷！ 百度首页 消息 设置 fa...

网页 新闻 贴吧 知道 音乐 图片 视频 地图 文库 更多»

百度为您找到相关结果约8,060个

中科大 万立骏的最新相关信息

热点

 万立骏任中科大校长 系中央候补委员

万立骏执掌中科大

中科大召开全校教授干部大会，中组部有关负责人宣布，万立骏任中科大校长。[详情>](#)

来源：凤凰网 海归院士赤子 执着科研之路
万立骏简历 回国创业奉献

万立骏任中科大校长 曾有二十余项发明专利(图) 新浪新闻
中科大新校长万立骏 尽心尽力当好服务员 环球网
中科大校长万立骏：不负组织所托 不负师生厚望 中国日报
中科大新校长万立骏走马上任 系中央候补委员 搜狐教育频道
中科大新任校长万立骏 系海归院士的杰出代表 网易新闻

1天前 1天前 2天前 2天前 2天前

中科大 万立骏的最新微博结果

 新安晚报 V : 【万立骏任中国科技大学校长 系中央候补委员】3月27日上午11时，中国科学技术大学召开全校教授干部大会，会上，中组部有关负责人宣布，**万立骏**任**中科大**校长。简历显示，2012年11月14日，**万立骏**当选中国共产党第十八届中央委员会候补委员(人民网) <http://t.cn/RAyHHAq> 查看全文>
2天前 - 新浪微博 转发(10) | 评论(2)

相关人物 展开 ▼

 曹淑敏
多少人羡慕不已的骄子

 袁帅
清华大学教师

 刘焕香
现任兰州大学教授

 翁冰莹
厦门大学人文学院博士

 肖益鸿
福大化学化工学院就职

 江雷
中国著名纳米材料专家

 白春礼
化学和纳米科技

 钱伟长
中国科学院院士

其他人物还搜 展开 ▼

 方璐

 侯建国

 徐良杰

 王吉红

中国科学技术大学校长 博士生导师 科学家

从示例看TF、IDF

Query给定的关键词顺序不同，结果不同？

Baidu 百度 万立骏 中科大 百度一下 添加百度到桌面，搜索更便捷！ 百度首页 消息 设置 ▾

网页 新闻 贴吧 知道 音乐 图片 视频 地图 文库 更多»

百度为您找到相关结果约35,500个

万立骏任中科大校长

热点



万立骏任中科大校长 系中央候补委员

万立骏执掌中科大

中科大召开全校教授干部大会，中组部有关负责人宣布，万立骏任中科大校长。[详情>](#)

来源：凤凰网
[海归院士赤子 执着科研之路](#)
[万立骏简历](#) [回国创业奉献](#)

中科大新校长万立骏走马上任 系中央候补委员-搜狐新闻



2天前 - 人民网合肥3月27日电(常国水韩震震)3月27日上午11时，[中国科学技术大学](#)召开全校教授干部大会，会上，中组部有关负责人宣布，[万立骏](#)任[中科大](#)校长。万...

news.sohu.com/20150327... [V3 - 百度快照](#)

万立骏 中科大的最新微博结果

 新安晚报 [V](#) : 【[万立骏任中国科技大学校长 系中央候补委员](#)】3月27日上午11时，[中国科学技术大学](#)召开全校教授干部大会，会上，中组部有关负责人宣布，[万立骏](#)任[中科大](#)校长。简历显示，2012年11月14日，[万立骏](#)当选中国共产党第十八届中央委员会候补委员(人民网) <http://t.cn/RAyHHAq> [查看全文>>](#)

2天前 - 新浪微博 [转发\(10\)](#) | [评论\(2\)](#)

Quills_Simon : <http://t.cn/zQMm97J> 万立骏[中科大](#)校长&src=http%3A%2F%2Fwww.ah.xinhuanet.com%2F2015-03%2F29%2Fc_1114796682.htm 科大新掌门 [查看全文>>](#)

相关人物 展开 ▾

 钱旭红 973首席科学家	 袁帅 清华大学教师	 曹淑敏 多少人羡慕不已的骄子	 侯建国 中国科学技术大学校长
 王吉红 科学家	 丁仲礼 第十届全国政协委员	 白春礼 化学和纳米科技	 朱清时 物理学家和自然科学家
 方璐 浙江副校长	 吴朝晖 中科院院士	 潘建伟 中科大教授	 王恩哥 科学院院士 前北大校长

其他人还搜 展开 ▾

 方璐 浙江副校长	 吴朝晖 中科院院士	 潘建伟 中科大教授	 王恩哥 科学院院士 前北大校长
--	---	---	--

从示例看TF、IDF

中文的自动分词

https://www.baidu.com/s?wd=%E4%B8%AD%F 百度一下 添加百度到桌面，搜索更便捷！ 百度首页 消息 设置 ▾

中科大新校长

网页 新闻 贴吧 知道 音乐 图片 视频 地图 文库 更多»

百度为您找到相关结果约1,180,000个

中科大新校长的最新相关信息

热点

万立骏任中科大校长 系中央候补委员

万立骏执掌中科大

中科大召开全校教授干部大会，中组部有关负责人宣布，万立骏任中科大校长。[详情>](#)

来源：凤凰网 [海归院士赤子 执着科研之路](#) [万立骏简历](#) [回国创业奉献](#)

万立骏任中科大校长 曾有二十余项发明专利(图) 新浪新闻
中国科大新掌门万立骏“当好大家的服务员” 凤凰网
中科大新校长万立骏:尽心尽力当好服务员 环球网
万立骏出任中科大新校长:不负组织所托和师生厚望 凤凰网
万立骏出任中科大校长 官网已更新 网易财经

1天前 1天前 1天前 2天前 2天前

中科大新校长的最新微博结果

江苏教育黄页 V : 【十所高校密集换帅 新校长多为他处调任】今天上午中组部宣布万立骏任**中科大**校长;而就在昨天,中组部宣布中央部门所属三所高校校长任免。上午,清华大学党委常务副书记、副校长邱勇院士接替陈吉宁,担任清华校长(副部长级);北京航空航天大学常务副校长徐惠彬院士接替怀进鹏,担任北航校长(副部长级) [查看全文>>](#)

23小时前 - 新浪微博

转发(0) | 评论(0)

Quills_Simon : <http://t.cn/zQMm97J> 万立骏**中科大校长**&src=http%3A%2F%

相关人物 展开 ▾

袁帅 清华大学教师
潘建伟 中科院院士
陈旭 清华大学党委书记
周涛 中国最年轻教授

冯祖 中国科学院博士
侯建国 中国科学技术大学校长
陈吉宁 环保部部长
万立骏 物理化学家

其他人还搜 展开 ▾

北大校花 方璐 白春礼 王恩哥

没有最美只有更美

从示例看TF、IDF

停用词权重为0
“的”？

Baidu 百度 中科大的新校长 百度一下 添加百度到桌面，搜索更便捷！ 百度首页 消息 设置 fa...7

网页 新闻 贴吧 知道 音乐 图片 视频 地图 文库 更多»

百度为您找到相关结果约911,000个

中科大的新校长的最新相关信息

万立骏任中科大校长曾有二十余项发明专利(图)

 在就职讲话中,新校长万立骏不仅“敬佩各位先生从容执着,严谨治学,兢兢业业,教书育人的大师风范”,“赞颂科大学子为科学发展学习,自己不断成长,也把...
新浪新闻 1天前

中科大新校长万立骏尽心尽力当好服务员 环球网 1天前

万立骏出任中科大新校长不负组织所托和... 凤凰网 2天前

万立骏出任中科大校长官网已更新 网易财经 2天前

中科大新校长万立骏走马上任系中央候补... 搜狐教育频道 2天前

中科大新校长万立骏走马上任|校长|大学_凤凰资讯

1天前 - 信息时报讯 昨日上午11时,中国科学技术大学召开全校教授干部大会。会上,中组部有关负责人宣布,万立骏任中科大校长。人民 news.ifeng.com/a/20150... ▼ 3 - 百度快照 - 75%好评

中科大的新校长的最新微博结果

 安徽身边事 V : 【中科大新校长万立骏到任】昨日11时,中国科学技术大学在东区理化大楼西三报告厅召开全校教授干部大会,宣布万立骏担任中国科学技术大学校长的决定。至此,中国五所著名大学的校长全部完成换届。<http://t.cn/RAUX715> 校长颜值蛮高的嘛[可爱]  查看全文>>

1天前 - 新浪微博 转发(4) | 评论(1)

Quills_Simon : <http://t.cn/zQMm97J> 万立骏 **中科大校长**&src=http%3A%2F%2Fwww.ah.xinhuanet.com%2F2015-03%2F29%2Fc_1114796682.htm 科大新掌门
查看全文>>

相关人物

			
袁帅 清华大学教师	潘建伟 中科院院士 中科大教授	陈旭 清华大学党委书记	周涛 中国最年轻教授
			
冯树 中国科学院博士	侯建国 中国科学技术大学校长	陈吉宁 环保部部长	万立骏 物理化学家

其他人还搜

			
北大校花 没有最美只有更美	方璐	白春礼	王恩哥 科学院院士 前北大校长

提纲

- ① 上一讲回顾
- ② 结果排序的重要性
- ③ 结果排序的实现
- ④ 完整的搜索系统

排序的重要性

http://cn.bing.com/search?q=%E4%B8%AD%E7%
中科院 新校长_百度搜索
中科院 新校长 - 必应

必应 中科大 新校长

网页 图片 视频 网典 资讯 地图 词典 更多

Bing.com in English 登录

中国科学技术大学

中国科技大学，即中国科学技术大学。中国科学院直属的一所以前沿科学和高新技术为主、兼有以科技为背景的管理和人文学科的综合性全国重点大学。
ustc.edu.cn ▾ 2015-3-26

中国科学技术大学 百度百科

中国科学技术大学是中国科学院直属的一所以前沿科学和高新技术为主、兼有特色管理和人文学科的理工类全国重点大学，是国家首批“211工程”、“985工程”重点建设院校，入…
[历史沿革](#) · [学术研究](#) · [办学规模](#) · [学校领导](#) · [知名校友](#) · [校区环境](#) · [文化传统](#)

[维基百科](#) [互动百科](#)
baike.baidu.com/view/4522 2015-1-16

新校长透露科大未来规划 - 中国科学技术大学新创校友基金会

新校长透露科大未来规划 侯建国高相校友龙门阵 校友龙门阵第二期相关新闻：“建国你好”延续民主办学氛围、侯建国的龙门阵、两会代表扎堆校友龙门阵
www.ustcif.org/default.php/content/523 ▾

中国科学技术大学的历任校长： 姚萌 新浪博客

上次更新日期: 2012-10-30 · 类别: 岁月回眸 · 博客等级: 16
中国科学技术大学的历任校长: 朱清时 福建平潭 中国科学院院士 中国科技大学 激光晶体
中国科学技术大学的历任校长: 郭沫若 1958.9-1978.6 严济慈 1980.2-1984.9 管…
blog.sina.com.cn/s/blog_635131bf010159mj.html ▾ 2012-10-30

侯建国获评中科大最好校长 上任就为学生装空调 建国 中科大…

维清华校长陈吉宁调任环保部长后，据共产党员网1月28日下午消息，中科院校长侯建国任科技部副部长。中国科大新创校友基金会微信公众号早在1月25日就…
news.atv.com.cn/system/2015/01/29/012109524.shtml ▾ 2015-1-29

微软专题活动

必应 全球生活汇 GO>
秒速5厘米的异国浪漫，樱花——你邂逅了么？

当前热点

- 习近平比尔盖茨扶贫
- 多国政要出席李光耀国葬
- 中国启动也门撤侨
- NASA耗资77亿元摘星
- 德翼副驾驶选择坠机地
- 澳洲砖工月薪3万人民币
- 备降昆明航班闻到糊味
- 汪涵成歌手3最大赢家

相关搜索

中科院 第一副校长

排序的重要性

http://cn.bing.com/search?q=%E4%B8%AD%E7%
中科院新校长_百度搜索
中科院新校长 - 必应
中科院新校长 - 搜狗搜索
中科院新校长 - Google 搜索

必应 中科大新校长

网页 图片 视频 网典 资讯 地图 词典 更多

Bing.com in English 登录

中科院新校长万立骏：尽心尽力当好服务员-搜狐新闻

中新社合肥3月27日电题：中科院新校长万立骏：尽心尽力当好服务员中新社记者吴兰“空档期”两个月的中科院，迎来了新一任校长。27日，中共中央组织部相关...

news.sohu.com 国内要闻 时事 2 天前

中科院新校长的最新相关资讯

中科院新校长万立骏到任 称自己和科大早就结缘 六安新闻网 · 1 天前
昨日11时，中国科学技术大学在东区理化大楼西三报告厅召开全校教授干部大会，宣布万立骏担任中国科学技术大学校长的决定。至此，中国五所著名大学的校长全部完成换届，巧合的是，万立骏与此前上任的北京大学新校长林建华、清华大学新校长...

万立骏出任中科院新校长：不负组织所托和师生厚望 东方网新闻 · 2 天前
中科院新校长万立骏：尽心尽力当好服务员 搜狐新闻 · 2 天前
bing.com/news

资讯继续看：焦点 国内 国际 财经 体育 娱乐 科技 教育 汽车 军事

中科院新任校长万立骏，系海归院士的杰出代表 网易新闻中心

中科院新任校长万立骏，系海归院士的杰出代表，中科院院士 杰出 应用 网易新闻 网易云音乐 网易云阅读 有道云笔记 网易花田 网易公开课 网易彩票 有道词典 ...
news.163.com 网易首页，新闻中心 2 天前

中科院新校长万立骏到任 称自己和科大早就结缘--安徽频道 ...

中科院新校长万立骏到任 称自己和科大早就结缘 2015年03月28日09:33 来源：中安在线-安徽商报 手机看新闻 打印网摘 纠错 商城 分享 推荐
ah.people.com.cn/n/2015/0328/c358428-24306435-2.html 1 天前

微软专题活动

GO >
秒速5厘米的异国浪漫，
樱花——你邂逅了么？

当前热点

- 习近平比尔盖茨扶贫
- 多国政要出席李光耀国葬
- 中国启动也门撤侨
- NASA耗资77亿元摘星
- 德翼副驾驶选择坠机地
- 澳洲砖工月薪3万人民币
- 备降昆明航班闻到糊味
- 汪涵成歌手3最大赢家

相关搜索

中科院新校长

排序的重要性

http://www.sogou.com/web?query=%E4%B8%AD

中科院 新校长_百度搜索 中科大 新校长 - 必应 中科大 新校长 - 搜狗搜索

搜狗搜索 新闻 网页 微信 问问 图片 视频 音乐 地图 论坛 更多>

中科院 新校长 搜狗搜索 全部时间 ▾

中科院 新校长的最新相关信息

万立骏任职中科院院长 环球网 1天前
2015年3月28日 - 为将中国科学技术大学早日建成世界一流研究型大学而努力奋斗。校友谈期待万立骏任院长的信息发布后，中科院微信号刊登了一些从中科院走出的校友对新院长...
【中国新闻网】中科院新院长万立骏：尽心尽力当... 中国科学技术大学 31分钟前
中科院新院长万立骏：尽心尽力当好服务员 搜狐新闻 2天前

中科院新院长万立骏到任 称自己和科大早就结缘-新闻频道-和讯网

2015年3月28日 - 中国三所顶尖高校院长也组成了“化学三掌门”。新院长与科大早就结缘 “... 在会上，万立骏发表了热情洋溢的就职讲话。在他看来，中科院是一所“具有...
和讯财经新闻 - news.hexun.com - 1天前 - 快照 - 预览

中科院新院长万立骏走马上任 - 滚动热点 - 21CN.COM

2015年3月28日 - 信息时报讯 昨日上午11时，中国科学技术大学召开全校教授干部大会。会上，中组部有关负责人宣布，万立骏任中科院院长。人民万立骏简历 万立骏，1957年7...
21CN - news.21cn.com/caiji... - 1天前 - 快照 - 预览

中科院新院长万立骏走马上任_网易新闻中心

2015年3月28日 - 信息时报讯 昨日上午11时，中国科学技术大学召开全校教授干部大会。会上，中组部有关负责人宣布，万立骏任中科院院长。人民万立骏简历 万立骏，1957年7...
网易新闻 - news.163.com/15/032... - 1天前 - 快照 - 预览

中科院 新校长的相关微信公众号文章

中科院新院长万立骏上任 系中央候补委员
3月27日上午11时，中国科学技术大学召开全校教授干部大会，会上，中组部有关负责人宣布，万立骏任中科院院长。(常国水 韩震震)万立骏简历:万

985高校 展开 ▾

中国科技大学 全国重点大学
兰州大学
华中科技大学 985211重点大学
浙江大学 中国著名顶尖学府之一

大连理工大学 卓越大学联盟
同济大学 中国著名高等学府
东北大学 东北大学秦皇岛分校
南京大学 顶尖大学九校联盟

大学校长 展开 ▾

侯建国
丁烈云
汪晋宽
李培根

科学家 展开 ▾

30

排序的重要性

Sogou search results for "中科大新校长" (Chinese University of Science and Technology new president) on March 28, 2015.

Search Results:

- 中科大新校长的最新相关信息**
 - 中科大新校长万立骏到任 称自己和科大早就结缘**
和讯财经新闻 1天前
2015年3月28日 - 从1958年初至今, 57年中, 中国科学技术大学共迎来了9位校长。据中科大新创校友基金会总结, 万立骏是约22年来第一次“空降”的中国科大...
 - 万立骏任职中科大校长 环球网**
1天前
 - 【中国新闻网】中科大新校长万立骏：尽心尽力当... 中国科学技术大学**
39分钟前
 - 中科大新校长万立骏：尽心尽力当好服务员 搜狐新闻**
2天前
 - 中科大新校长万立骏走马上任 系中央候补委员 搜狐新闻**
2天前
- 中科大新校长万立骏到任 称自己和科大早就结缘 (图) 网易新闻中心**
 - [图文] 2015年3月28日 - 从1958年初至今, 中国科学技术大学共迎来了9位校长。据中科大新创校友基金会总结, 万立骏是约22年来第一次“空降”的中国科大校长, 上一次...**
网易新闻 - news.163.com - 1天前 - 快照 - 预览
- 中科大新校长万立骏走马上任 - 滚动热点 - 21CN.COM**
 - 2015年3月28日 - 信息时报讯 昨日上午11时, 中国科学技术大学召开全校教授干部大会。会上, 中组部有关负责人宣布, 万立骏任中科大校长。人民 万立骏简历 万立骏, 1957年7...**
21CN - news.21cn.com/caiji... - 1天前 - 快照 - 预览
- 中科大新校长万立骏走马上任-新闻频道-和讯网**
 - 2015年3月28日 - 信息时报讯 昨日上午11时, 中国科学技术大学召开全校教授干部大会。会上, 中组部有关负责人宣布, 万立骏任中科大校长。人民 万立骏简历 万立骏, 1957年7...**
和讯财经新闻 - news.hexun.com/2015... - 1天前 - 快照 - 预览
- 中科大新校长的相关微信公众号文章**

885高校

	中国科学技术大学 全国重点大学		兰州大学		华中科技大学 985211重点大学		浙江大学 中国著名顶尖学府之一
	大连理工大学		同济大学 中国著名高等学府		东北大学 东北大学秦皇岛分校		南京大学 顶尖大学九校联盟

大学校长

	侯建国		丁烈云		汪晋宽		李培根
--	------------	--	------------	--	------------	--	------------

科学家

	薛其坤		屠呦呦		袁隆平		杨振宁
--	------------	--	------------	--	------------	--	------------

排序的重要性

The screenshot shows a search results page from Google. The search query is "中科大 新校长". The results are filtered by "网页" (Web). There are approximately 435,000 results found in 0.43 seconds.

中科大新校长万立骏：尽心尽力当好服务员--教育--人民网
edu.people.com.cn > 教育 ▾
2 天前 - 中新社合肥3月27日电题：**中科大新校长**万立骏：尽心尽力当好服务员中新社记者吴兰“空档期”两个月的**中科大**，迎来了新一任校长。27日，中共中央 ...

中科大新校长万立骏走马上任 系中央候补委员（图/简历）--教...
edu.people.com.cn > 教育 ▾
2 天前 - 人民网合肥3月27日电（常国水韩震震）3月27日上午11时，**中国科学技术大学**召开全校教授干部大会，会上，中组部有关负责人宣布，万立骏任**中科大** ...

时事新闻

 **中科大新校长万立骏走马上任**
搜狐 - 2 天前
此次万立骏履新**中科大校长**也成为近2日内上任的第4位高校**校长**。此前，邱勇已出任清华大学 ...

中科大新校长万立骏：尽心尽力当好服务员
新浪网 - 2 天前

中科大新校长万立骏到任称自己和科大早就结缘
中安在线 - 2 天前

更多关于“中科大 新校长”的新闻

排序的重要性

+你 搜索 图片 地图 Play YouTube 新闻 Gmail 更多 ▾

中科大新校长

网页 新闻 图片 视频 更多 ▾ 搜索工具

找到约 866,000 条结果 (用时 0.35 秒)

中科大新校长万立骏：尽心尽力当好服务员--教育--人民网
edu.people.com.cn, 教育 ▾
2 天前 - 中新社合肥3月27日电题：中科大新校长万立骏：尽心尽力当好服务员中新社记者吴兰“空档期”两个月的中科大，迎来了新一任校长。27日，中共中央 ...

中科大新校长万立骏走马上任系中央候补委员-搜狐教育
learning.sohu.com, 新闻, 高等教育 ▾
2 天前 - 据人民网安徽频道消息，3月27日上午11时，中国科学技术大学召开全校教授干部大会，会上，中组部有关负责人宣布，万立骏任中科大校长。

中科大新校长万立骏到任称自己和科大早就结缘-安徽新闻 ...
ah.anhuinews.com, 科教教育, 科教新闻 ▾
2 天前 - 昨日11时，中国科学技术大学在东区理化大楼西三报告厅召开全校教授干部大会，宣布万立骏担任中国科学技术大学校长的决定。至此，中国五所 ...

中科大新校长万立骏上任系中央候补委员(图/简历) -中新网
www.chinanews.com/edu/2015/03-27/7164350.shtml ▾
2 天前 - 3月27日上午11时，中国科学技术大学召开全校教授干部大会，会上，中组部有关负责人宣布，万立骏任中科大校长。

中国科大换帅！侯建国进京履新科技部- 中国科学技术大学新 ...
www.ustcif.org/default.php/content/1971/ ▾
2015年1月25日 - 中国特殊的体制决定大学校长可能悬空——2013年2月，浙大校长杨卫调任自然科学基金委主任；6月26日新校长林建华方上任；校友一度曾致信建议

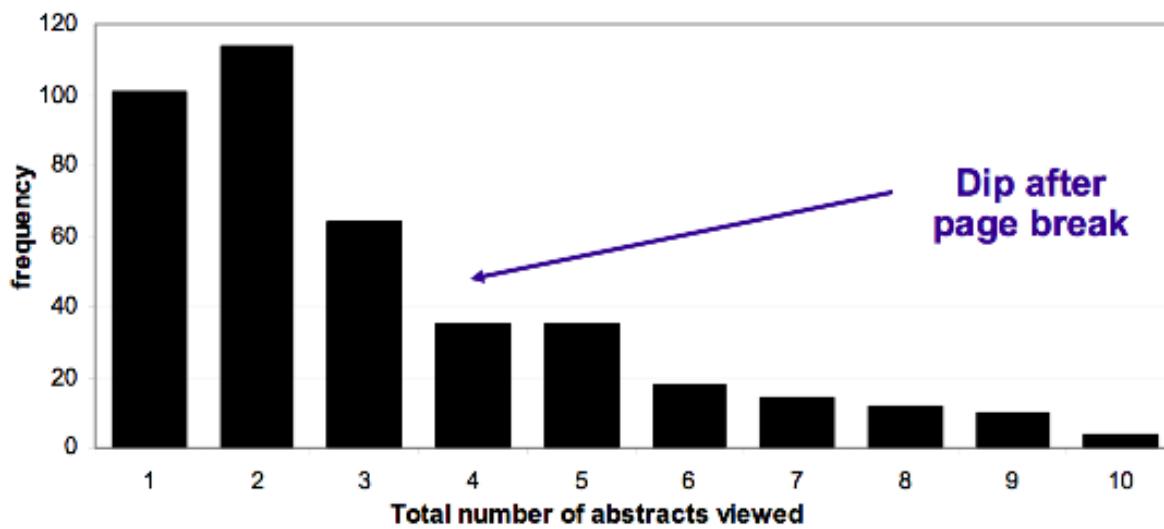
排序的重要性

- 上一讲：**不排序的问题严重性**
 - 用户只希望看到一些而不是成千上万的结果
 - 很难构造只产生一些结果的查询
 - 即使是专家也很难
 - → 排序能够将成千上万条结果缩减至几条结果，因此非常重要
- 实际上，大部分用户只看1到3条结果

用户浏览的链接数

How many links do users view?

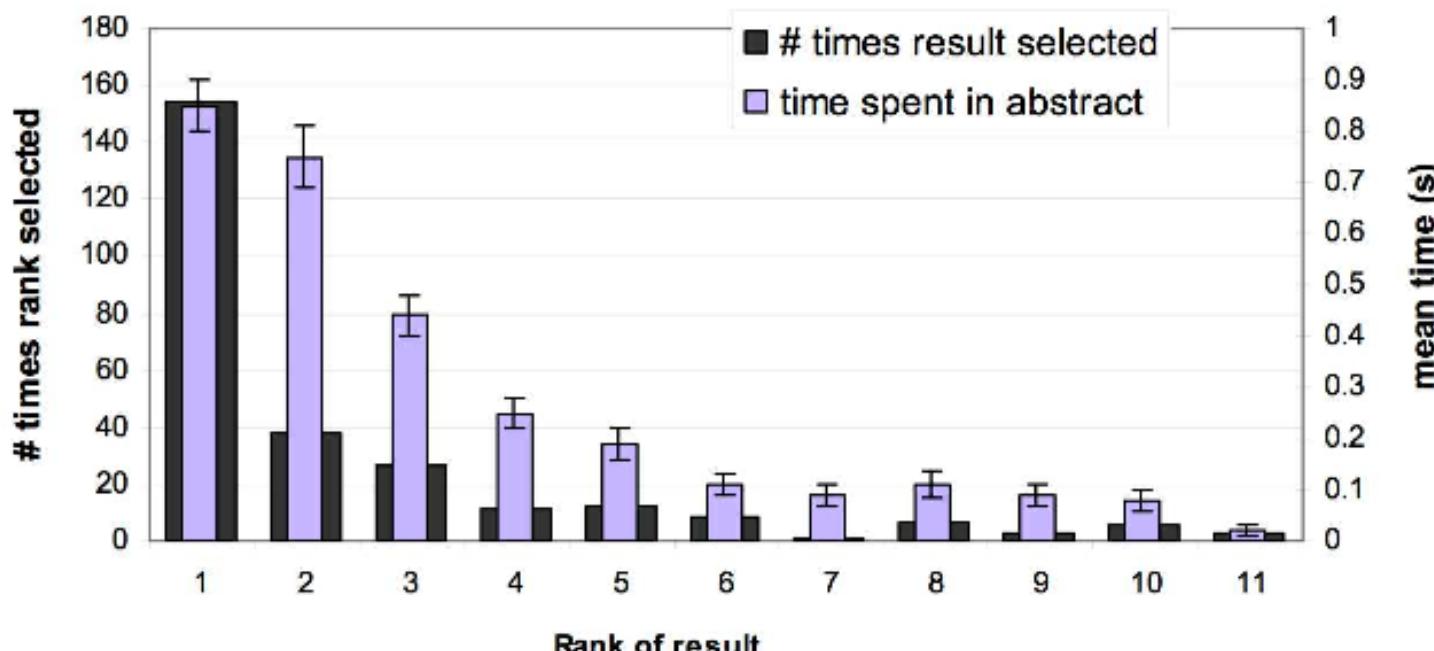
Total number of abstracts viewed per page



Mean: 3.07 Median/Mode: 2.00

浏览 vs. 点击

Looking vs. Clicking

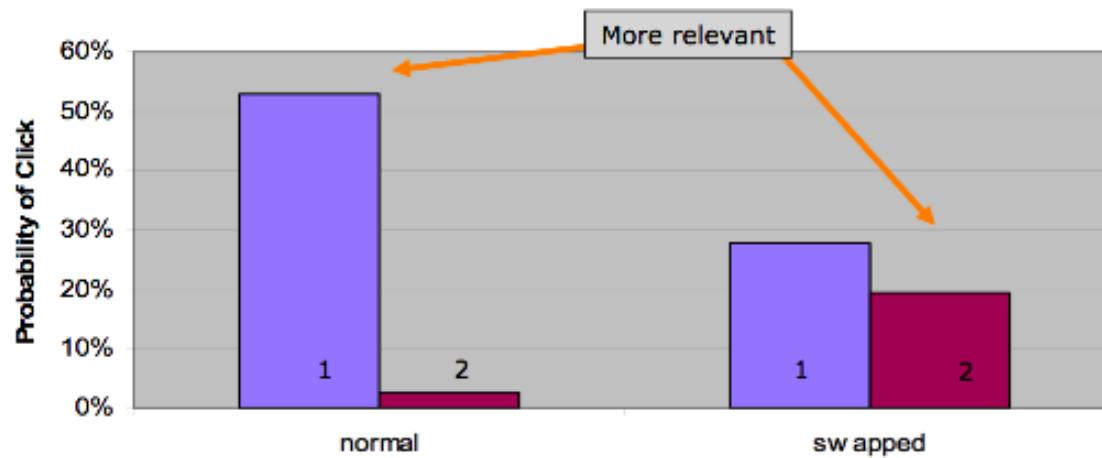


- Users view results one and two more often / thoroughly
- Users click most frequently on result one

结果显示顺序对行为的影响

Presentation bias – reversed results

- Order of presentation influences where users look
AND where they click



排序的重要性：小结

- 摘要阅读(Viewing abstracts)：用户更可能阅读前几页(1, 2, 3, 4)的结果的摘要
- 点击(Clicking)：点击的分布甚至更有偏向性
 - 一半情况下，用户点击排名最高的页面
 - 即使排名最高的页面不相关，仍然有30%的用户会点击它
- → 正确排序相当重要
- → 把最相关的页面放在首页非常重要

提纲

- ① 上一讲回顾
- ② 结果排序的重要性
- ③ 结果排序的实现

精确top K 检索及其加速办法

非精确top K检索

- ④ 完整的搜索系统

精确top K 检索及其加速办法

- 方法一：快速计算余弦
- 方法二：堆排序法N中选K
- 方法三：提前终止计算

词项频率tf和文档频率idf的存储

- 词典中保存每个词的idf值
- 词项频率tf存入倒排索引中

BRUTUS →

1 ,2	7 ,3	83 ,1	87 ,2	...
------	------	-------	-------	-----

CAESAR →

1 ,1	5 ,1	13 ,1	17 ,1	...
------	------	-------	-------	-----

CALPURNIA →

7 ,1	8 ,2	40 ,1	97 ,3
------	------	-------	-------

- 当然也需要位置信息，上面没显示出来

倒排索引中的词项频率存储

- 每条倒排记录中，除了 docID_d 还要存储 $\text{tf}_{t,d}$
- 通常存储是原始的整数词频，而不是对数词频对应的实数值
- 这是因为取对数后的实数值不易压缩
- 对 tf 采用一元码编码效率很高 ←为什么？
- 总体而言，额外存储 tf 所需要的开销不是很大：采用位编码压缩方式，每条倒排记录增加不到一个字节的存储量
- 或者在可变字节码方式下每条倒排记录额外需要一个字节即可

余弦相似度计算算法

COSINESCORE(q)

- 1 $\text{float Scores}[N] = 0$
- 2 $\text{float Length}[N]$
- 3 **for each** query term t
- 4 **do** calculate $w_{t,q}$ and fetch postings list for t
- 5 **for each** pair(d , $\text{tf}_{t,d}$) in postings list
- 6 **do** $\text{Scores}[d] += w_{t,d} \times w_{t,q}$
- 7 Read the array Length
- 8 **for each** d
- 9 **do** $\text{Scores}[d] = \text{Scores}[d] / \text{Length}[d]$
- 10 **return** Top K components of $\text{Scores}[]$

能不能加快？

精确top K 检索及其加速办法

- 目标：从文档集的所有文档中找出 K 个离查询最近的文档
- (一般)步骤：对每个文档评分(余弦相似度)，按照评分高低排序，选出前 K 个结果
- 如何加速：
 - 思路一：加快每个余弦相似度的计算
 - 思路二：不对所有文档的评分结果排序而直接选出Top K 篇文档
 - 思路三：能否不需要计算所有 N 篇文档的得分？

精确top K检索加速方法一：快速计算余弦

- 检索排序就是找查询的K近邻
- 一般而言，在高维空间下，计算余弦相似度没有很高效的方法
- 但是如果查询很短，是有一定办法加速计算的，而且普通的索引能够支持这种快速计算

特例 - 不考虑查询词项的权重

- 查询词项无权重
 - 相当于假设每个查询词项都出现1次
- 于是，不需要对查询向量进行归一化
 - 可以对上一讲给出的余弦相似度计算算法进行轻微的简化

$$\cos(\vec{q}, \vec{d}) = \frac{\vec{q} \bullet \vec{d}}{\|\vec{q}\| \|\vec{d}\|} = \frac{\vec{q}}{\|\vec{q}\|} \bullet \frac{\vec{d}}{\|\vec{d}\|} = \frac{\sum_{i=1}^{|V|} q_i d_i}{\sqrt{\sum_{i=1}^{|V|} q_i^2} \sqrt{\sum_{i=1}^{|V|} d_i^2}}$$

快速余弦相似度计算：无权重查询

FASTCOSINESCORE(q)

- 1 float $Scores[N] = 0$
- 2 **for each** d
- 3 **do** Initialize $Length[d]$ to the length of doc d
- 4 **for each** query term t
- 5 **do** calculate $w_{t,q}$ and fetch postings list for t
- 6 **for each** pair(d , $tf_{t,d}$) in postings list
- 7 **do** add $wf_{t,d}$ to $Scores[d]$
- 8 Read the array $Length[d]$
- 9 **for each** d
- 10 **do** Divide $Scores[d]$ by $Length[d]$
- 11 **return** Top K components of $Scores[]$

Figure 7.1 A faster algorithm for vector space scores.

精确top k检索加速方法二：堆排序法N中选K

- 检索时，通常只需要返回前K条结果
 - 可以对所有的文档评分后排序，选出前K个结果，但是这个排序过程可以避免
- 令 $J = \text{具有非零余弦相似度值的文档数目}$
 - 从J中选K个最大的

堆排序法

- 堆排序法

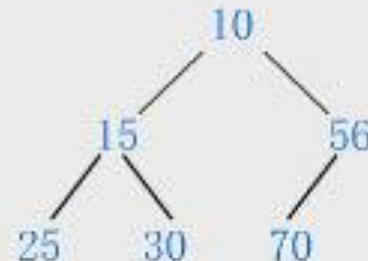
- 1991年的计算机先驱奖获得者、斯坦福大学计算机科学系教授罗伯特·弗洛伊德(Robert W. Floyd) 和威廉姆斯(J. Williams) 1964年共同发明了堆排序算法 (Heap Sort)

- 小根堆，大根堆

- n 个关键字序列 K_1, K_2, \dots, K_n 称为 (Heap)，当且仅当该序列满足如下性质 (简称为堆性质)： $k_i \leq k(2i)$ 且 $k_i \leq k(2i+1)$ ($1 \leq i \leq n/2$)，这是小根堆，大根堆则换成 \geq 号。
- 大根堆和小根堆：根结点 (亦称为堆顶) 的关键字是堆里所有结点关键字中最小者的堆称为小根堆，又称最小堆。根结点 (亦称为堆顶) 的关键字是堆里所有结点关键字中最大者，称为大根堆，又称最大堆。

堆排序法: 小根堆, 大根堆

堆可以被看成是一棵树，结点在堆中的**高度**可以被定义为从本结点到叶子结点的最长简单下降路径上边的数目；
定义堆的高度为树根的高度。

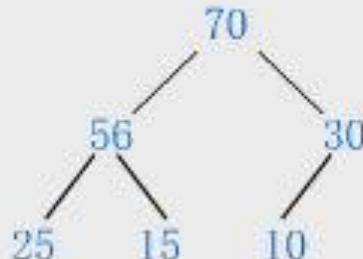


(a) 逻辑结构



(b) 存储结构

小根堆示例



(a) 逻辑结构



(b) 存储结构

Baidu Baidu

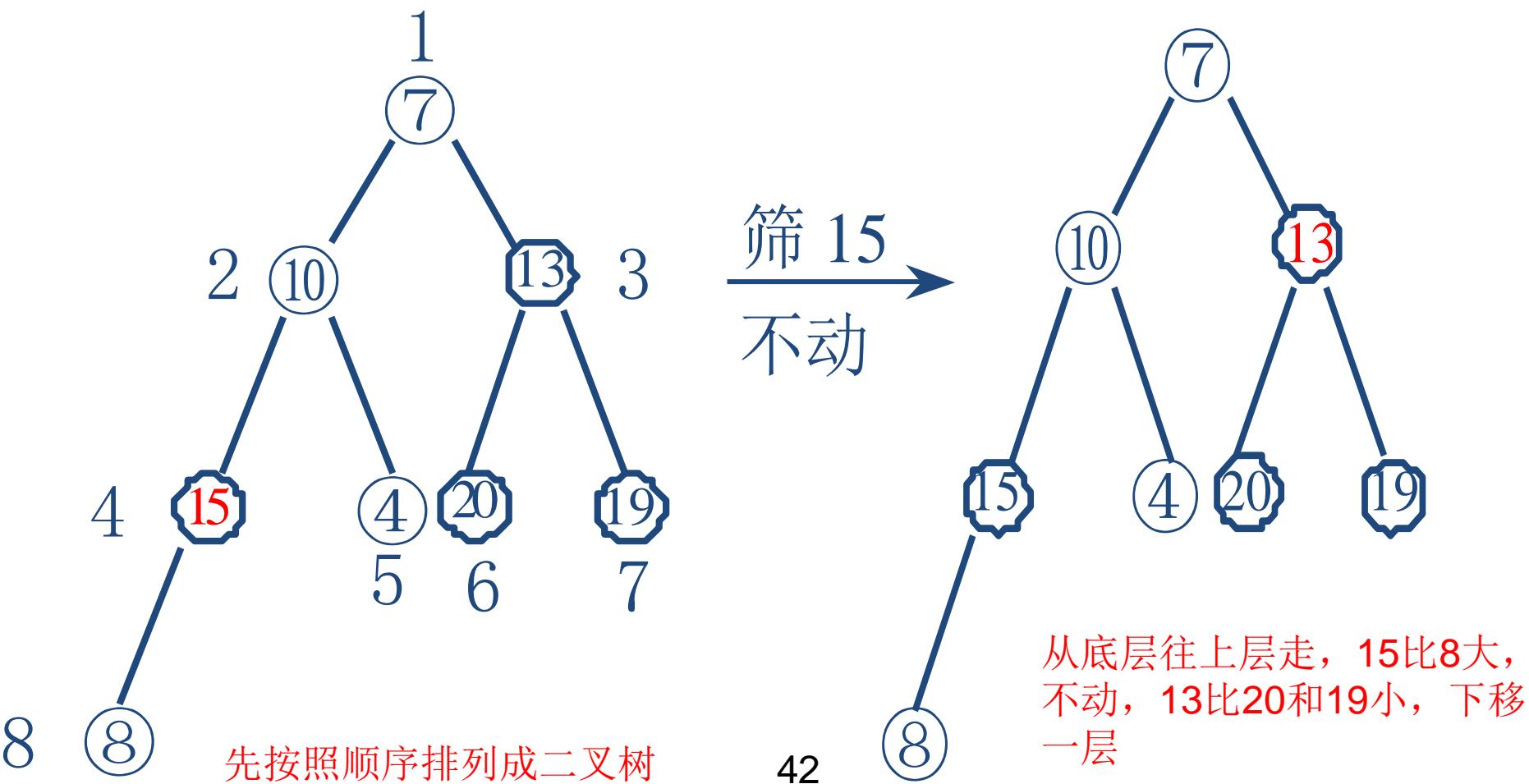
大根堆示例

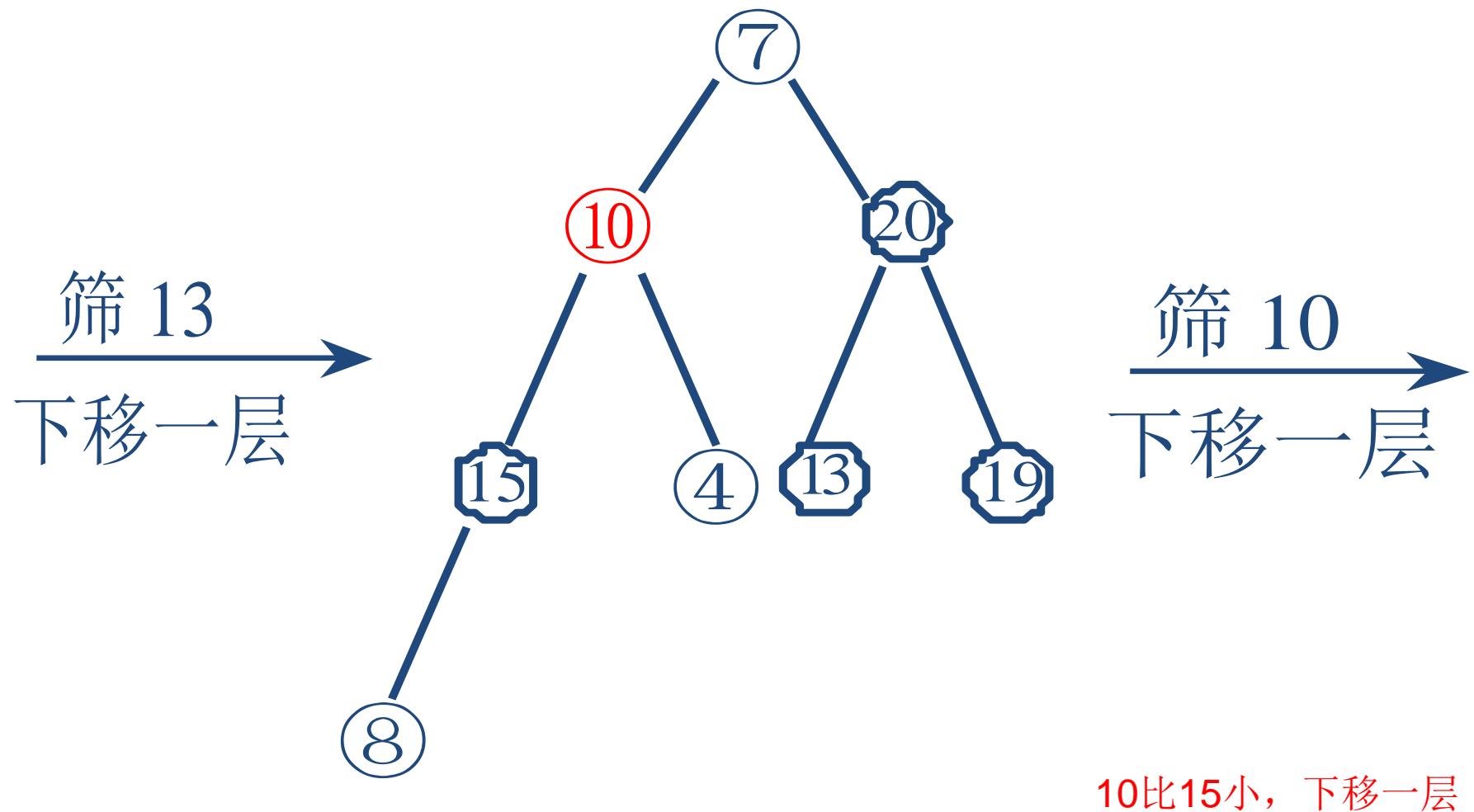
堆方法: 运算量

- 堆: 二叉树的一种, 每个节点上的值 $>$ 子节点上的值 (Max Heap)
- 步骤:
 - 堆构建: 需要 $2J$ 次操作
 - 选出前 K 个结果: 每个结果需要 $2\log J$ 步
- 如果 $J=1M$, $K=100$, 那么代价大概是全部排序代价的 10%

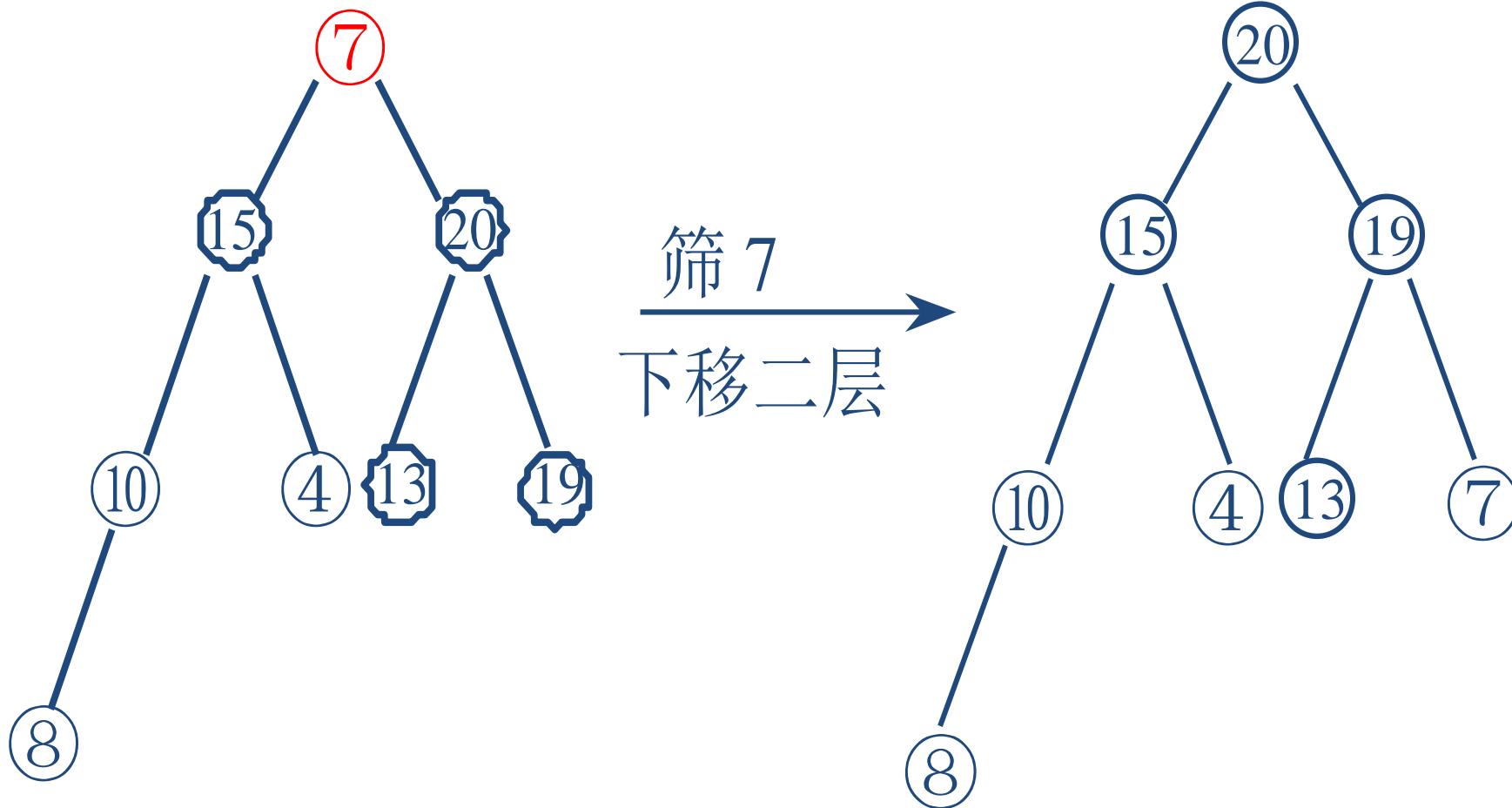
(最大)堆构建样例(筛选shift法-摘自网上课件)

- 7, 10, 13, 15, 4, 20, 19, 8 (数据个数n=8)

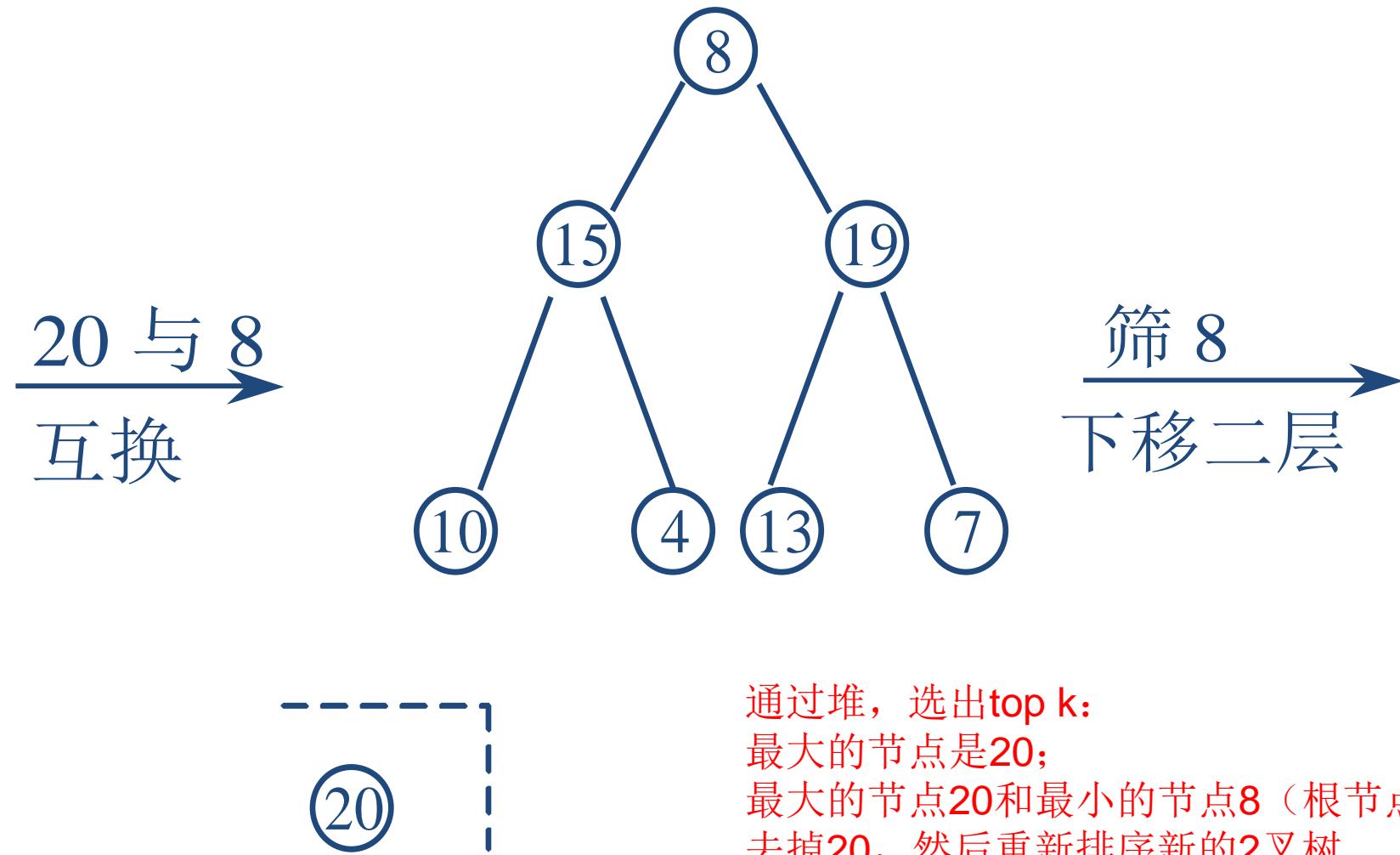




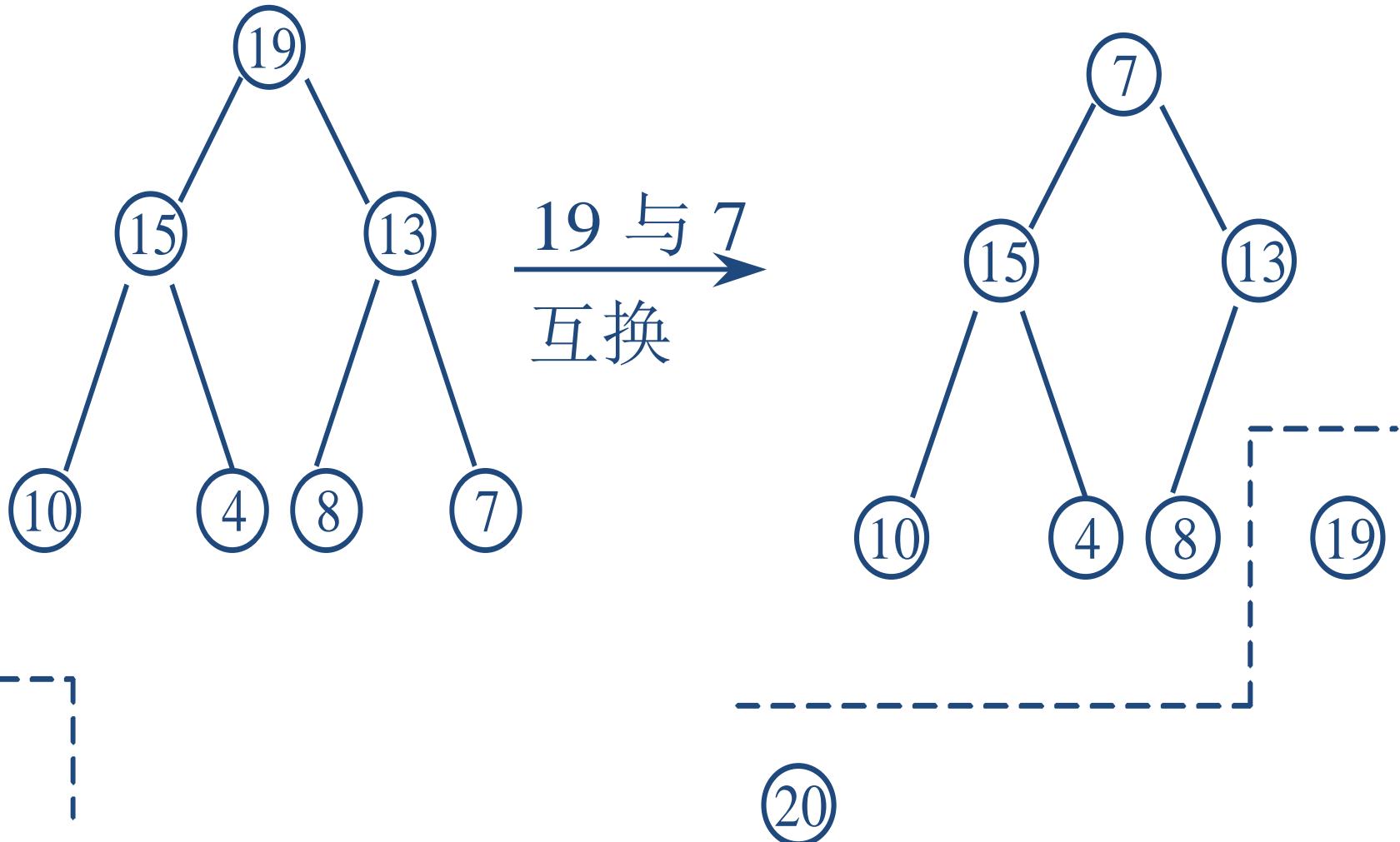
7比15和20小，把最大的子节点20和7换，7
还是比子节点13,19小，再和19换



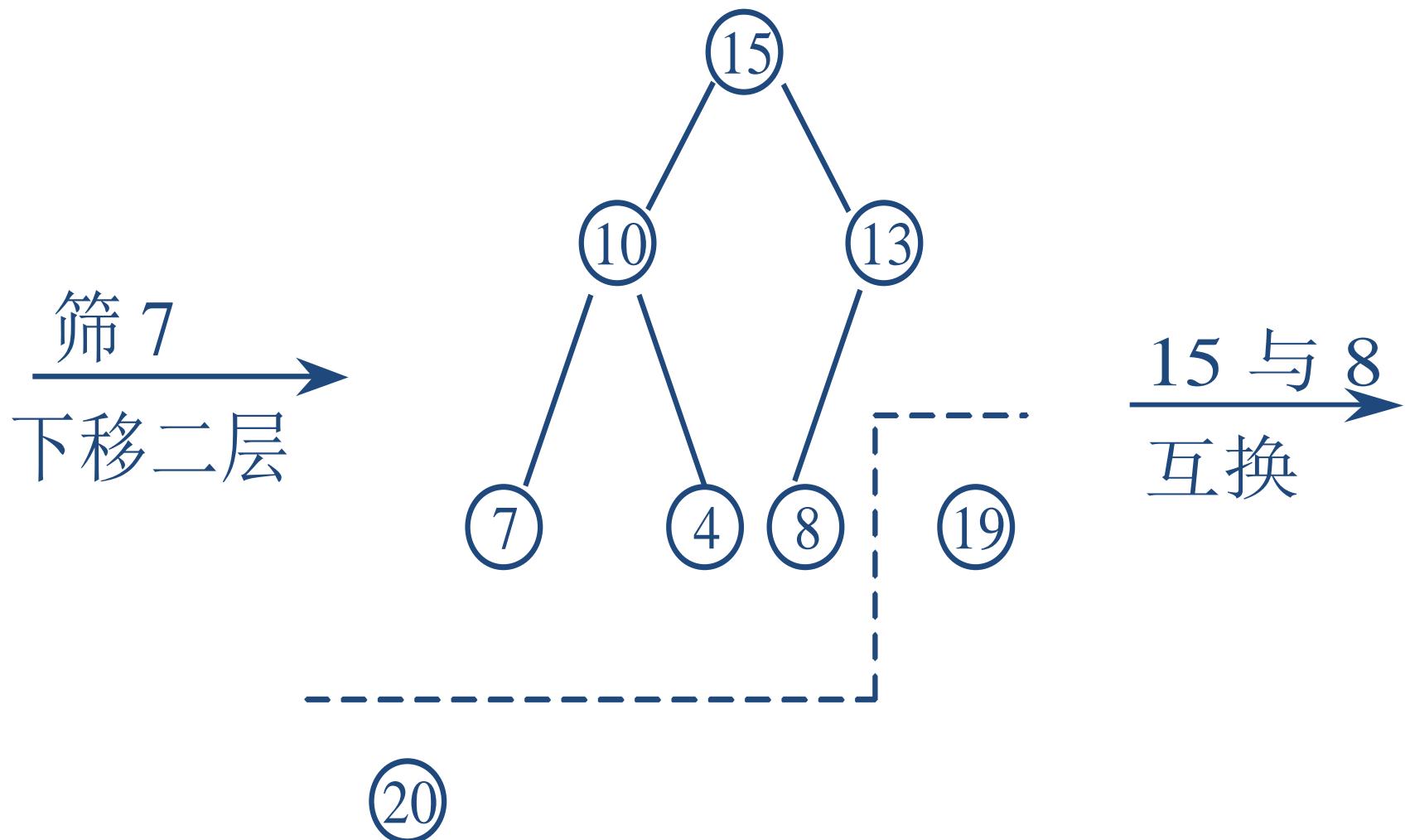
利用堆选出Top K (=4)

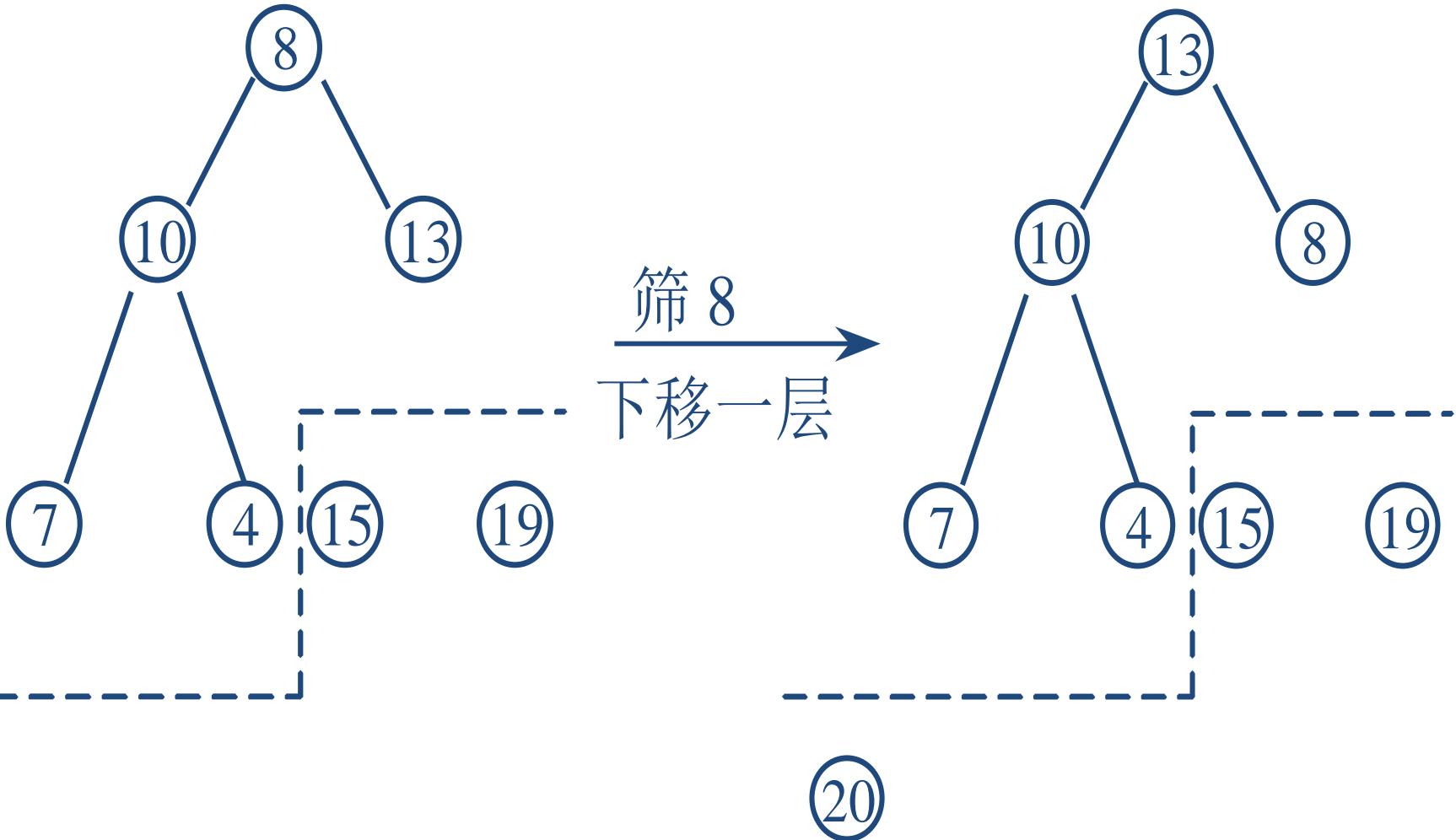


选第二最大值同样的处理，
直到选出k个最大值

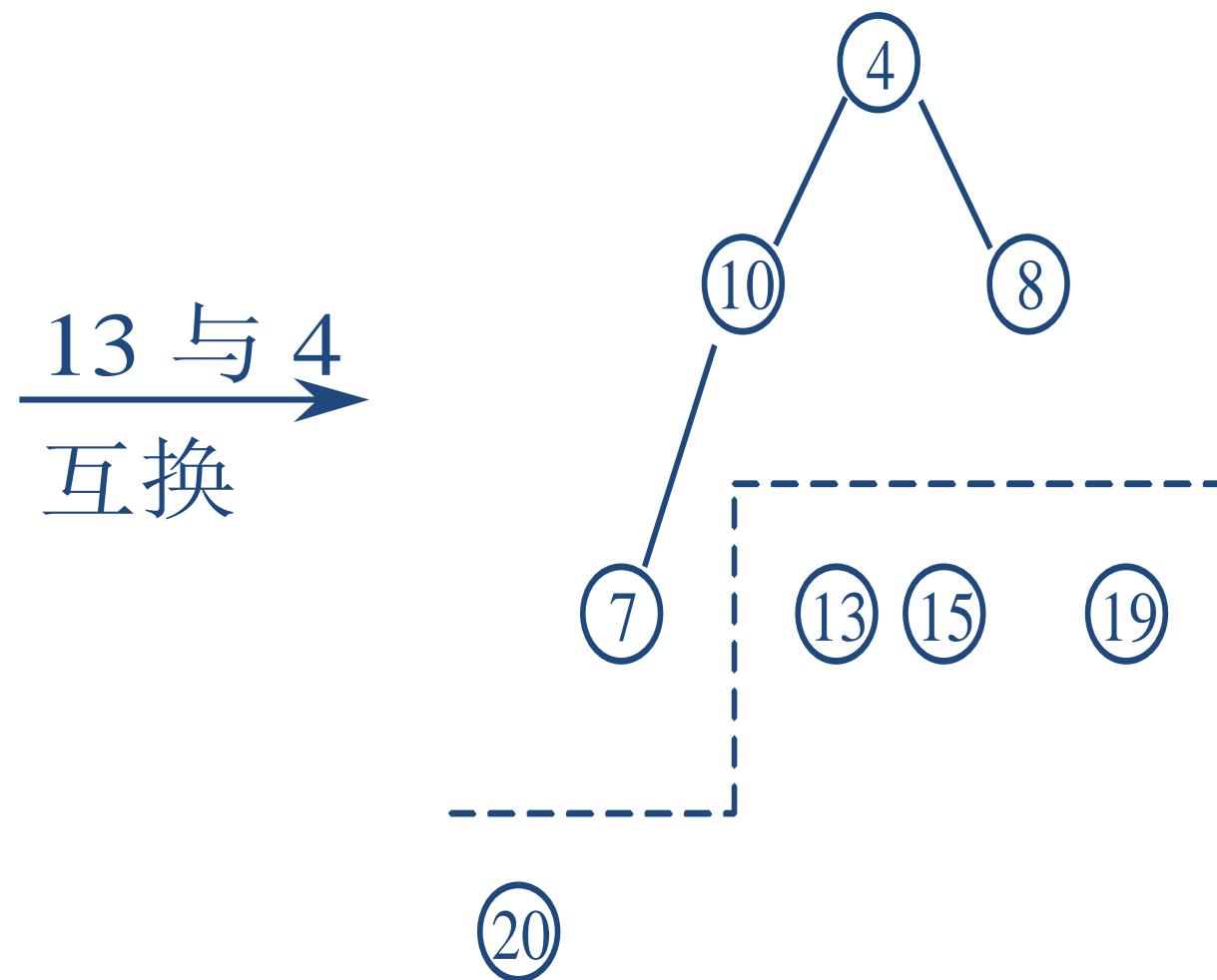


选第3最大值（例子中为15）同样的处理，直到选出k个最大值





选第4最大值（例子中为13）
 $K=4$, 结束



小结：堆排序

- 特点

- 堆排序其实也是一种选择排序，是一种树形选择排序。只不过直接选择排序中，为了从 $R[1...n]$ 中选择最大记录，需比较 $n-1$ 次，然后从 $R[1...n-2]$ 中选择最大记录需比较 $n-2$ 次。事实上这 $n-2$ 次比较中有很多已经在前面的 $n-1$ 次比较中已经做过，而树形选择排序恰好利用树形的特点保存了部分前面的比较结果，因此可以减少比较次数。

- 运算量

- 对于 n 个关键字序列，最坏情况下每个节点需比较 $\log_2(n)$ 次，因此其最坏情况下时间复杂度为 $n \log n$ 。

冒泡排序（Bubble Sort），时间复杂度为 $O(n^2)$ ，空间复杂度为 $O(1)$ ，即就地排序

精确top K 检索加速方法三： 提前终止计算

- 到目前为止的倒排记录表都按照docID排序
- 接下来将采用与查询无关的另外一种反映结果好坏程度的指标(静态质量)
 - 例如： 页面d的PageRank $g(d)$ ， 就是度量有多少好页面指向d的一种指标（参考第 21 章）
 - 于是可以将文档按照PageRank排序 $g(d_1) > g(d_2) > g(d_3) > \dots$
 - 将PageRank和余弦相似度线性组合得到文档的最后得分
$$\text{net-score}(q, d) = g(d) + \cos(q, d)$$

提前终止计算

- 假设：
 - (i) $g \rightarrow [0, 1]$;
 - (ii) 检索算法按照 d_1, d_2, \dots , 依次计算(为文档为单位的计算, document-at-a-time), 当前处理的文档的 $g(d) < 0.1$;
 - (iii) 而目前找到的 top K 的得分中最小的都 > 1.2
- 由于后续文档的得分不可能超过 1.1 ($\cos(q, d) < 1$)
- 所以, 我们已经得到了 top K 结果, 不需要再进行后续计算

小结：精确top K 检索及其加速

- 方法一：快速计算余弦
 - 无权重
- 方法二：堆排序法**N**中选**K**
 - 堆构建：需要 $2J$ 次操作
 - 选出前**K**个结果：每个结果需要 $2\log J$ 步
- 方法三：提前终止计算
 - $\text{net-score}(q, d) = g(d) + \cos(q, d)$

提纲

- ① 上一讲回顾
- ② 结果排序的重要性
- ③ 结果排序的实现

精确top K 检索及其加速办法

非精确top K检索

- ④ 完整的搜索系统

非精确top K检索

- 策略一：索引去除(**Index elimination**)
- 策略二：胜者表
- 策略三：静态得分
- 策略四：影响度排序
- 策略五：簇剪枝方法——预处理
- 策略六：参数化索引以及域索引
- 策略七：层次索引

精确top K检索的问题

- 仍然无法避免大量文档参与计算
- 一个自然而然的问题就是能否尽量减少参与计算文档数目，即使不能完全保证正确性也在所不惜。
 - 即采用这种方法得到的top K虽然接近但是并非真正的 top K——非精确top K检索

非精确top K检索的可行性

- 检索是为了得到与查询匹配的结果，该结果要让用户满意
- 余弦相似度是刻画用户满意度的一种方法
- 非精确top K的结果如果和精确top K的结果相似度相差不大，应该也能让用户满意

一般思路

- 找一个文档集合A， $K < |A| \ll N$ ，利用A中的top K结果代替整个文档集的top K结果
 - 即给定查询后，A是整个文档集上近似剪枝得到的结果
- 上述思路不仅适用于余弦相似度得分，也适用于其他相似度计算方法

策略一：索引去除 (Index elimination)

- 对于一个包含 **多个词项** 的查询来说，很显然我们可以仅仅考虑那些**至少包含一个查询词项的文档**
- 可以进一步拓展这种思路
 - 只考虑那些词项的 **idf 值** 超过一定阈值的文档
 - 只考虑包含 **多个查询词项**（一个特例是包含全部查询词项）的文档

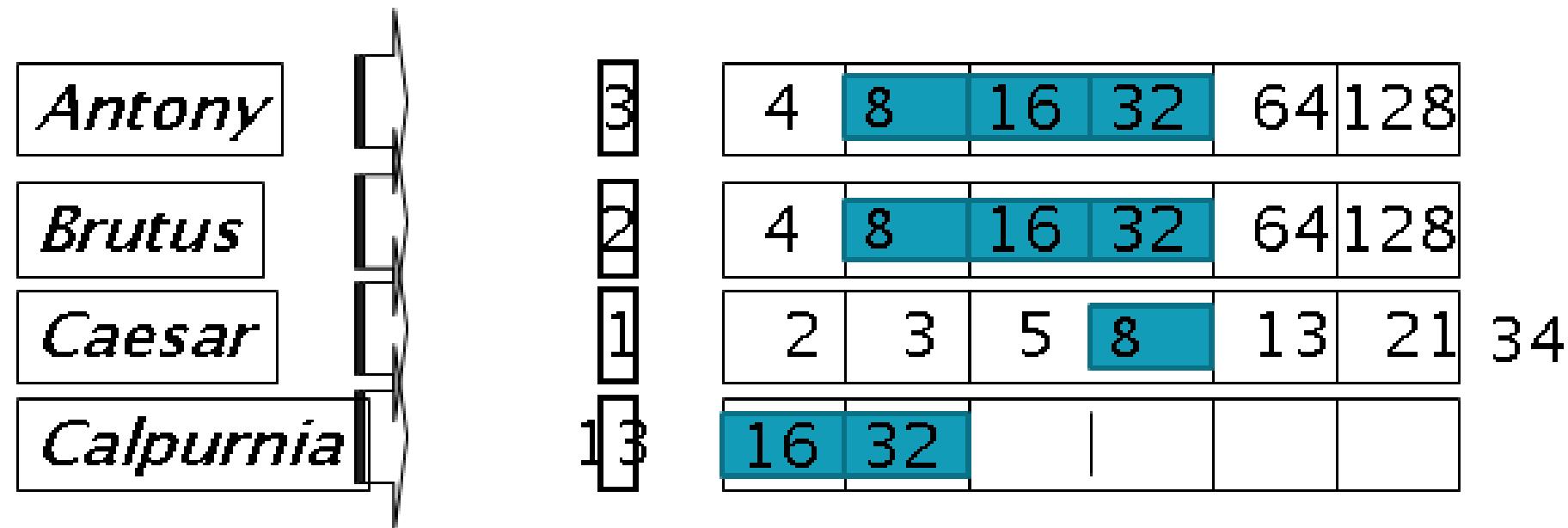
词项的 idf 值超过一定阈值的文档

- 在查询 *catcher in the rye* 时
- 只有 *catcher* 和 *rye* 的倒排记录表才会被遍历
- 显而易见: *in* 和 *the* 的作用很小
- 优点:
 - 含有低 idf 值的词项的文档非常多, 采用这种方法可以将大量无关的文档从候选集合 A 中去除

包含多个查询词项的文档

- 那些至少包含一个查询词项的文档才有可能成为候选文档
- 对于多词项查询，只考虑那些包含较多查询词项的文档
 - 比如，至少含有超过 $3/4$ 的查询词项
- 可以在倒排记录表遍历过程中实现

4中含3

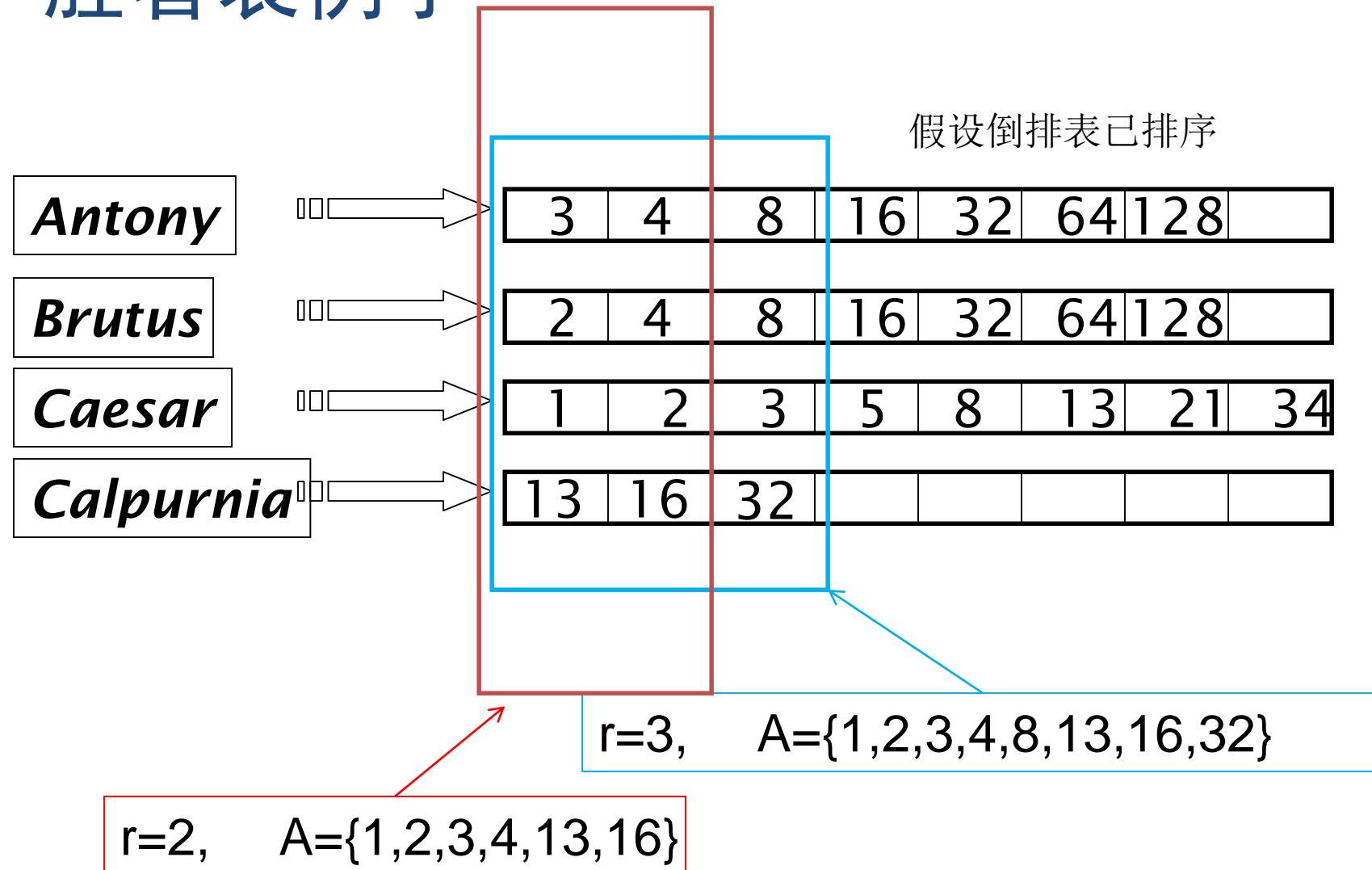


仅对文档8、16和32进行计算

策略二：胜者表

- 对于词典中的每个词项 t , 预先计算出 r 个最高权重的文档
 - 词项 t 所对应的 tf 值最高的 r 篇文档构成 t 的胜者表
 - 也称为**优胜表 (fancy list)** 或**高分文档 (top doc)**
- 其中 r 的值需要在索引建立之时给定
 - 因此, 有可能出现 $r < K$ 的情况
- 给定查询 q , 对查询 q 中所有词项的胜者表求并集, 并可以生成集合 A
 - 根据余弦相似度大小从 A 中选取前 $\text{top } K$ 个文档

胜者表例子



课堂思考

- 胜者表方式和前面的索引去除方式有什么关联？如何融合它们？
- 如何在一个倒排索引当中实现胜者表？
 - 提醒：胜者表与docID大小无关

策略三：静态得分

- 我们希望排序靠前的文档既是相关的又是权威的
- 相关性通过余弦相似度得分来判断
- 权威性是与query无关的文档本身的属性决定的
- 权威性标志举例
 - 维基百科
 - 报纸上的文章
 - 很多引用的文章
 - ~~del.icio.us diggs等网站~~
 - PageRank值

Quantitative



权威度计算

- 为每篇文档赋予一个与查询无关的(query-independent) [0, 1]之间的值，记为 $g(d)$
- 同前面一样，最终文档排名基于 $g(d)$ 和相关度的线性组合。
 - $\text{net-score}(q, d) = g(d) + \text{cosine}(q, d)$
 - 可以采用等权重，也可以采用不同权重
 - 可以采用任何形式的函数，而不只是线性函数
- 接下来我们的目标是找net-score最高的top K文档（非精确检索）

基于net-score的Top K文档检索

- 首先按照 $g(d)$ 从高到低将倒排记录表进行排序
- 该排序对所有倒排记录表都是一致的(只与文档本身有关)
- 因此，可以并行遍历不同查询词项的倒排记录表来
 - 进行倒排记录表的合并
 - 余弦相似度的计算

利用 $g(d)$ 排序的优点

- 这种排序下，高分文档更可能在倒排记录表遍历的前期出现
- 在时间受限的应用当中（比如，任意搜索需要在50ms内返回结果），上述方式可以提前结束倒排记录表的遍历

全局胜者表

- 对于选择好的 r 值，对每个词项 t 构建一个全局胜者表
- 其中包含了 $g(d) + \text{tf-idf}_{td}$ 得分最高的 r 篇文档
- 当查询提交以后，对所有全局胜者表的并集中的文档计算其最后得分
- 根据最终得分选择 top K

高端表 (High list) 和低端表 (Low list)

- 对每个词项，维护两个倒排记录表，分别成为高端表和低端表
 - 比如可以将高端表看成胜者表
- 遍历倒排记录表时，仅仅先遍历高端表
 - 如果返回结果数目超过K，那么直接选择前K篇文档返回
 - 否则，继续遍历低端表，从中补足剩下的文档数目
- 上述思路可以直接基于词项权重，不需要全局量 $g(d)$
- 实际上，相当于将整个索引分层

策略四：影响度排序

- 将词项 t 对应的所有文档 d 按照 tf_{td} 值降序排列
- 不同词项倒排记录表中文档所采用的排序方式就不是统一的
- 不能通过并发扫描多个倒排记录表的方式来计算文档的得分
- 有两种思路可以显著降低用于累加得分的文档数目

思路1： 提前结束

- 对某个查询词项 t 对应的倒排记录表进行从前往后扫描时，可以在某个阶段停止
 - 停止条件一：扫描了 r 篇固定数目的文档或者采用
 - 停止条件二：是当前记录的 tf_{td} 已经低于某个阈值

思路2：词项按照idf 降序排列

- 词项按照idf 降序排列 (query里面的词项)
 - 对最终得分贡献最大的查询词项首先被考虑
- 在查询处理过程中进行自适应处理
 - 当遇到具有较低idf 值的查询词项时，可以根据和前一个查询词项的文档得分的改变值来决定是否需要处理，改变值达到最小限度的时候终止。
- 例如：查询*catcher in the rye* 时，按idf排序应该是 catcher, rye, in, the，当我们依次处理，处理到in时发现改变值很小，对排序影响很小，所以终止。
- 上述思路给出所有方法所共有的一个一般形式。我们也可以实现另外的静态得分情况下的倒排索引

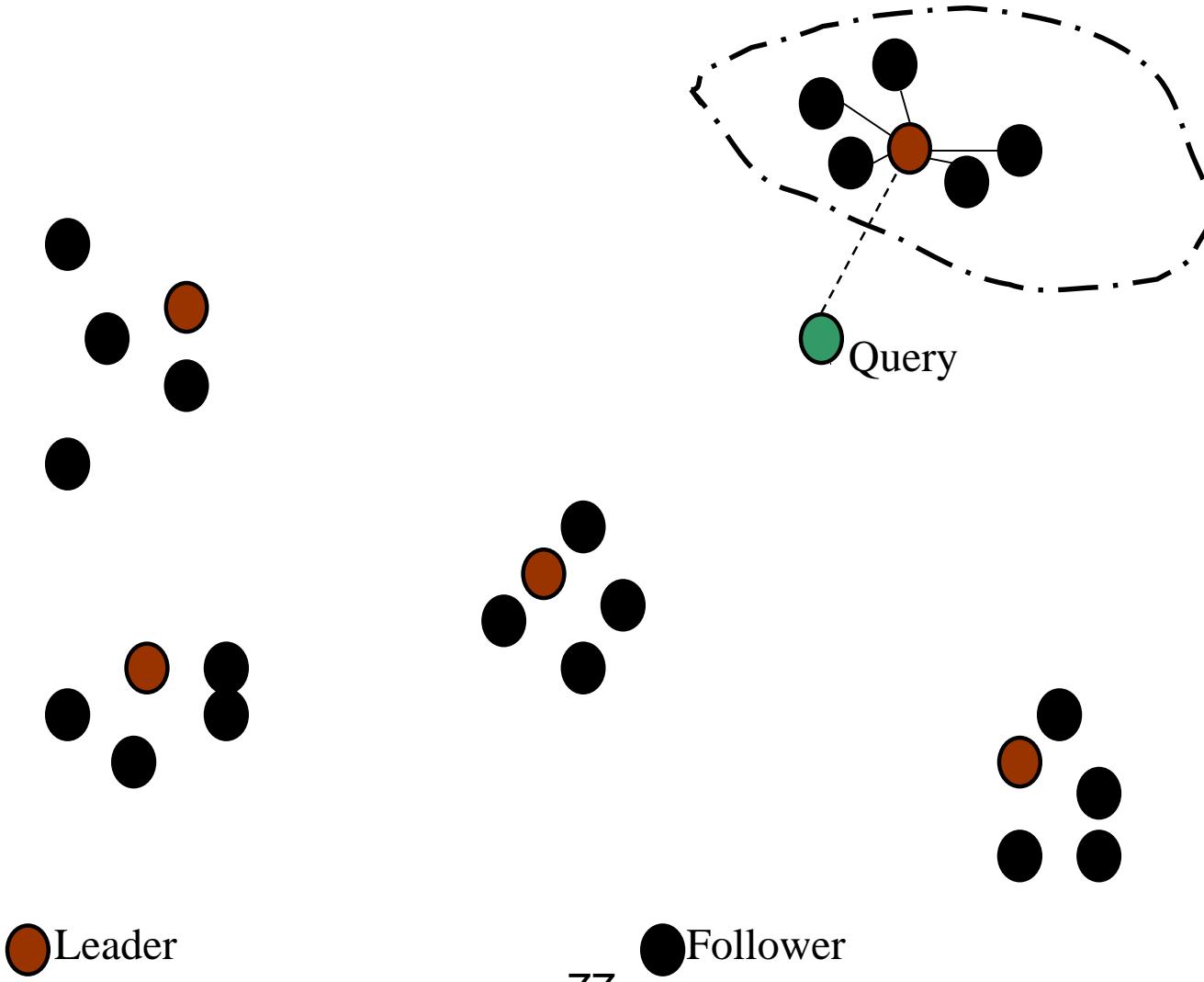
策略五：簇剪枝方法——预处理

- 从 N 篇文档组成的文档集中随机选出 \sqrt{N} 篇文档，它们称为先导者（leader）集合
- 对于每篇不属于先导者集合的文档，计算离之最近的先导者
 - 不属于先导者集合的文档称为追随者
 - 对于 N 篇随机选出的先导者文档，其期望分配到的追随者个数大约为 \sqrt{N}

簇剪枝方法——查询处理

- 查询处理：
 - 给定查询 q , 通过与先导者计算余弦相似度, 找出和它最近的先导者 L
 - 候选集合 A 包括 L 及其追随者, 然后对 A 中的所有文档计算余弦相似度

可视化示意图



为什么采用随机抽样？

- 速度快
- 先导者能够反映数据的分布情况

常见变形

- 预处理时，我们将每个追随者分配给离它最近的 b_1 个先导者
- 查询处理时，我们将考虑和查询 q 最近的 b_2 个先导者
- 很显然，前面讲到的方法只是该方法在 $b_1 = b_2 = 1$ 情况下的一个特例

思考一下

- 在簇剪枝方法中，第一步发现最近先导者过程中需要多少次余弦计算？
- 前一页中变量 $b1, b2$ 的作用？

策略六：参数化索引以及域索引

- 迄今，我们一直将文档看成由词项组成的有序排列
- 事实上，一个文档一般是由几个部分组成，这些部分有不同的意义，如：
 - 作者
 - 题目
 - 正文
 - 语言
 - 格式
 -
- 这些构成了一个文档的元数据

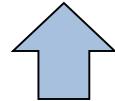
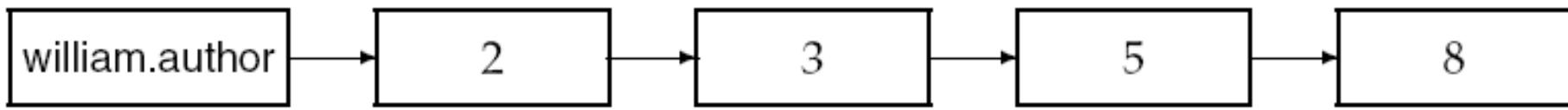
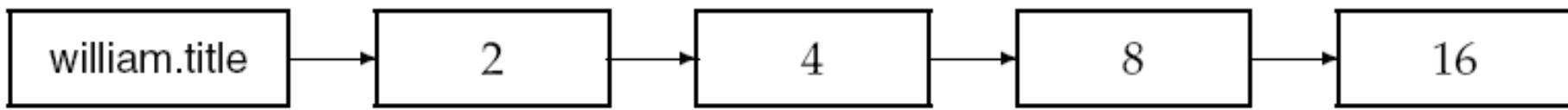
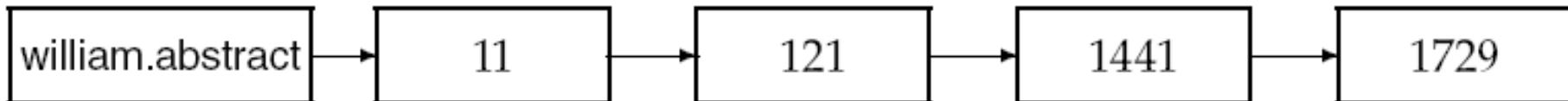
字段 (Field)

- 我们经常希望检索这些元数据
 - 如： 寻找莎士比亚在1601年写的小说， 文中包含 *alas poor Yorick* 这几个词
- ($\text{Year}=1601$) 就是一个字段
- 同样， ($\text{作者}=$ 莎士比亚) 也是一个字段
- 按字段查询一般都属于联合查询
 - 即每个域条件都满足

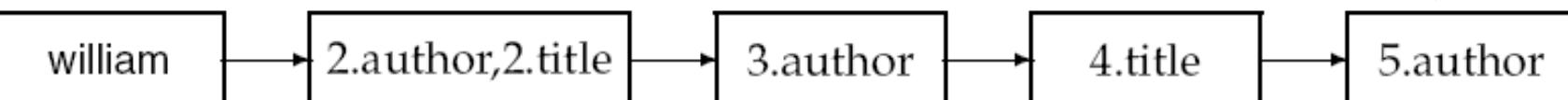
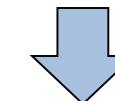
域 (zone)

- 域是一个可以包含任意内容的区域，如：
 - 题目
 - 摘要
 - 引用…
- 在域上建立倒排索引同样可以进行查询
- 例如：“寻找标题中出现merchant、作者中出现william且正文中出现gentle rain的文档”

域索引



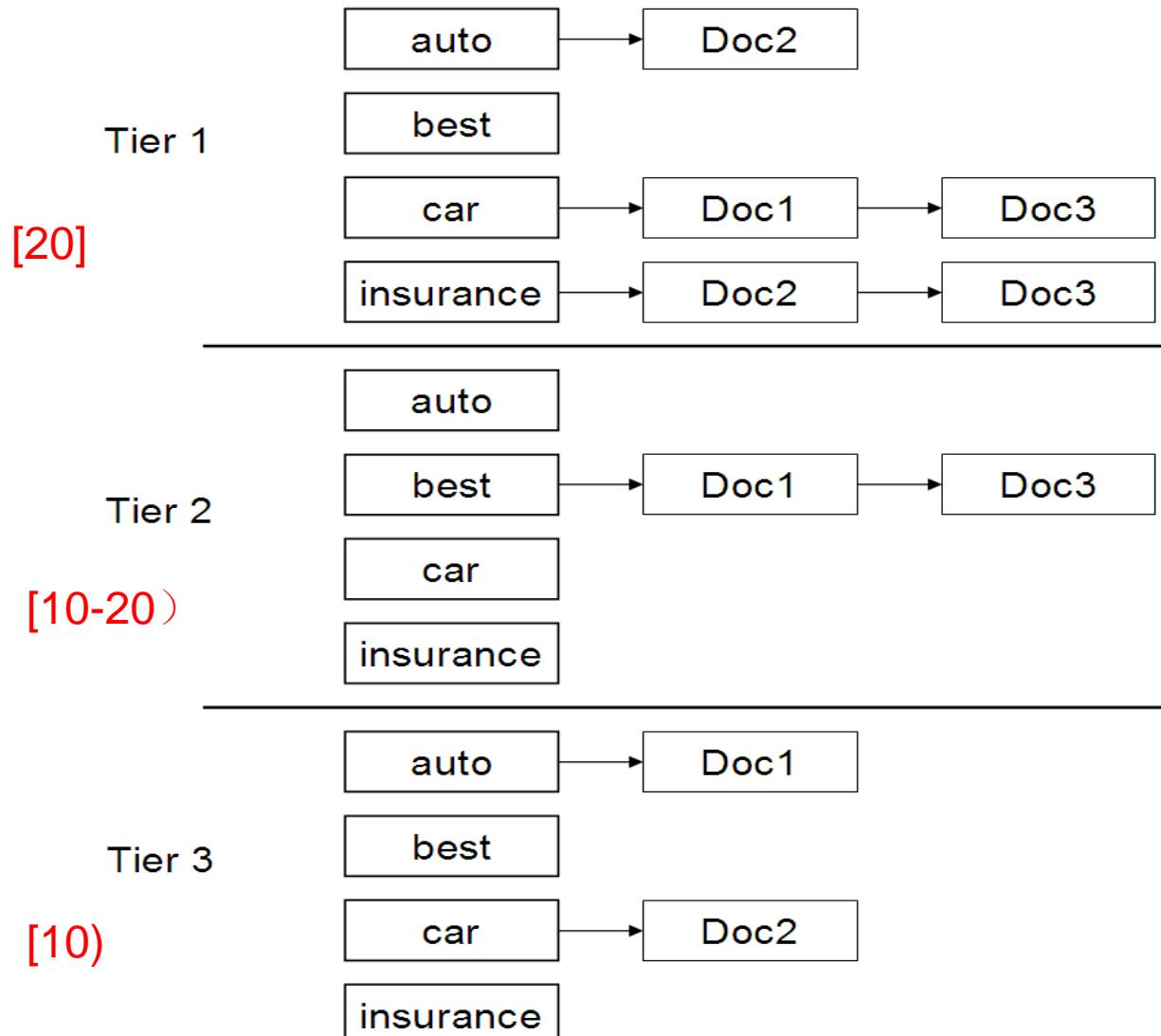
Encode zones in dictionary vs. postings.



策略七：层次索引

- 可以看成是优胜表的一般化形式
 - 最重要
 - ...
 - 最不重要
- 可以用静态得分或者其它得分衡量
- 倒排记录表按照重要性降序转化成层次索引
- 查询是只用上层索引，除非上层索引返回结果小于 K
 - 上层返回结果小于K则再从下一层中检索

层次索引



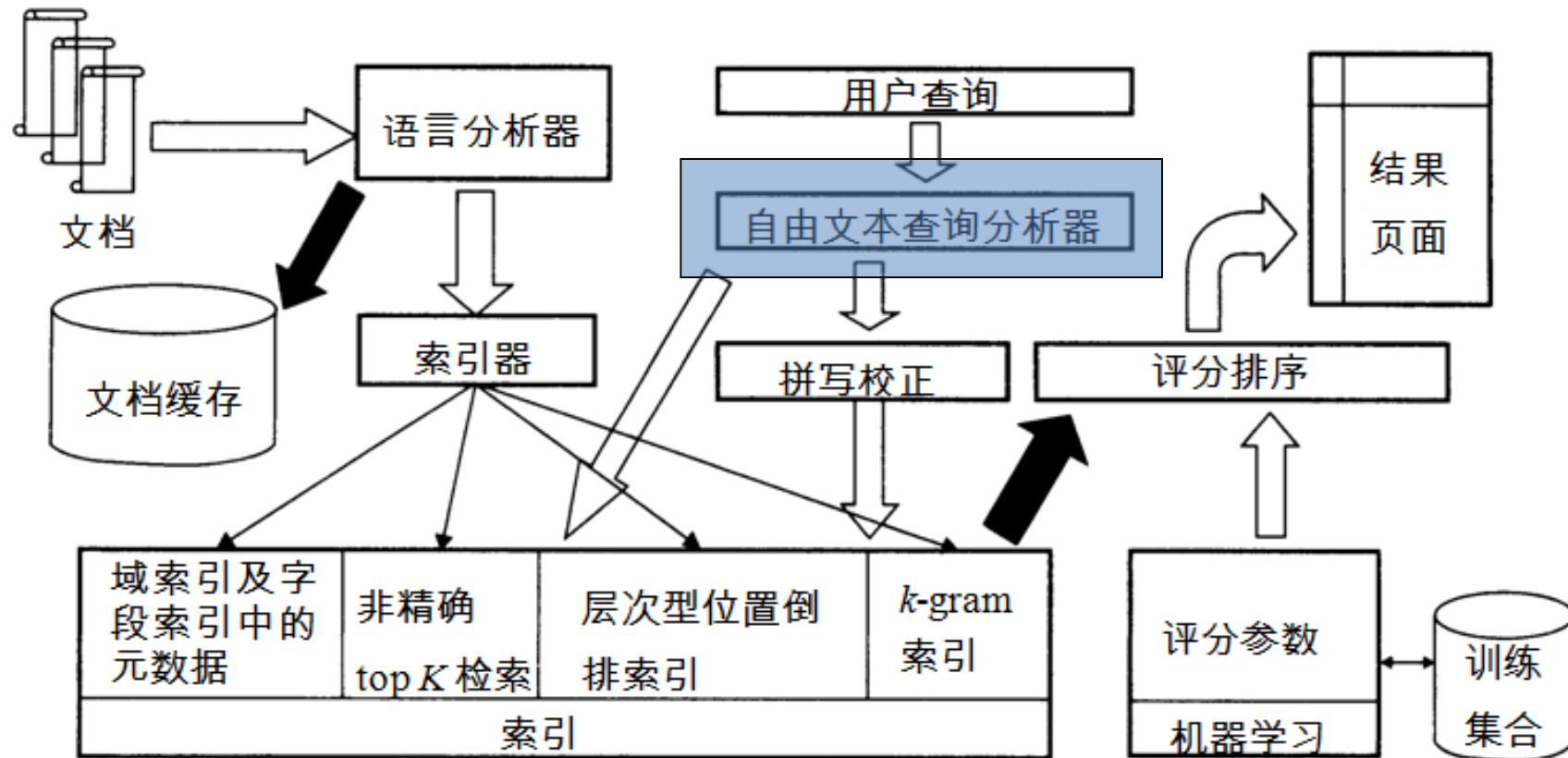
小结：非精确top K检索

- 策略一：索引去除(**Index elimination**)
- 策略二：胜者表
- 策略三：静态得分
- 策略四：影响度排序
- 策略五：簇剪枝方法——预处理
- 策略六：参数化索引以及域索引
- 策略七：层次索引

提纲

- ① 上一讲回顾
- ② 结果排序的重要性
- ③ 结果排序的实现
- ④ 完整的搜索系统

搜索系统组成



查询词项的邻近性

- 自由文本查询：用户输入几个词项到搜索框——一般的互联网检索
- 用户往往希望返回的文档中大部分或者全部查询词项之间的距离比较近
- 令文档 d 中包含所有查询词项的最小窗口大小为 ω ，其取值为窗口内词的个数
- 假设某篇文档仅仅包含一个句子 $The\ quality\ of\ mercy\ is\ not\ strained$, 那么查询 $strained\ mercy$ 在此文档中的最小窗口大小是4
- 用窗口大小来度量位置关系

词项的邻近性示例

https://www.baidu.com/s?wd=%E4%B8%87%E7% 万立骏 化学_百度搜索 侯建国副部长做主题报告

Baidu 百度 万立骏 化学 百度一下 添加百度到桌面，搜索更便捷！ fa...7@tom.com

网页 新闻 贴吧 知道 音乐 图片 视频 地图 文库 更多»

百度为您找到相关结果约65,800个

万立骏_百度百科

姓名: **万立骏**
生日: 1957年7月 职业: 物理**化学家**
简介: **万立骏**, 博士, 物理**化学家**, 中国科学院**化学**研究所研究员。
1987年获大连理工大学硕士学位, 1996年获日本东北大学博士学位。
主要从事**电化学**、SPM技术、纳米**化学**和纳米材料的研究...
个人简介 **履历** **学术成就** **所获奖项** **代表性论著**
baike.baidu.com/2014-04-17

万立骏 化学的最新相关信息

万立骏任中科大校长 曾有二十余项发明专利(图)

 至此,中国五所著名大学的校长全部完成换届,巧合的是,**万立骏**与此前上任的北京大学新校长林建华、清华大学新校长邱勇同为**化学家**出身,中国三所顶尖高校...
新浪新闻 1天前

万立骏：“当好大家的服务员” [中国教育新闻网](#) 1天前
万立骏任中科大校长 [环球网](#) 1天前
中科大新任校长万立骏系海归院士的杰出.... [网易新闻](#) 2天前
万立骏 - 万立骏 简历 资料 新闻-中国党.... [中国共产党新闻网](#) 3天前

中国3所顶尖高校校长组成“化学三掌门”怎么解释 万立骏院士资料

 2天前 - 中国3所顶尖高校校长组成“化学三掌门”怎么解释 **万立骏**院士资料。值得一提的是,**万立骏**与此前上任的北京大学[微博]新校长林建华、清华大学[微博]新...
edu.qlw.com.cn/0327/3... - 百度快照 - 83%好评

相关人物 展开

**曹俊** 多少人羡慕不已的骄子
**袁帅** 清华大学教师
**刘换香** 现任兰州大学教授
**翁冰莹** 厦门大学文学院博士

**肖益鸿** 福大化学化工学院任职
**江雷** 中国著名纳米材料专家
**白春礼** 化学和纳米科技
**钱逸泰** 中国科学院院士

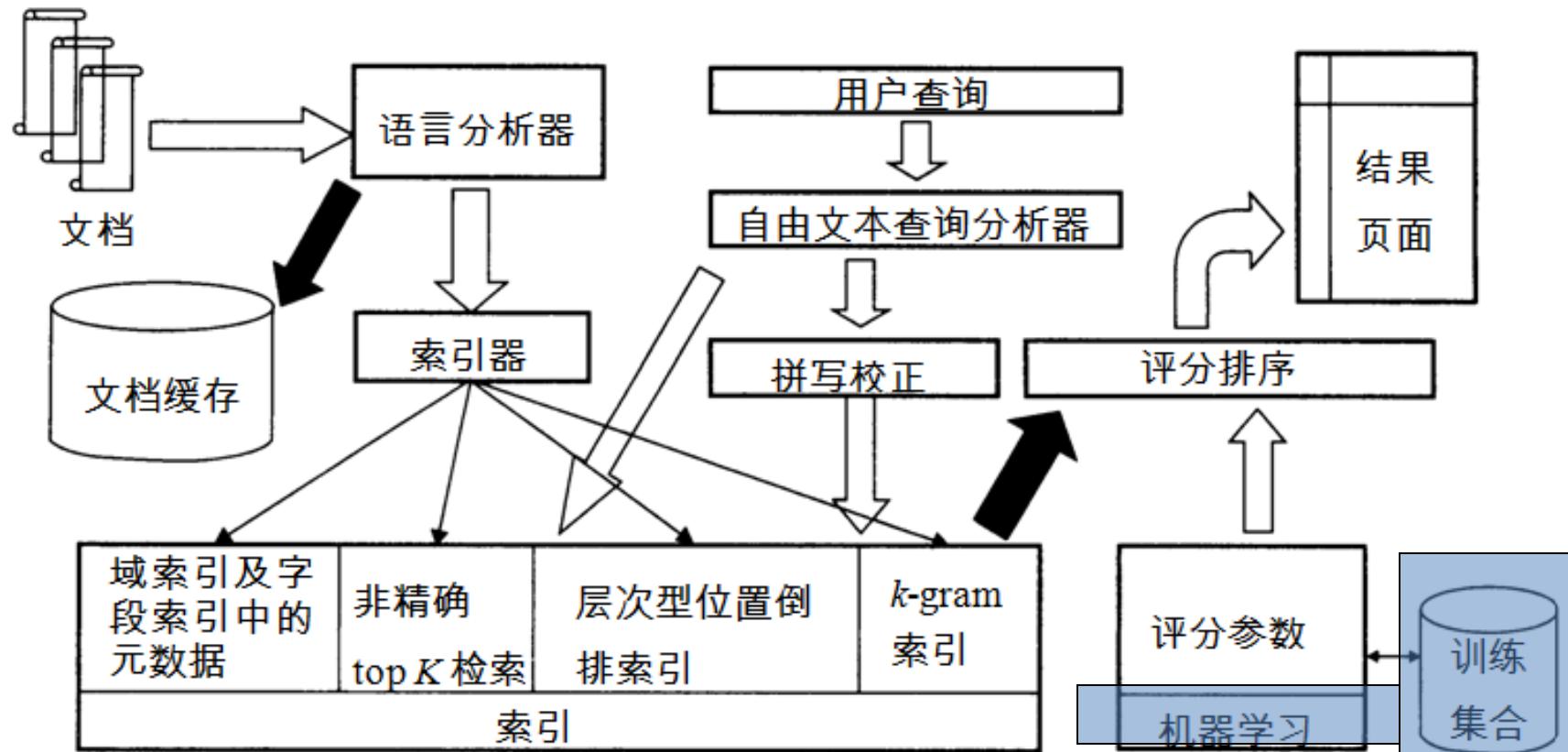
其他人还搜 展开

**方璐**
**侯建国** 中国科学技术大学校长
**徐良杰** 博士生导师
**王吉红** 科学家

查询分析器

- 自由文本查询对用户输入的关键词可能需要基于底层索引结果对多个查询进行处理，如查询 *rising interest rates* 之类 query 时，**查询分析器** 可能做如下操作：
 - 1. 将用户输入的查询字符串看成一个短语查询
 - 2. 如果包含短语 rising interest rates 的文档数目少于 10 篇，那么会将原始查询看成 rising interest 和 interest rates 两个查询短语，同样通过向量空间方法来计算。
 - 3. 如果结果仍然少于 10 个，重新利用向量空间模型求解，认为 3 个查询词项之间是互相独立的

搜索系统组成



综合评分

- 已经介绍的评分函数有余弦相似度、静态得分、近邻性等。
- 如何将这些评分组合才是最优的？
- 通用方法——机器学习

机器学习有下面几种定义：“机器学习是一门人工智能的科学，该领域的主要研究对象是人工智能，特别是如何在经验学习中改善具体算法的性能”。“机器学习是对能通过经验自动改进的计算机算法的研究”。“机器学习是用数据或以往的经验，以此优化计算机程序的性能标准。”一种经常引用的英文定义是：A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

课后练习

- 习题7-3
- 习题7-5
- 习题7-7

思考：

各种查询操作在向量空间模型中的实现

谢谢大家!