计算语言学

第2讲 词典

刘群
中国科学院计算技术研究所
liuqun@ict.ac.cn

中国科学院研究生院2002~2003学年第二学期课程讲义

词典与词典编撰的研究

- 词典学 lexicology
  - Theory and description of lexical information
- 计算词典学 computational lexicology
  - formal modelling of lexical information
- 词典编撰学 lexicography
  - Construction of dictionaries (databases, handbooks)
- 计算词典编撰学 computational lexicography
  - construction and production of dictionaries using electronic publishing

机读词典与人读词典

- 人读词典（Human Readable Dictionary）
  - 格式不规范
  - 数据完整性和一致性不好
  - 非结构化
- 机读词典（Machine Readable Dictionary）
  - 格式规范
  - 数据完整性和一致性较好
  - 结构化

人读词典（demo）

- 金山词霸

story
中古英语 storie <古法语 estoire <拉丁语 historia
n
-ries
(1)故事，小说；传闻； 轶事
Please read us a story!
请给我们读个故事！
(2) 谎话，假话
(3)（书籍、电影、戏剧等的）情节
(4)（报刊、杂志文章的）素材，题材

机读词典的分类

- 按信息类型分类
  - 语法词典
  - 语义词典（包括同义词典）
  - 双语词典
  - ……
- 按领域分类
  - 通用词典
  - 专业词典（术语词典）
  - 专名词典
  - ……

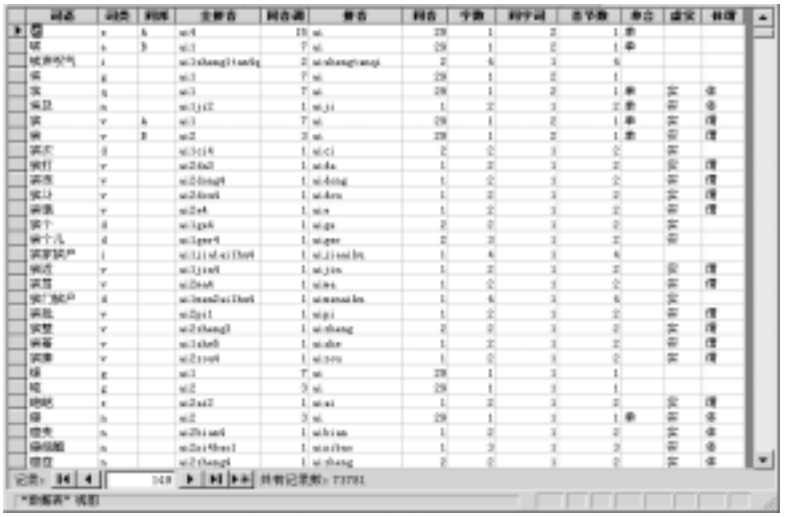
汉语语法信息词典

- 开发单位：北京大学计算语言学研究所
- 参考文献：
  - 俞士汶 等（1998）《现代汉语语法信息词典详解》，清华大学出版社、广西科学技术出版社1998年版。
- 规模：7万多词条
  - 总库
  - 词性库

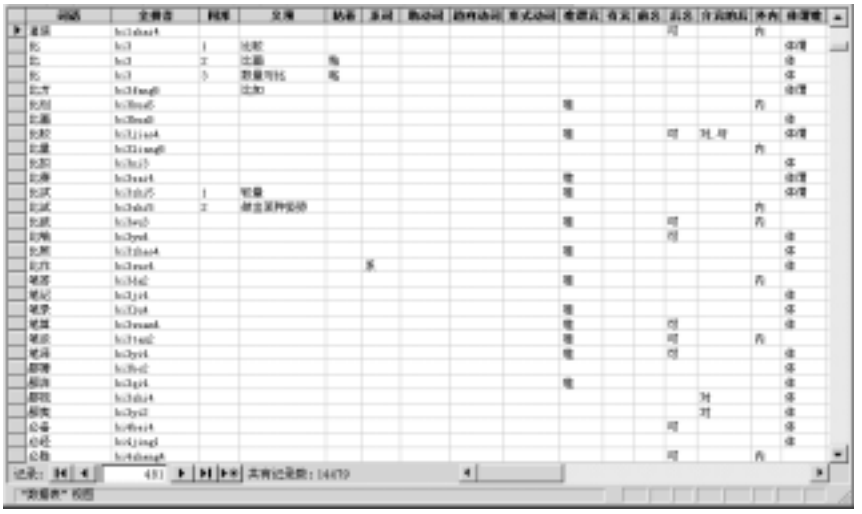
    名词 时间词 处所词 方位词 数词 量词 区别词 代词 动词 形容词 状态词 副词 介词 连词 助词 语气词 前接成分 后接成分 成语 简称略语 习用语 语素 标点符号
  - 词性分库

    动词 代词

汉语语法信息词典·总库



汉语语法信息词典·动词库



汉语语法信息词典·谓宾动词分库

新华社词语数据库

全库分为中文和外文两个大类，主要包括中文新闻库、经济信息库、证券库、人物库、组织机构库、专题资料库等中文数据库，还包括Xinhua News Bulletin 、Who’s Who in China等英文数据库。共有28个库100多个子库，数据量达80多亿汉字，并以日均150万汉字的速度增长。

新华社词语数据库·国际组织

- “２０００年问题”联合委员会 /joint year 2000 council/ International
- “４·１９”运动 /movement april 19/ Colombia
- “阿尔法６６” /"alpha 66"/ Cuba
- “俄罗斯地区”社会联盟 /regions of russia group/ Russia
- “法中－２０００年”协会 /france-china association for the year 2000/ France
- “繁荣”党 /prosperity/ Russia
- “光明的日本”国会议员联盟 /parliamentary union for a bright japan/ Japan
- “基地”组织 /al qaeda/ Saudi Arabia
- 《财富》杂志 /fortune/ USA
- 《朝日新闻》 /asahi shimbun/ Japan
- 国际献血组织联合会 /international federation of blood donor organizations/ International
- 国际宪法学协会 /international association of constitutional law/ International
- 国际香料集团 /international spice group/ International
- 经济和外贸部 /ministry of economy and external trade of syria/ Syria
- 经济和外贸部 /ministry of economy and foreign trade of egypt/ Egypt

新华社词语数据库·人名

依	阿布申科		abushenko		
捷	阿布希诺夫		abuzinov		
土	阿布什卡		abushka		
土	阿布什奥卢		abuoglu		
阿拉伯	阿布－萨马赫		abuzmah		
意	阿布西		abuzi		
土	阿布特		abut		
依	阿布塔利布		abutalib		
依	阿布塔利普		abutalip		
依	阿布泰德		abutideo		
依	阿布亚		abutia		
依	阿布扎利诺夫		abutzlinov		
日	阿佛		abutsu		处
依	阿布耶夫		abuyev		
土	阿布兹		abuz		
依	阿布扎罗夫		abuzarov		
阿拉伯	阿布－扎伊德		abuzaid		
土	阿布泽尔		abuzer		
依	阿布津		abuzin		
依	阿布佐夫		abuzov		
依	阿布佐罗夫		abuzyarov		
扎	阿布瓦萨萨		abuzizza		
英	阿比		abr		
阿拉伯	阿比阿德		abyad		其它解读：abi-a
阿拉伯	阿比阿德		abyadh		
阿拉伯	阿比安		abyan		

知网（Hownet）1

- 作者：董振东 董强
- 网站：http://www.keenage.com
- 概念描述举例

  NO.=017144
  W_C=打
  G_C=V
  E_C=~网球，~牌，~秋千，~太极，球~得很棒
  W_E=play
  G_E=V
  E_E=
  DEF=exercise|锻练,sport|体育
- 其中DEF是核心，采用特定的“知识描述语言”

知网（Hownet）2

打	017144	exercise|锻练,sport|体育
男人	059349	human|人,family|家,male|男
高兴	029542	aValue|属性值,circumstances|境况,happy|福,desired|良
生日	072280	time|时间,day|日,@ComeToWorld|问世,$congratulate|祝贺
写信	089834	write|写,ContentProduct=letter|信件
北京	003815	place|地方,capital|国都,ProperName|专,(China|中国)
爱好者	000363	human|人,*FondOf|喜欢,#WhileAway|消闲
必须	004932	{modality|语气}
串	015204	NounUnit|名量,&(grape|葡萄),&(key|钥匙)
从良	016251	cease|停做,content=(prostitution|卖淫)
打对折	017317	subtract|削减,patient=price|价格,commercial|商,(range|幅度=50%)
儿童基金会	024083	part|部件,%institution|机构,politics|政,#young|幼,#fund|资金,(institution|机构=UN|联合国)

知网（Hownet）3

- 义原总数：1500多个
- 义原分类：共8类
  - 基本义原
    - 事件、实体、次要特征
    - 属性、属性值、数量、数量值
  - 语法义原：描述语法特征，如POS
    - 语法
  - 关系义原：描述意义关系，类似于格关系
    - 动态角色
    - 动态属性

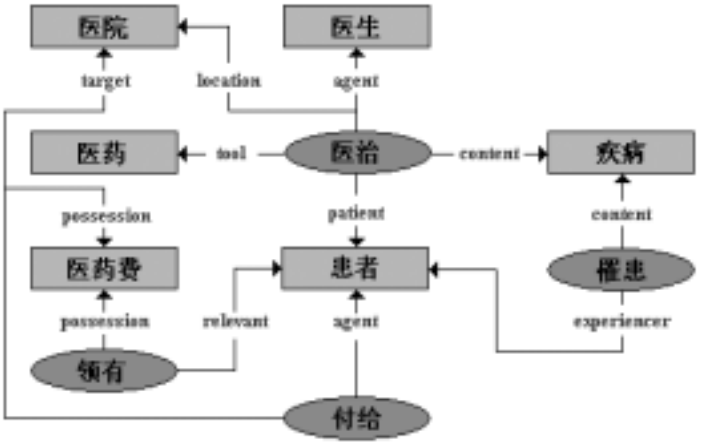
知网（Hownet）4

义原的上下位关系构成树结构

- entity|实体
  - ├ thing|万物
  - … ├ physical|物质
  - … ├ animate|生物
  - … ├ AnimalHuman|动物
  - … ├ human|人
  - │ └ humanized|拟人
  - └ animal|兽
  - ├ beast|走兽
  - …

知网（Hownet）5

知网中的关系



同义词词林 1

- 梅家驹 等，1983，上海辞书出版社
- 为克服写作和翻译时的词穷现象而编写
- 目前广泛应用于自然语言处理中
- 收词近7万（按义项统计）
- 按义项编排
  - 12大类
  - 94中类
  - 1428小类
  - 3925词群
- 词群内部的词是同义词
- 大类、中类、小类之间不一定是上下位关系（有些是领域）

同义词词林 2

Ag100101	旅客
Ag100101	客人
Ag100101	旅人
Ag100101	客子
Ag100101	客行子
Ag100101	游子
Ag100101	行人
Ag100101	行者
Ag100101	行旅
Ag100101	行客
Ag100101	行子
Ag100101	征人
Ag100101	征夫
Ag100101	征客
Ag100101	羁客
Ag100101	羁旅
Ag100101	客
Ag100102	过路人
Ag100102	过客
Ag100103	游人
Ag100103	游客
Ag100103	游者
Ag100103	旅游者
Ag100103	观光者

大类：A

中类：g

小类：10

词群：01

最小同义词集：01，02，03

WordNet 1

- 网址：
  - http://www.cogsci.princeton.edu/~wn/
- 开发单位：
  - 普林斯顿大学心理语言学实验室
  - 初衷是作为研究人类词汇记忆的心理语言学成果
  - 在自然语言处理中得到广泛的应用
- 免费的在线词汇数据库
- 世界很多语种都开发了相应的版本
  - 各种欧洲语言：EuroNet
  - 汉语：CCD（Chinese Concept Dictioanry）

WordNet 2

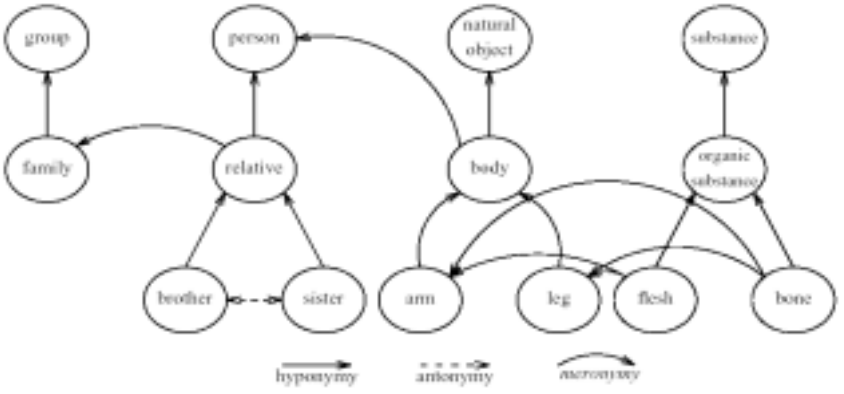
- 同义词集 Synset
  - 用一组同义词的集合Synset来表示一个概念
  - 每一个概念有一段描述性的说明
- 关系
  - 上下位关系（hyponymy，troponymy）
  - 同义反义关系（synonymy，antonymy）
  - 部分整体关系（entailment，meronymy）
  - ……

Wordnet 3

- 规模
  - 名词：80,000 words, 60,000 synsets
  - 形容词：16,000 synsets
  - 动词：11,500 synsets
  - 还在不断发展之中

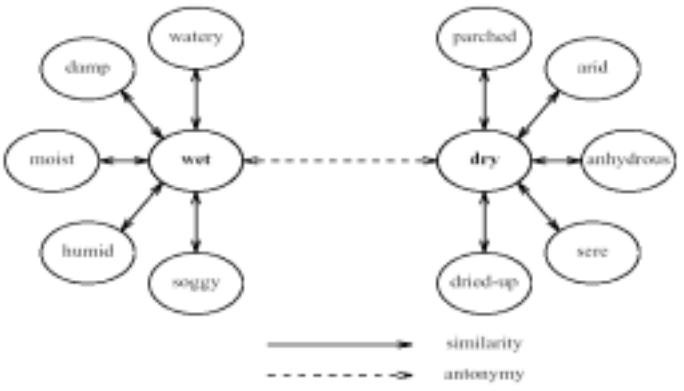
WordNet 4

名词概念的组织：

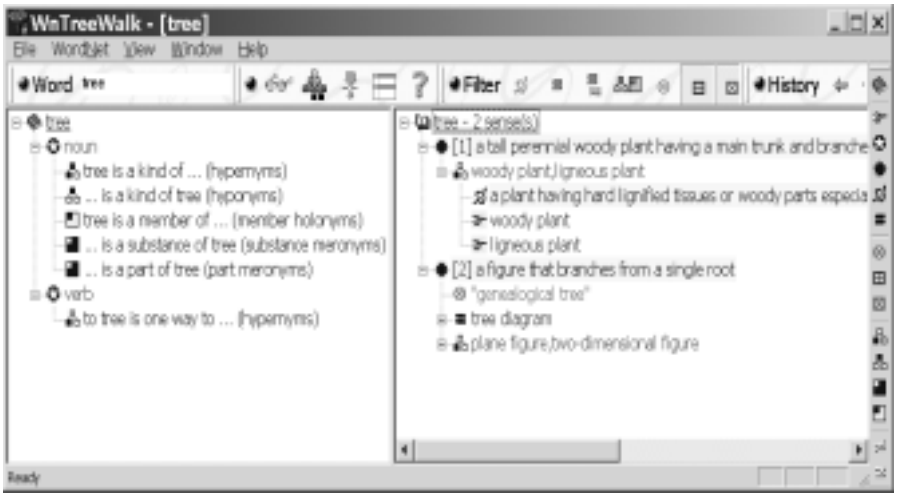


WordNet 5

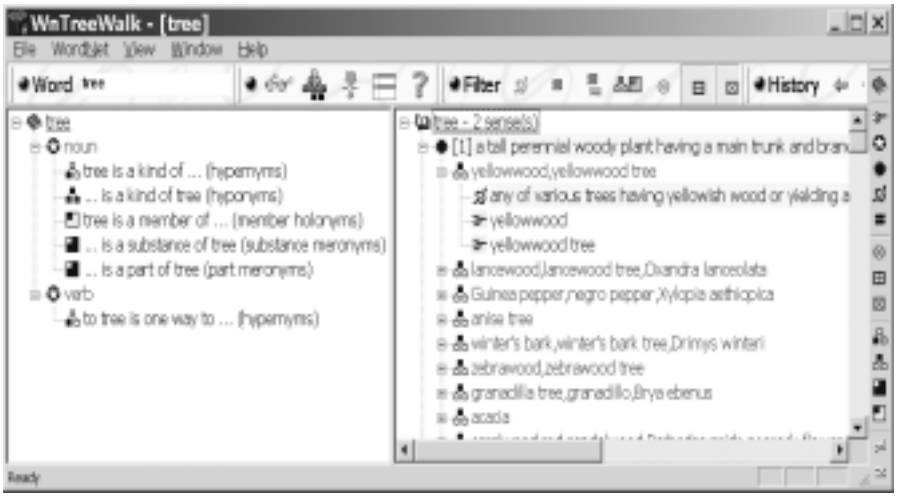
形容词概念的组织：



WordNet 6



WordNet 7



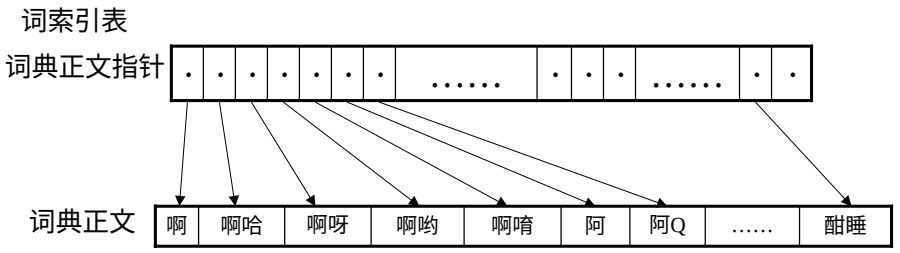
词典检索算法 1

- 词典检索算法的性能评价
  - 时间复杂度
  - 空间复杂度
  - 检索方式
    - 直接用词语检索
    - 检索句子中某个位置开始的所有词
    - 检索句子中某个位置开始的最长词
    - 模糊检索
    - ……
  - 增量式索引

词典检索算法 2

- 两个问题
  - 索引结构
  - 查找算法
- 一种索引结构可以对应不同的查找算法

词典顺序索引

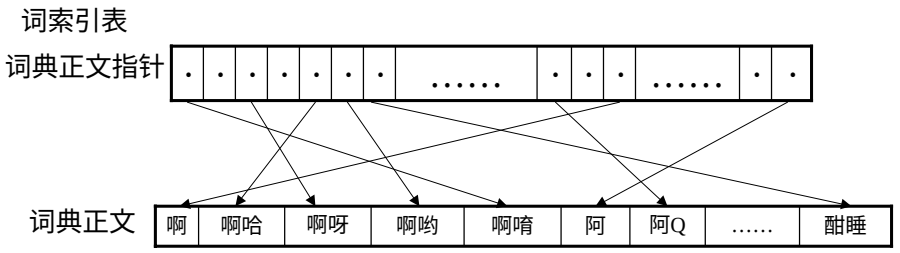


- 索引结构简单，占用空间小
- 不能实现增量式索引：每增加一个词需重新排序

词典顺序索引的查找算法

- 整词二分查找
  - 时间复杂度$O(\log_2 N)$
  - 无法按前缀查找
- 改进的整词二分查找
  - 时间复杂度$O(\log_2 N)$
  - 可以实现按前缀查找

词典散列索引



- 索引结构简单，占用空间小（比顺序索引稍大）
- 可以实现增量式索引

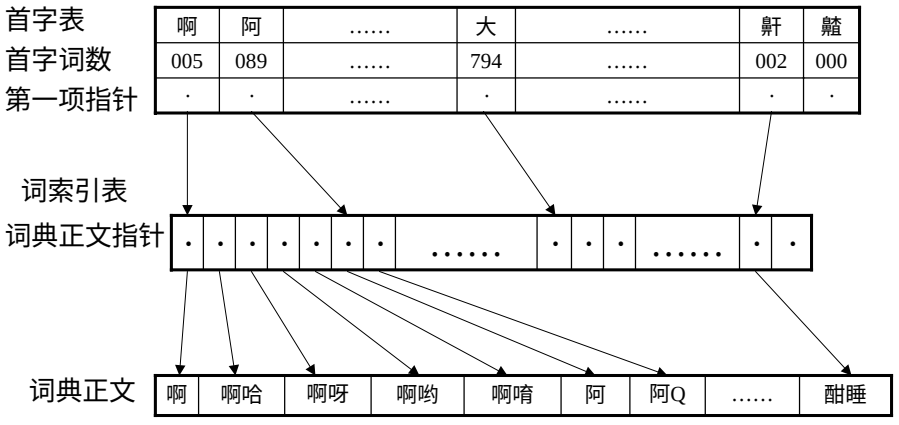
词典散列索引的检索算法

- 利用散列（hash）函数直接定位
- 效率高：常数
- 不能按前缀查找
- 冲突的解决
  - 使用冲突队列
  - 使用再散列
- 散列函数（hash）的选择
- 算法改进：逐词散列，可以实现按前缀查找

词典分级索引

- 将词语分成若干部分，为每一部分分别建立索引
- 在分级索引中，每一级索引都可以采用各种不同的索引和查找算法
- 对于汉语而言，第一级索引一般使用词语的首字，所以又常称为首字索引
- 汉语的首字数量有限，可以使用直接定位法，效率最高，空间也不大

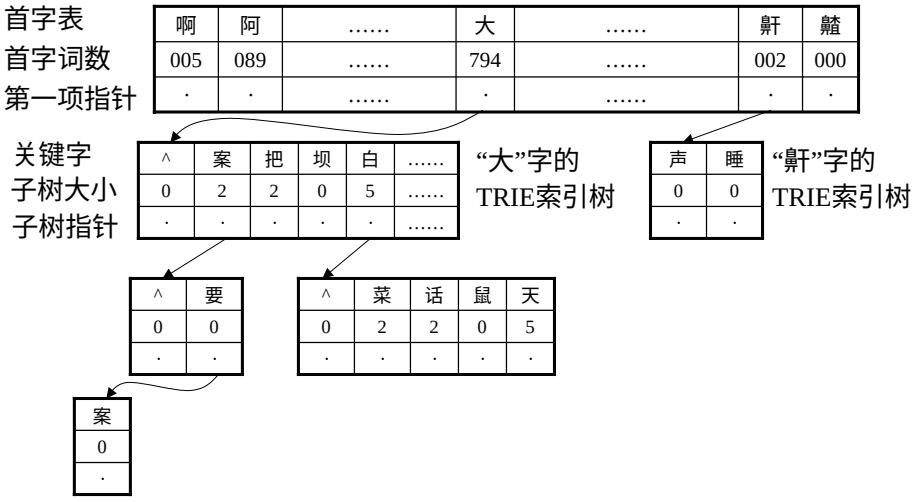
汉语词典按首字顺序索引



首字二分检索 2

- 时间复杂度：$O(\log_2 N)$
- 空间复杂度：O(N)
- 可以按前缀查找
- 不能增量式索引：每次要重新排序

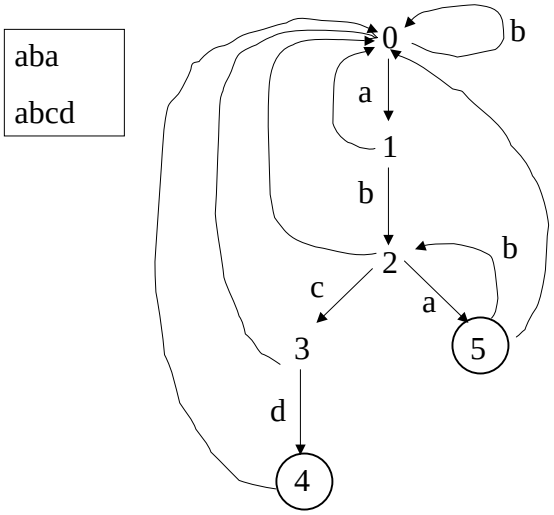
汉语词典TRIE树索引



AC算法 1

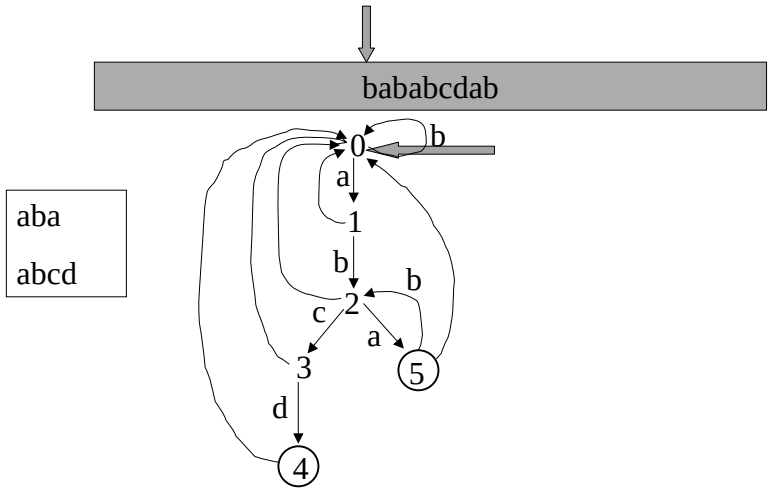
- 问题
  - 假设词典中有两个词：aba，abcd
  - 考虑输入串：bababcdab
  - 如何迅速找出输入串中词典词的所有出现？
- 简单解决办法
  - 逐字查词典：效率太低
- AC算法
  - 将词典构造成一个自动机，一次扫描完成

AC算法 2

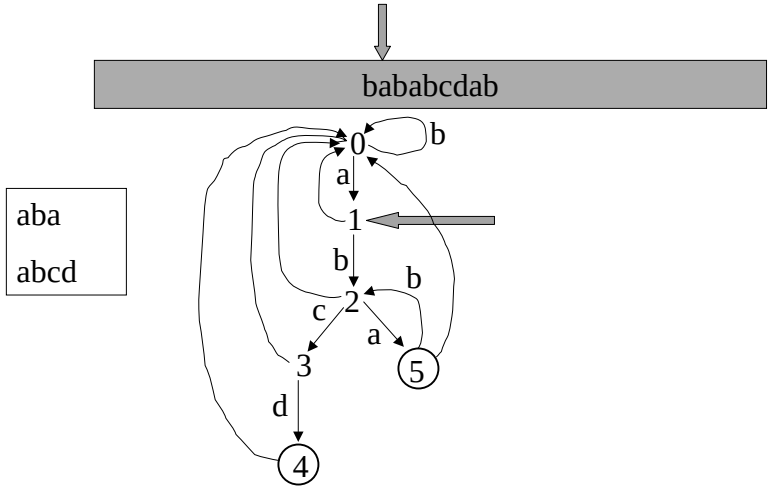


空边对应所有
不匹配的字符

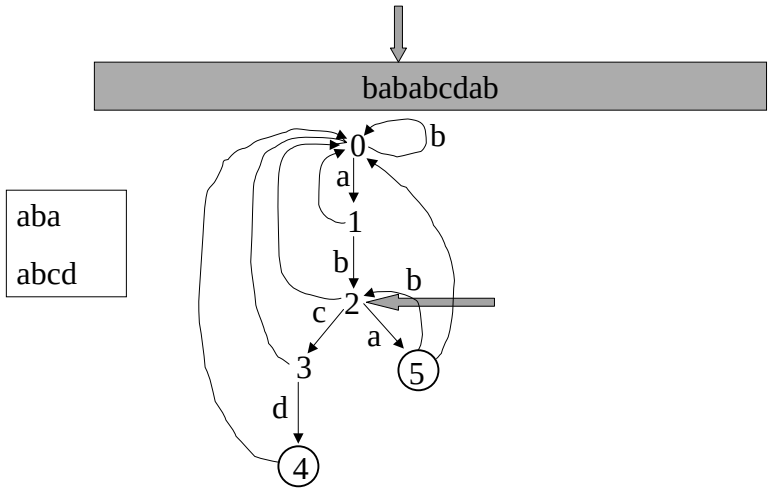
AC算法 3


bababcdab
aba
abcd
0
1
2
3
4
5
a
b
c
d
b
b
a
b

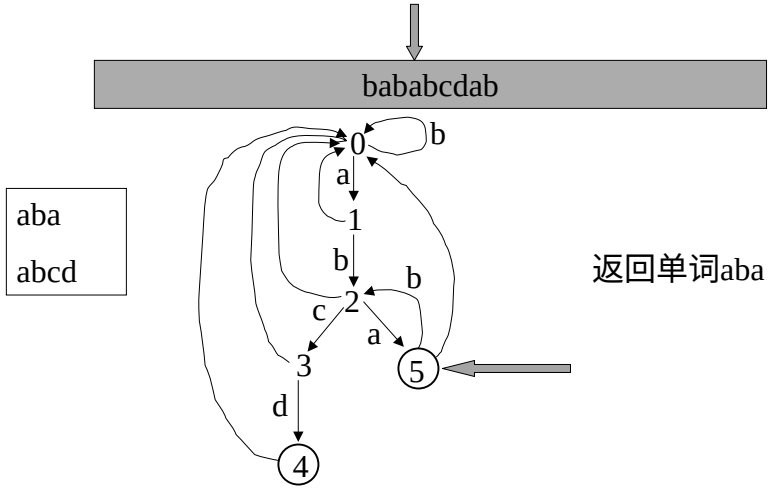
AC算法 4


bababcdab
aba
abcd
0
1
2
3
4
5
a
b
c
d
b
b
a
b

AC算法 5

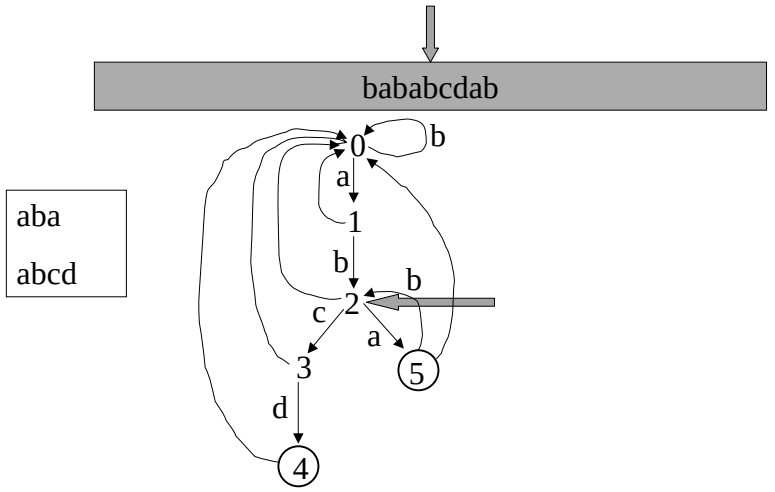


AC算法 5

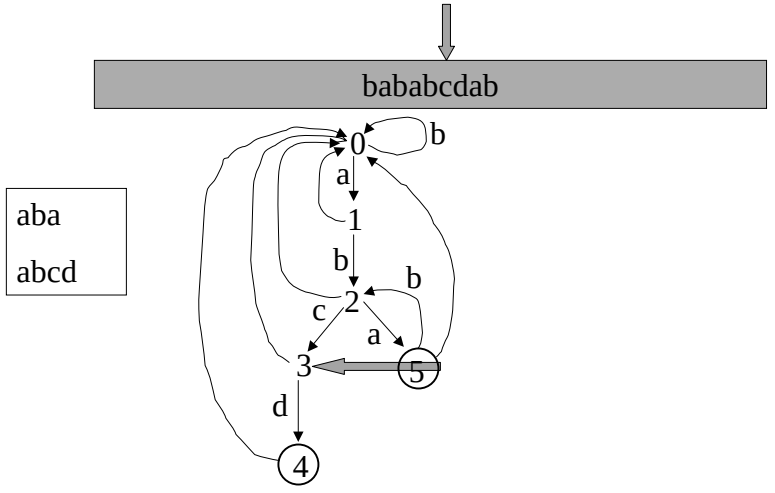


返回单词aba

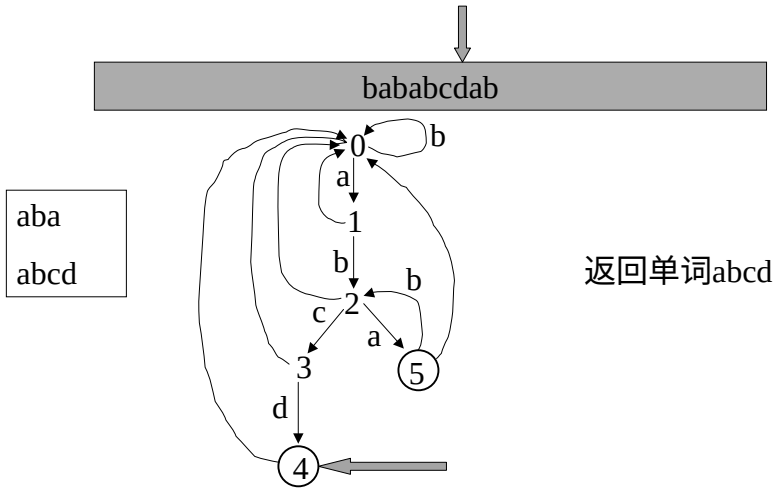
AC算法 6


bababcdab
aba
abcd
0
1
2
3
4
5
a
b
b
b
c
a
d
b

AC算法 7


bababcdab
aba
abcd
0
1
2
3
4
5
a
b
b
b
c
a
d
b

AC算法 8

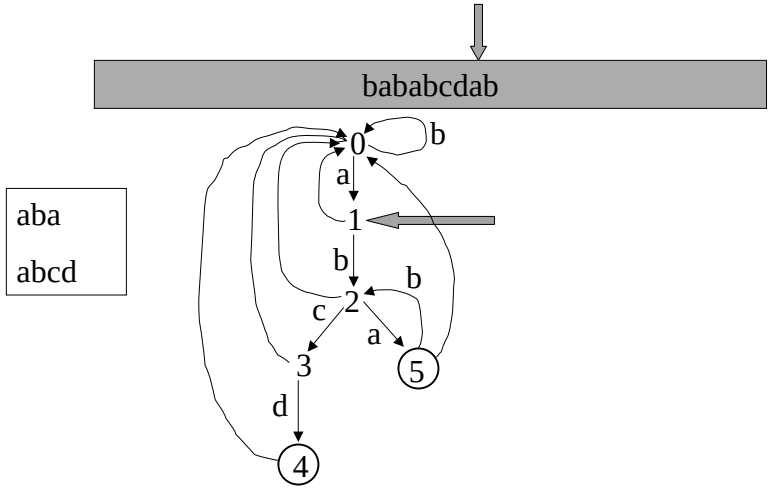


aba

abcd

返回单词abcd

AC算法 9



aba

abcd

重复子串识别

目标：识别出文本中所有出现两次以上的子串

据香港《文汇报》报道，北京的台湾问题专家李家泉受访时指出，台北、高雄两市市长选举，尽管蓝、绿两政治势力进行了激烈的斗争，但“北蓝南绿”的政治格局未被打破，由此可以预见，未来一段时间内两岸关系的改善很难有突破。李家泉指出，此次北高两市选举在两个大背景下进行，一是民进党执政两年来政绩相当差，自身危机感非常强；二是距离2004年“大选”只有一年多时间，两派都格外重视此次交锋，对泛绿阵营来说是政权保卫战，而对泛蓝阵营来说则是夺权演习战。因此可以看到斗争形势相当严峻而激烈。

逐词递增算法 1

- 首先记录所有二字串的出现位置和频度
- 删除只出现一次的二字串记录
- 对于出现两次以上的二字串，向后扩展一个字，记录所有三字串的出现位置和频度
- 删除只出现一次的三字串
- 重复上述过程，直到不再有重复串为止

逐词递增算法 2

- 性能
  - 最坏情况：前后两段文字完全相同
  - 在最坏情况下，时间复杂度：$O(n^2)$
- 算法改进
  - 时间复杂度可以达到O(n)？
- 演示

基于重复子串的新词发现

- 对于《人民日报》2002年和2001年语料分别进行重复子串识别
- 用2002年的重复子串集合减去2001年的重复子串集合
- 2002年出现词数大于20的词语而2001年没有出现过的重复子串：1005个
- Top 10

十 六大 精神	1289	中 共 十 六大	342
学习 贯彻 十 六大 精神	238	核查 人员	223
干部 任用 条例	220	建设 中国 特色 社会主义	194
一 边 一 国	189	贯彻 十 六大 精神	156
胡锦涛 当选 为 中共中央 总书记	155	军品 出口	151

复习思考题

- 如果有一部人读的双语词典，你如何将它转换成机读词典？
- 如何利用语义词典进行词语相似度计算？
- 请实现逐字散列的词典检索算法。
- 汉语词典和英语词典在实现上有什么不同？
- 请查找文献，看看如何寻找一个好的散列函数。