信息检索导论
Introduction to Information Retrieval

第2讲 布尔检索

授课人：李永可
邮箱：lyk@xjau.edu.cn

提 纲

- 信息检索概述
- 倒排索引
- 布尔查询的处理

信息检索Information Retrieval

- Information Retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).
- 信息检索是从大规模非结构化数据（通常是文本）的集合（通常保存在计算机上）中找出满足用户信息需求的资料（通常是文档）的过程。

信息检索Information Retrieval

- Document –文档
- Unstructured – 非结构化
- Information need –信息需求
- Collection—文档集、语料库

常见信息检索工具

名称	地址
百度	http://www.baidu.com/
谷歌	https://www.google.com.hk/
必应	http://cn.bing.com/
izda	http://www.izda.com/

IR vs数据库: 结构化 vs 非结构化数据

- 结构化数据即指“表”中的数据

Employee	Manager	Salary
Smith	Jones	50000
Chang	Smith	60000
Ivy	Smith	50000

数据库常常支持范围或者精确匹配查询 。 e.g.,
Salary < 60000 AND Manager = Smith.

非结构化数据

- 通常指自由文本
- 允许
  - -关键词加上操作符号的查询
  - -更复杂的 概念性查询
  - -找出所有的有关药物滥用(drug abuse)的网页
- 经典的检索模型一般都针对自由文本进行处理
- 非结构化数据,包括所有格式的办公文档、文本、图片、XML、HTML、各类报表、图像和音频/视频信息等等

半结构化数据

介于完全结构化数据（如关系型数据库、面向对象数据库中的数据）和完全无结构的数据（如声音、图像文件等）之间的数据，HTML文档就属于半结构化数据。它一般是自描述的，数据的结构和内容混在一起，没有明显的区分。

非结构化数据发展

随着网络技术的发展，特别是Internet和Intranet技术的飞快发展，使得非结构化数据的数量日趋增大。这时，主要用于管理结构化数据的关系数据库的局限性暴露地越来越明显。因而，数据库技术相应地进入了“后关系数据库时代”，发展进入基于网络应用的非结构化数据库时代。

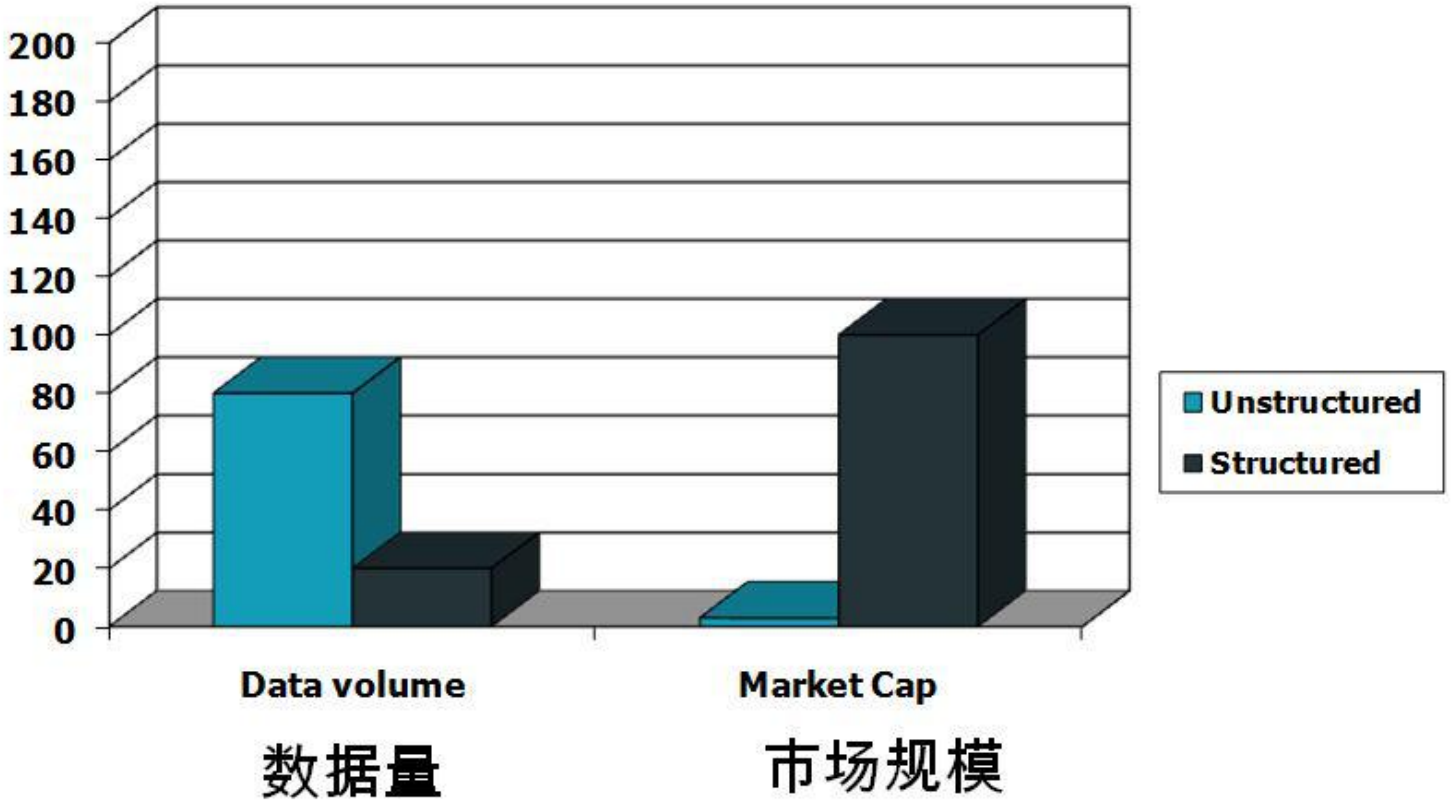
IBase

我国非结构化数据库以北京国信贝斯(iBase)软件有限公司的IBase数据库为代表。IBase数据库是一种面向最终用户的非结构化数据库，在处理非结构化信息、全文信息、多媒体信息和海量信息等领域以及Internet/Intranet应用上处于国际先进水平，在非结构化数据的管理和全文检索方面获得突破

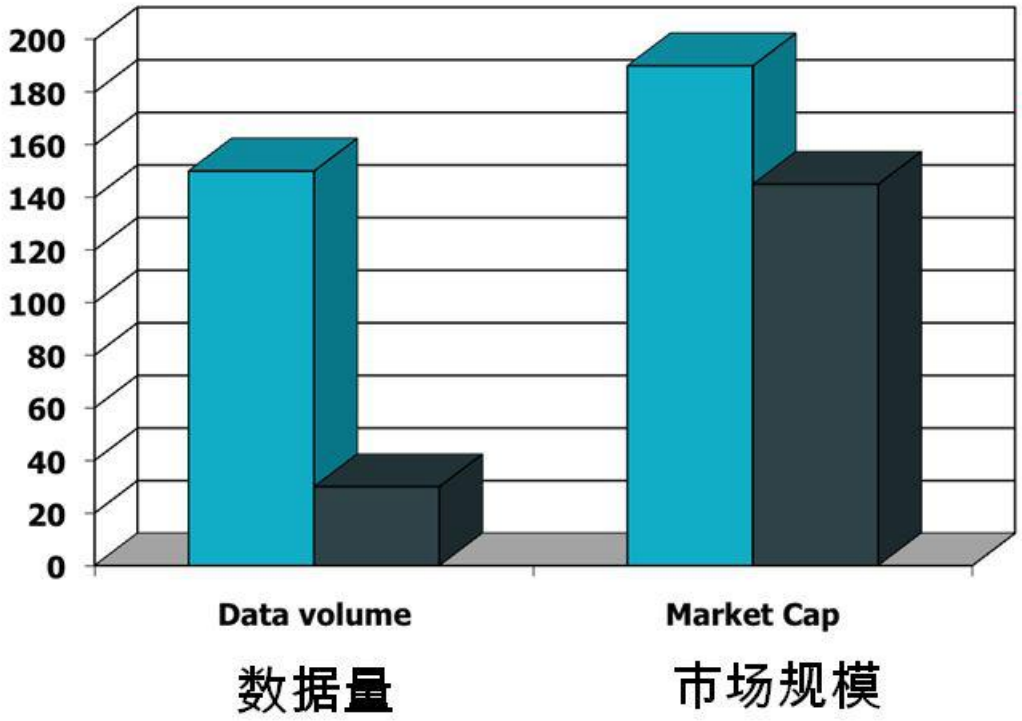
传统信息检索 vs. 现代信息检索

- 传统信息检索主要关注结构化、半结构化数据
- 现代信息检索中也处理结构化数据，但主要处理非结构化、半结构化数据

非结构化数据(文本) vs. 结构化数据(数据库) @ 1996年



非结构化数据(文本) vs. 结构化数据(数据库) @ 2009年



布尔检索

- 针对布尔查询的检索，布尔查询是指利用 AND, OR 或者 NOT操作符将词项 连接起来的查询
- 信息 AND 检索
- 信息 OR 检索
- 信息 AND 检索 AND NOT 教材

一个简单的例子(《三国演义》)

- 《三国演义》的那一章包括关羽、张飞不包括赵云？
- 布尔表达式：关羽 and 张飞 and not 赵云
- 笨方法： 从头到尾扫描所有章节，对每个章节判断它是否包含“关羽”and“ 张飞”，同时又不包含“赵云”
- 笨方法为什么不好?
  - 速度超慢 (特别是大型文档集)
  - 处理NOT “赵云” 并不容易（一旦包含即可停止判断）
  - 不太容易支持其他操作
  - 不支持检索结果的排序 (即只返回较好的结果)

词项-文档(term-doc)的关联矩阵

关羽 and 张飞 and not 赵云

	第一章	第二章	第三章	第四章	第五章	第六章
关羽	1	1	0	1	0	1
张飞	0	1	1	1	1	0
赵云	1	0	1	0	1	1

若某章节包含检索单词
则为1，否则为0

关联向量(incidence vectors)

- 关联矩阵的每一列都是 0/1向量，每个0/1都对应一个词项
- 给定查询**关羽 and 张飞 and not 赵云**
- 取出三个行向量，并对赵云的行向量求补，最后按位进行与操作
- 110101 and 011110 and 010100=010100

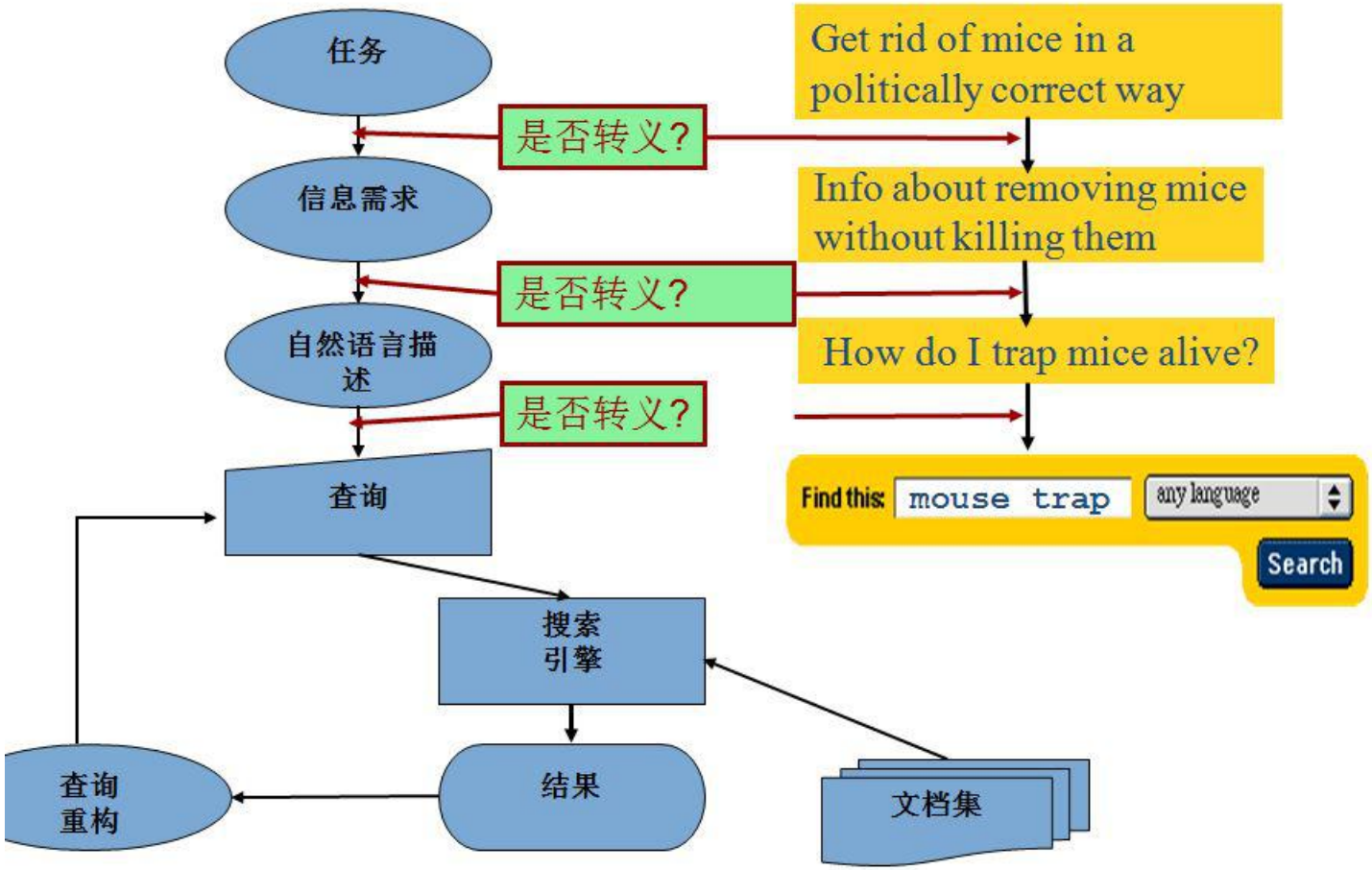
上述查询的结果文档

关羽 and 张飞 and not 赵云？

第二章

第四章

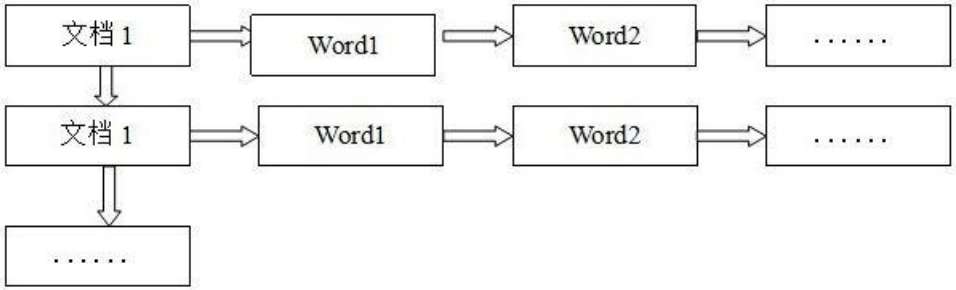
典型的搜索过程


任务
是否转义?
Get rid of mice in a politically correct way
信息需求
Info about removing mice without killing them
是否转义?
自然语言描述
How do I trap mice alive?
是否转义?
查询
Find this: mouse trap
any language
Search
搜索引擎
查询重构
结果
文档集

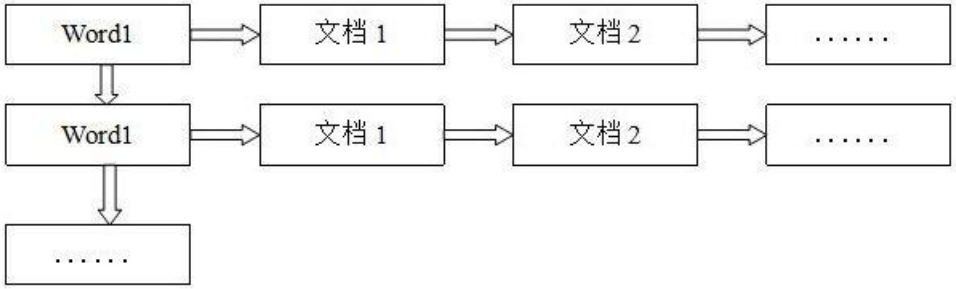
倒排索引vs正排索引

索引名称	结构特点
正排索引	以文档的ID为关键字，表中记录文档中每个字的位置信息，查找时扫描表中每个文档中字的信息直到找出所有包含查询关键字的文档
倒排索引	倒排表以字或词为关键字进行索引，表中关键字所对应的记录表项记录了出现这个字或词的所有文档，一个表项就是一个字表段，它记录该文档的ID和字符在该文档中出现的位置情况

倒排索引vs正排索引



正排索引



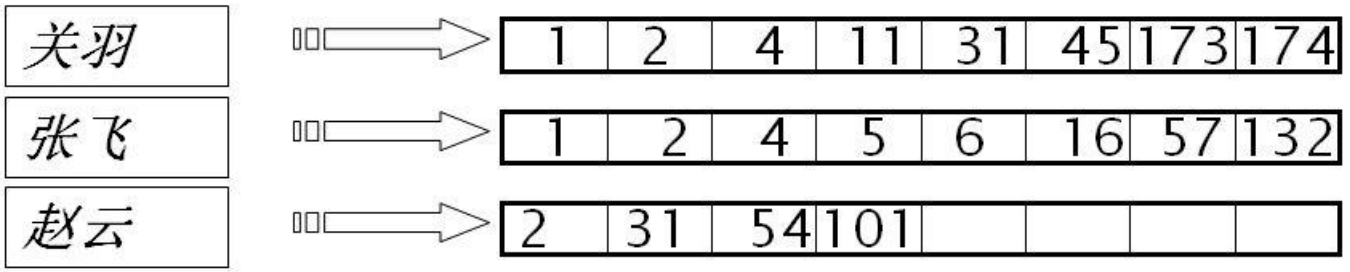
倒排索引

倒排索引vs正排索引

索引名称	效率
正排索引	需逐篇文章扫描，效率底下
倒排索引	直接定位包含关键字的文章，效率高

倒排索引(Inverted index)

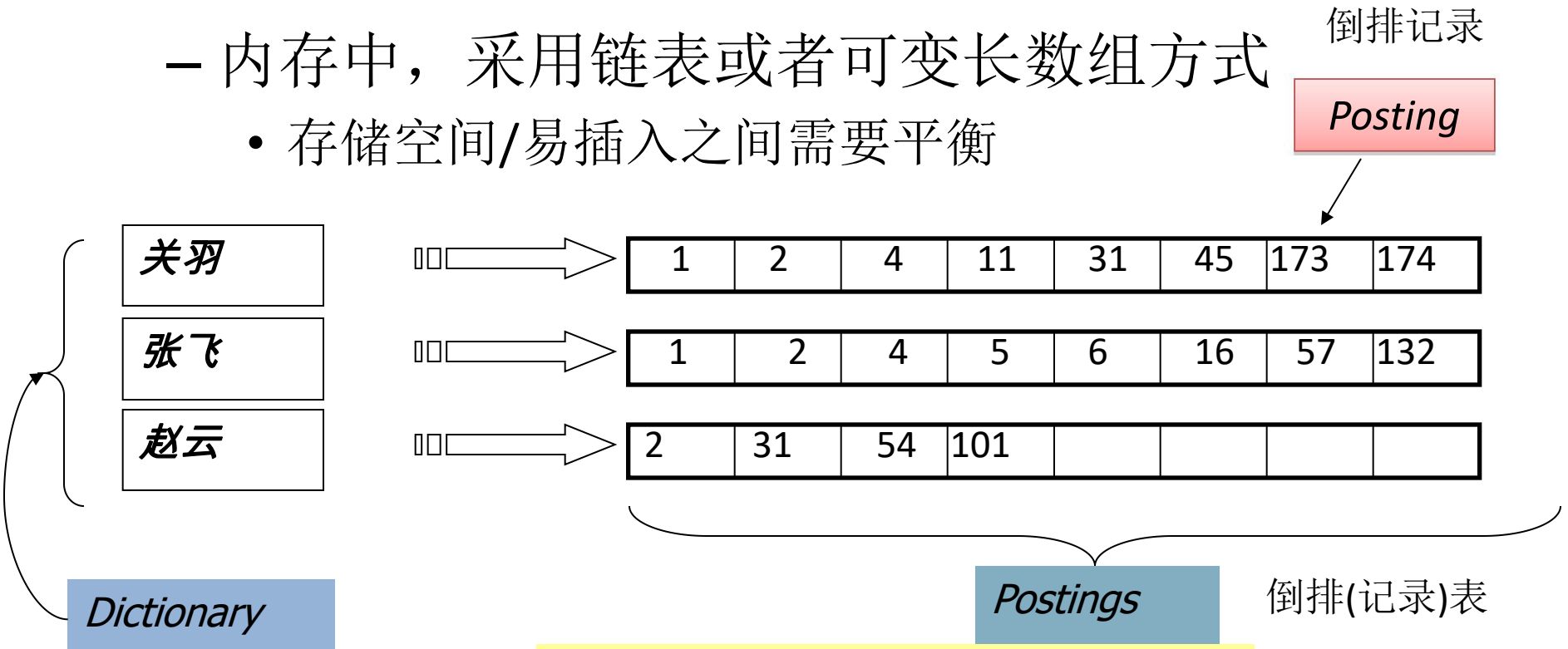
- 对每个词项t, 记录所有包含t的文档列表.
  - 每篇文档用一个唯一的 docID来表示，通常是正整数，如1,2,3…
- 能否采用定长数组的方式来存储docID列表



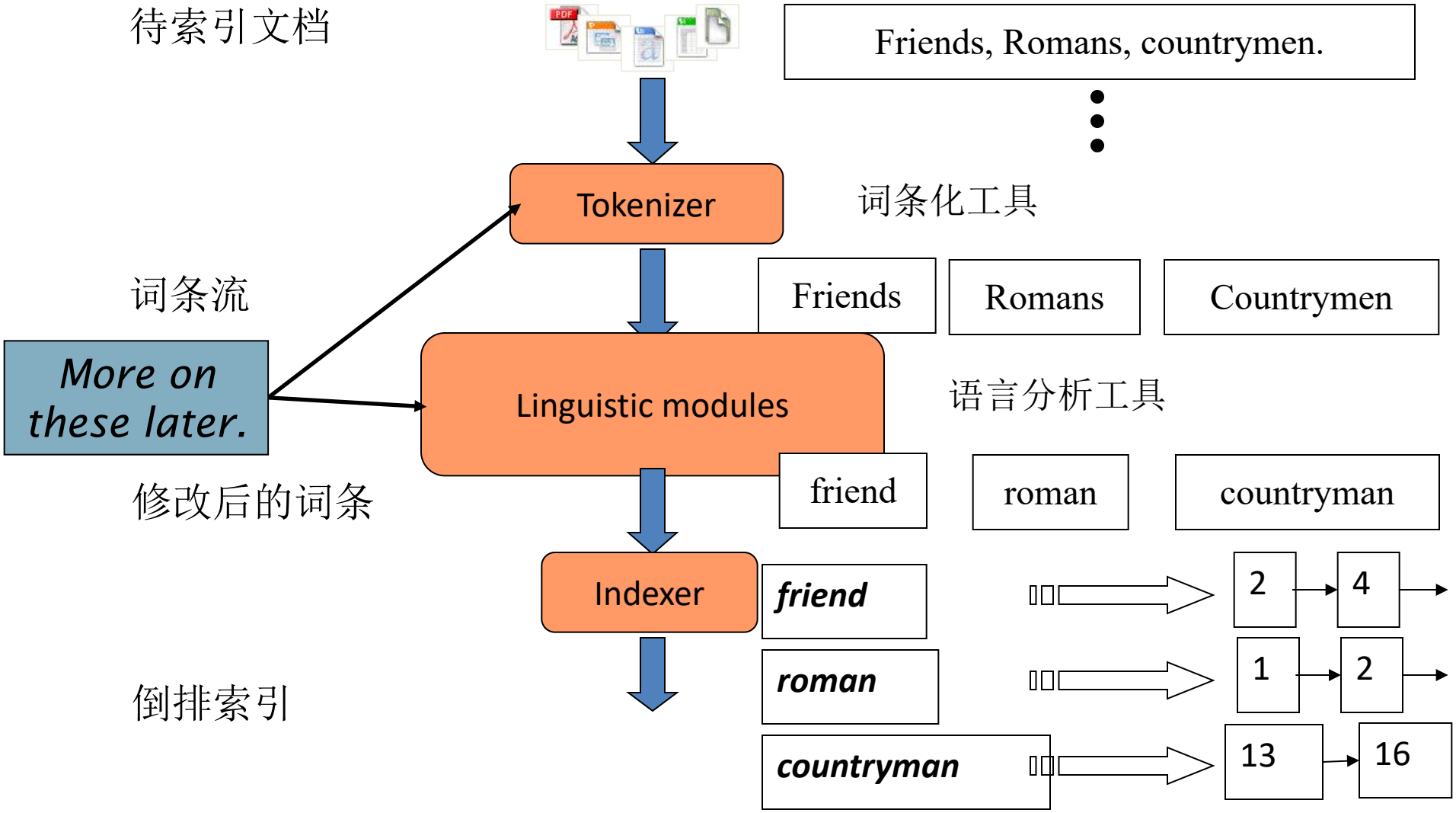
第10章加入单词张飞之后怎么办？

倒排索引(续)

- 通常采用变长表方式
  - 磁盘上，顺序存储方式比较好，便于快速读取
  - 内存中，采用链表或者可变长数组方式
    - 存储空间/易插入之间需要平衡



倒排索引构建


待索引文档
Friends, Romans, countrymen.
Tokenizer
词条化工具
词条流
Friends
Romans
Countrymen
More on these later.
Linguistic modules
语言分析工具
修改后的词条
friend
roman
countryman
Indexer
倒排索引
friend
2
4
roman
1
2
countryman
13
16

索引构建过程: 词条序列

- <词条，docID>二元组

Doc 1

I did enact Julius Caesar I was killed i' the Capitol; Brutus killed me.

Doc 2

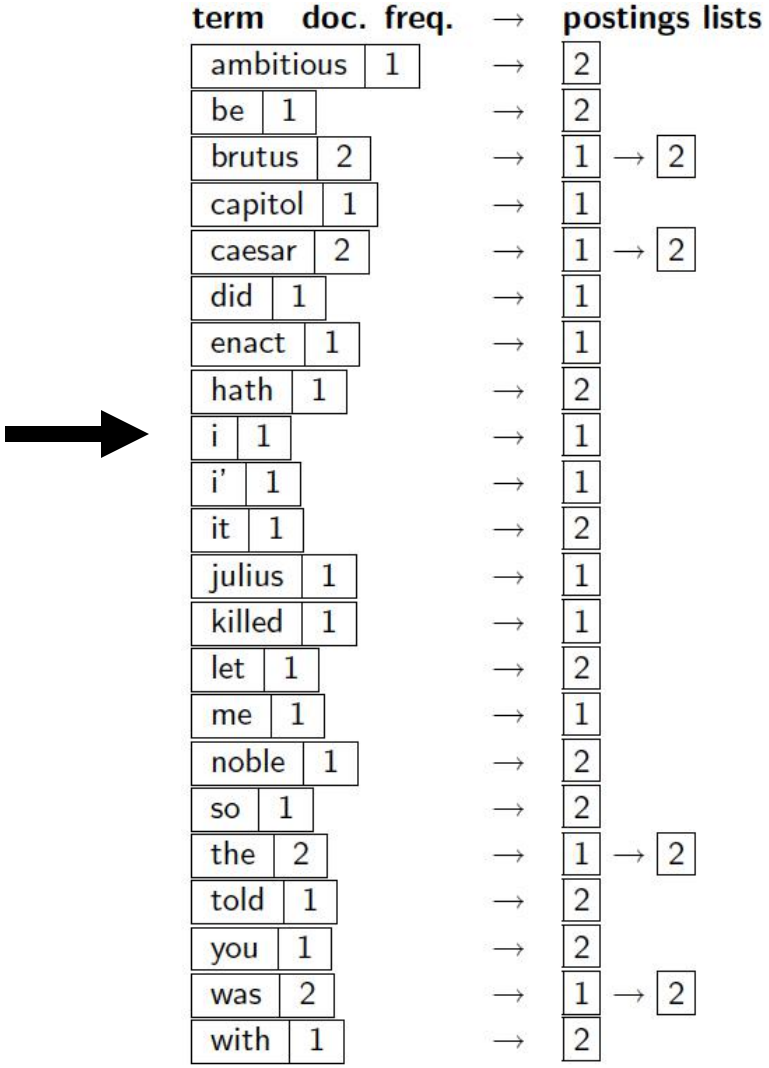
So let it be with Caesar. The noble Brutus hath told you Caesar was ambitious

Term	docID
I	1
did	1
enact	1
julius	1
caesar	1
I	1
was	1
killed	1
i'	1
the	1
capitol	1
brutus	1
killed	1
me	1
so	2
let	2
it	2
be	2
with	2
caesar	2
the	2
noble	2
brutus	2
hath	2
told	2
you	2
caesar	2
was	2
ambitious	2

索引构建过程: 词典 & 倒排记录表

- 某个词项在单篇文档中的多次出现会被合并
- 拆分成词典和倒排记录表两部分
- 每个词项出现的文档数目 (doc. frequency, DF) 会被加入

Term	docID
ambitious	2
be	2
brutus	1
brutus	2
capitol	1
caesar	1
caesar	2
caesar	2
did	1
enact	1
hath	1
I	1
I	1
i'	1
it	2
julius	1
killed	1
killed	1
let	2
me	1
noble	2
so	2
the	1
the	2
told	2
you	2
was	1
was	2
with	2



提纲

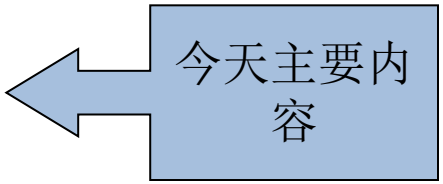
① 信息检索概述

② 倒排索引

③ 布尔查询的处理

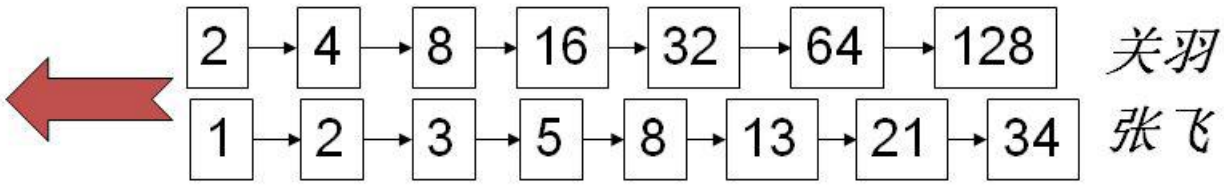
假定索引已经构建好

- 如何利用该索引来处理查询?



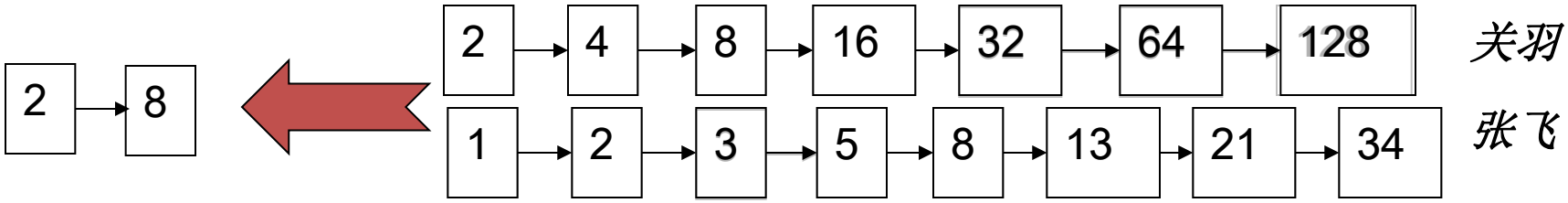
AND查询的处理

- 考虑如下查询（从简单的布尔表达式入手）：
  - 关羽 AND 张飞
  - 在词典中定位 关羽
    - 返回对应倒排记录表(对应的docID)
  - 在词典中定位张飞
    - 再返回对应倒排记录表
  - 合并(Merge)两个倒排记录表，即求交集



合并过程

- 每个倒排记录表都有一个定位指针，两个指针同时从前往后扫描，每次比较当前指针对应倒排记录，然后移动某个或两个指针。合并时间为两个表长之和的线性时间



假定表长分别为x 和y, 那么上述合并算法的复杂度为 O(*x*+*y*)

关键原因: 倒排记录表按照docID排序

布尔检索的优点

- 构建简单，或许是构建IR系统的一种最简单方式
  - 在30多年中是最主要的检索工具
  - 当前许多搜索系统仍然使用布尔检索模型:
    - 电子邮件、文献编目、Mac OS X Spotlight工具

布尔检索例子: www.cnki.net

布尔检索例子: www.cnki.net

全球学术快报

资源类型
- 硕士 (282)
- 期刊 (45)
- 国内会议 (3)
- 博士 (1)

文献来源
- 电子科技大学 (23)
- 北京邮电大学 (19)
- 重庆大学 (10)
- 西安电子科技大学 (9)
- 北京交通大学 (8)

关键词
- Lucene (132)
- 信息检索 (85)
- 搜索引擎 (69)
- 全文检索 (64)
- 本体 (33)

文献类型
- 综述类文献 (1)

检索历史
- 信息检索
- 水控蓄热大棚
- 李永可

检索痕迹 清空

浏览历史
- 非对称水控酿热大棚性能研究

每页显示: 10 20 50

已选文献: 0 清除 | 批量下载 | 导出/参考文献 | 计量可视化分析 | 找到 331 条结果 1/17

	题名	作者	来源	发表时间	数据库	被引	下载	阅读
1	基于Lucene的信息检索的研究与应用	孙西全; 马瑞芳; 李燕灵	情报理论与实践	2006-01-30	期刊	72	722	
2	基于网络爬虫的网站信息采集技术研究	孙骏雄	大连海事大学	2014-09-01	硕士	19	1150	
3	Lucene搜索引擎	周登朋; 谢康林	计算机工程	2007-09-20	期刊	75	1110	
4	一种基于Lucene的Web全文信息检索系统的设计与实现	张晓卫; 朱巧明	计算机与现代化	2006-12-30	期刊	33	447	
5	基于Lucene的英汉跨语言信息检索	陈士杰; 张玥杰	计算机工程	2005-07-05	期刊	44	662	
6	基于Lucene的数据库全文信息检索	王富强; 王青山; 张立朝; 朱浩群; 王锐	测绘科学	2008-05-20	期刊	20	491	
7	基于Lucene的全文检索系统的研究与实现	周锦程; 王丹; 余泉; 张维	计算机技术与发展	2011-03-10	期刊	17	284	
8	基于Lucene的全文检索系统的设计与实现	范蕾	厦门大学	2014-04-01	硕士	11	434	
9	基于知识图谱的搜索引擎技术研究与应用	邵领	电子科技大学	2016-03-28	硕士	1	616	
10	基于Lucene面向主题搜索引擎的研究与设计	姜华	华东师范大学	2007-03-01	硕士	57	2486	
11	基于本体的语义检索原型系统的设计与实现	段寿建	云南师范大学	2008-05-20	硕士	14	521	
12	基于Lucene的面向商业应用的搜索引擎研究与实现	潘亭沥	电子科技大学	2007-04-01	硕士	31	1126	
13	基于Lucene的应用系统内部搜索的研究与设计	张琦玉	南京理工大学	2013-03-01	硕士	8	180	
14	基于Lucene的全文检索系统的研究与应用	苏景春	北京交通大学	2010-05-30	硕士	25	654	
15	基于Lucene的中文自然语言搜索引擎	胡长春	上海交通大学	2009-01-01	硕士	25	1035	
16	垂直搜索引擎的研究与设计	刘运强	计算机应用与软件	2010-07-15	期刊	29	448	

Google支持布尔查询

- 想查关于2011年快女 6进5 比赛的新闻，用布尔表达式怎么构造查询？

- (2011 OR 今年) AND (快乐女声 OR 快女 OR 快乐女生) AND (6进5 OR 六进五 OR 六 AND 进 AND 五)

- 表达式相当复杂，构造困难！
- 不严格的话结果过多，而且很多不相关；非常严格的话结果会很少，漏掉很多结果。

布尔检索的缺点

- 布尔查询构建复杂，不适合普通用户。构建不当，检索结果过多或者过少
- 没有充分利用词项的频率信息
  - 1 vs. 0 次出现
  - 2 vs. 1次出现
  - 3 vs. 2次出现, …
  - 通常出现的越多越好，需要利用词项在文档中的词项频率(term frequency, tf)信息
- 不能对检索结果进行排序