计算语言学

第2讲·上 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2007~2008学年第二学期课程讲义

内容提要：词典

- 词典与词典编纂
- 人读词典与机读词典
- 词典检索算法
- 应用之一：汉语新词语检测
- 应用之二：词汇语义相似度计算

词典与词典编纂的研究

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

内容提要：词典

- 词典与词典编纂
- 人读词典与机读词典
- 词典检索算法
- 应用之一：汉语新词语检测
- 应用之二：词汇语义相似度计算

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（Demo）

英文词语	story		
中古英语	storie		
古法语	estoire		
拉丁语	historia		
词性	n		
复数	-ries		
释义	中文解释	故事	
	中文解释	小说	
	中文解释	传闻	
	中文解释	逸事	
	例句	中文	Please read us a story!
		英文	请给我们读个故事！
释义	中文解释	谎话	
	中文解释	假话	
释义	中文解释	情节	
	限制条件	书籍、电影、戏剧等	
释义	中文解释	素材	
	中文解释	题材	
	限制条件	报刊、杂志文章	

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

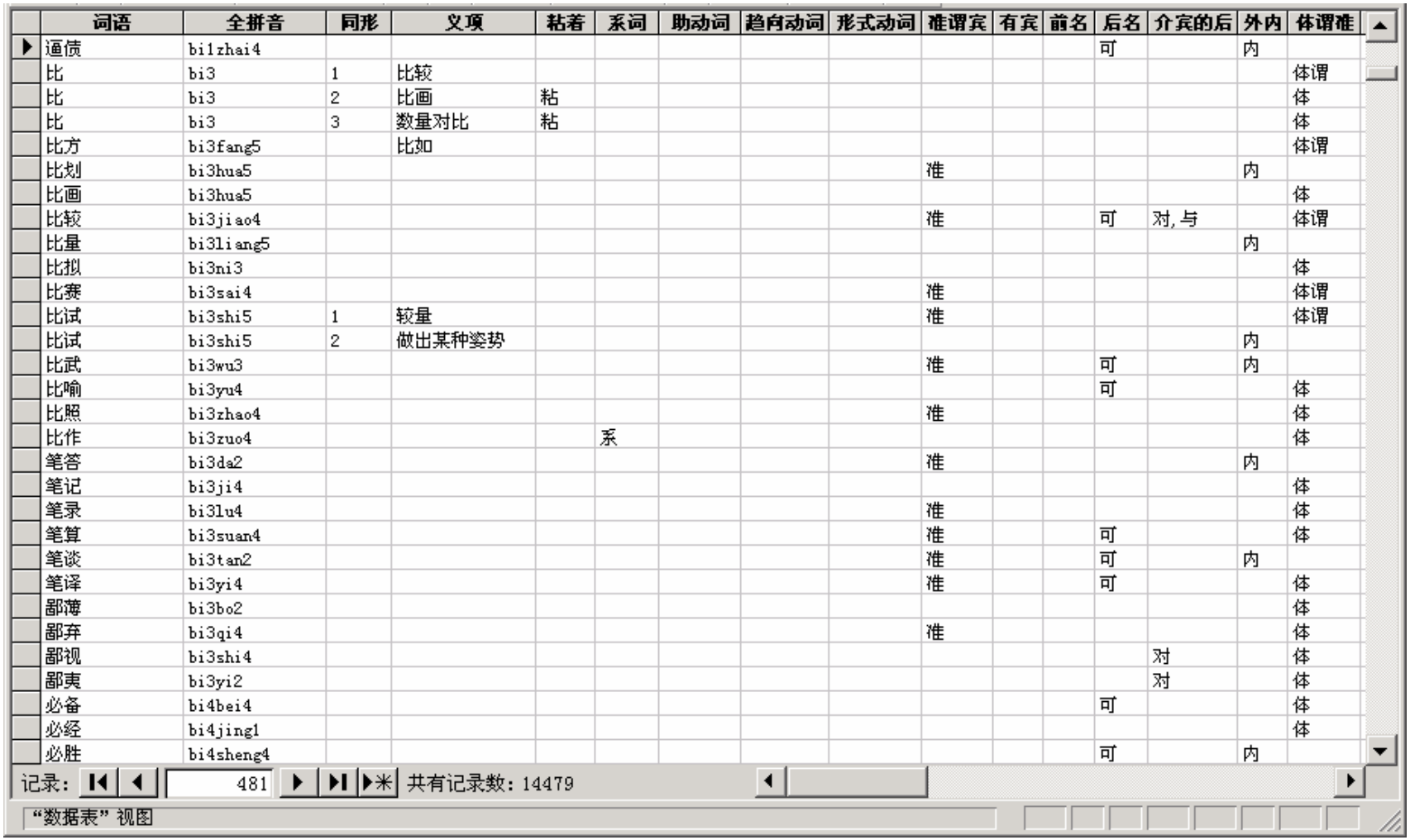
汉语语法信息词典·总库

词语	词类	同形	全拼音	同音调	拼音	同音	字数	同字词	音节数	单合	虚实	体谓
嗳	e	A	ai4	15	ai	29	1	2	1	单		
唉	e	B	ai1	7	ai	29	1	2	1	单		
唉声叹气	i		ai1sheng1tan4q	2	aishengtanqi	2	4	1	4			
埃	g		ai1	7	ai	29	1	2	1			
埃	q		ai1	7	ai	29	1	2	1	单	实	体
埃及	n		ai1ji2	1	aiji	1	2	1	2	单	实	体
挨	v	A	ai1	7	ai	29	1	2	1	单	实	谓
挨	v	B	ai2	3	ai	29	1	2	1	单	实	谓
挨次	d		ai1ci4	1	aici	2	2	1	2		实	
挨打	v		ai2da3	1	aida	1	2	1	2		实	谓
挨冻	v		ai2dong4	1	aidong	1	2	1	2		实	谓
挨斗	v		ai2dou4	1	aidou	1	2	1	2		实	谓
挨饿	v		ai2e4	1	aie	1	2	1	2		实	谓
挨个	d		ai1ge4	1	aige	2	2	1	2		实	
挨个儿	d		ai1ger4	1	aiger	2	3	1	2		实	
挨家挨户	i		ai1jialai1hu4	1	aijiaaihu	1	4	1	4			
挨近	v		ai1jin4	1	aijin	1	2	1	2		实	谓
挨骂	v		ai2ma4	1	aima	1	2	1	2		实	谓
挨门挨户	d		ai1men2ai1hu4	1	aimenaihu	1	4	1	4		实	
挨批	v		ai2pi1	1	aipi	1	2	1	2		实	谓
挨整	v		ai2zheng3	1	aizheng	2	2	1	2		实	谓
挨着	v		ai1zhe5	1	aizhe	1	2	1	2		实	谓
挨揍	v		ai2zou4	1	aizou	1	2	1	2		实	谓
锿	g		ai1	7	ai	29	1	1	1			
皑	g		ai2	3	ai	29	1	1	1			
皑皑	z		ai2ai2	1	aiai	1	2	1	2		实	谓
癌	n		ai2	3	ai	29	1	1	1	单	实	体
癌变	n		ai2bian4	1	aibian	1	2	1	2		实	体
癌细胞	n		ai2xi4bao1	1	aixibao	1	3	1	3		实	体
癌症	n		ai2zheng4	1	aizheng	2	2	1	2		实	体

记录: 149 共有记录数: 73781

"数据表"视图

汉语语法信息词典·动词库



词语	全拼音	同形	义项	粘着	系词	助动词	趋向动词	形式动词	准谓宾	有宾	前名	后名	介宾的后	外内	体谓准
逼债	bi1zhai4											可		内	
比	bi3	1	比较												体谓
比	bi3	2	比画	粘											体
比	bi3	3	数量对比	粘											体
比方	bi3fang5		比如												体谓
比划	bi3hua5								准					内	
比画	bi3hua5														体
比较	bi3jiao4								准			可	对,与		体谓
比量	bi3liang5													内	
比拟	bi3ni3														体
比赛	bi3sai4								准						体谓
比试	bi3shi5	1	较量						准						体谓
比试	bi3shi5	2	做出某种姿势											内	
比武	bi3wu3								准			可		内	
比喻	bi3yu4											可			体
比照	bi3zhao4								准						体
比作	bi3zuo4				系										体
笔答	bi3da2								准					内	
笔记	bi3ji4														体
笔录	bi3lu4								准						体
笔算	bi3suan4								准			可			体
笔谈	bi3tan2								准			可		内	
笔译	bi3yi4								准			可			体
鄙薄	bi3bo2														体
鄙弃	bi3qi4								准						体
鄙视	bi3shi4												对		体
鄙夷	bi3yi2												对		体
必备	bi4bei4											可			体
必经	bi4jing1														体
必胜	bi4sheng4											可		内	

记录: 481 共有记录数: 14479

"数据表"视图

汉语语法信息词典·谓宾动词分库

词语	同形	动宾	形宾	状宾	小句宾	疑问	备注
回想		动			句		
回忆		动			句	问	
悔恨					句		
汇报		动			句	问	
会	B2	动					~英语/~讲三种方言
活像					句		
获悉					句		
获准		动					
讥嘲					句		
讥诮					句		
讥笑					句		
即		动					
亟待		动			句		~我们解决
亟盼		动			句		~你早日回来
亟需		动			句		~政府发救济款
急需		动			句		~救治病人
急于		动					
嫉妒					句		
计划		动					
计较	1	动			句	问	~合算不合算/~个人利益/不~自己是否得到
计算	1	动				问	~年龄多大/~下一步怎么走/~符号
计算	2	动			句	问	
记		动			句		
记得					句	问	
记恨					句		~老李欠账不还
记录		动			句	问	
记住		动			句	问	
纪录					句		
纪念					句		
忌		动	形		句		晚上~一个人出门/~吃油性大的食品

记录: 389 共有记录数: 1313

"数据表"视图

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

俄	阿布申科		abushenko		
捷	阿布希诺夫		abusinov		
土	阿布什卡		abuska		
土	阿布什奥卢		abusoglu		
阿拉伯	阿布－萨马赫		abussamah		
意	阿布西		abussi		
土	阿布特		abut		
俄	阿布塔利布		abutalib		
俄	阿布塔利普		abutalip		
俄	阿布季泽		abutidze		
俄	阿布京		abutin		
俄	阿布托利诺夫		abutolinov		
日	阿佛		abutsu		姓
俄	阿布耶夫		abuyev		
土	阿布兹		abuz		
俄	阿布扎罗夫		abuzarov		
阿拉伯	阿布－扎伊德		abuzeid		
土	阿布泽尔		abuzer		
俄	阿布津		abuzin		
俄	阿布佐夫		abuzov		
俄	阿布贾罗夫		abuzyarov		
扎	阿布瓦伊萨萨		abwaisasa		
英	阿比		aby		
阿拉伯	阿比阿德		abyad		其它拼法：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
- 最新版本：2005版（子集MiniHownet免费）
- 网站：http://www.keenage.com
- 概念描述举例：“打”（022303～022357，55个概念）
  - NO.=022331
  - W_C=打
  - G_C=V [da3]
  - E_C=~球，~网球，~篮球，~羽毛球，~牌，~扑克，~麻将，~秋千，~太极拳，球~得很棒
  - W_E=play
  - G_E=V
  - E_E=
  - DEF={exercise|锻炼:domain={sport|体育}}
- 其中DEF是核心，采用特定的“知识描述语言”

知网（Hownet）(2a)

打	play	{exercise|锻炼:domain={sport|体育}}
男人	husband	{human|人:belong={family|家庭},modifier={male|男}{spouse|配偶}}
高兴	cheerful	{joyful|喜悦}
生日	birthday	{time|时间:TimeSect={day|日},{ComeToWorld|问世:time={~}}}
写信	write a letter	{compile|编辑:ContentProduct={letter|信件}}
北京	Beijing	DEF={place|地方:PlaceSect={capital|国都},belong="China|中国",modifier={ProperName|专}}
爱好者	amateur	{human|人:{FondOf|喜欢:experiencer={~}}}

知网（Hownet）(2b)

必须	must	{FuncWord|功能词:comment={?}}
串	cluster	{NounUnit|名量:host={inanimate|无生物}}
从良	get married and start a new life	{cease|停做:content={affairs|事务:modifier={lascivious|淫}{unlawful|非法}}}
打对折	offer a 50% discount	{subtract|削减:domain={commerce|商业},patient={Price|价格}}
儿童基金会	UNICEF	{part|部件:domain={economy|经济},whole={institution|机构:belong="UN|联合国",domain={politics|政},modifier={ProperName|专}{international|国际}}}

知网（Hownet）(3)

- 义原总数：2360个(不考虑专有名词、反义、对义)
- 义原分类：共11类
  - Antonym（反义）484×2（重复）
  - Attribute（属性）247
  - AttributeValue（属性值）889
  - Converse（对义）224×29（重复）
  - Entity（实体）151
  - Event（事件）812
  - EventRoleAndFeatures（事件角色与特征）113
  - ProperNoun（专有名词）248
  - SecondaryFeatures（次要特征）121
  - Sigh（符号）6
  - Syntax（句法）21

知网（Hownet）(4)

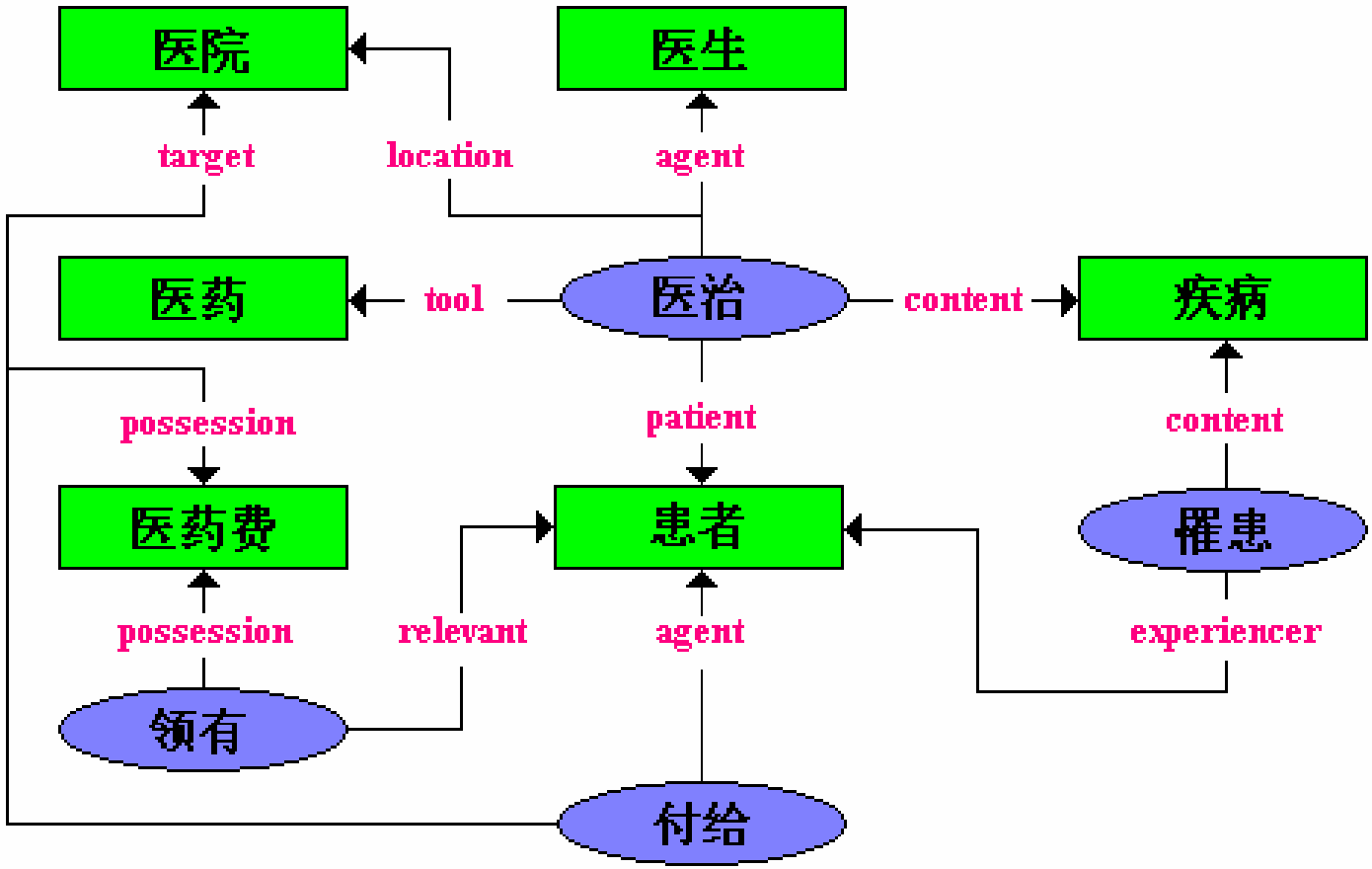
义原的上下位关系构成树结构

- 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)

cate	word
Aa010101	人
Aa010101	士
Aa010101	人物
Aa010101	人士
Aa010101	人氏
Aa010101	人选
Aa010102	人类
Aa010102	噍类
Aa010102	生人
Aa010102	横目
Aa010102	圆颅方趾
Aa010102	方趾圆颅
Aa010103	人手
Aa010103	人员
Aa010103	人口
Aa010103	人丁
Aa010103	口
Aa010103	丁口
Aa010103	食指
Aa010104	劳力
Aa010104	劳动力
Aa010105	匹夫
Aa010105	个人

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)

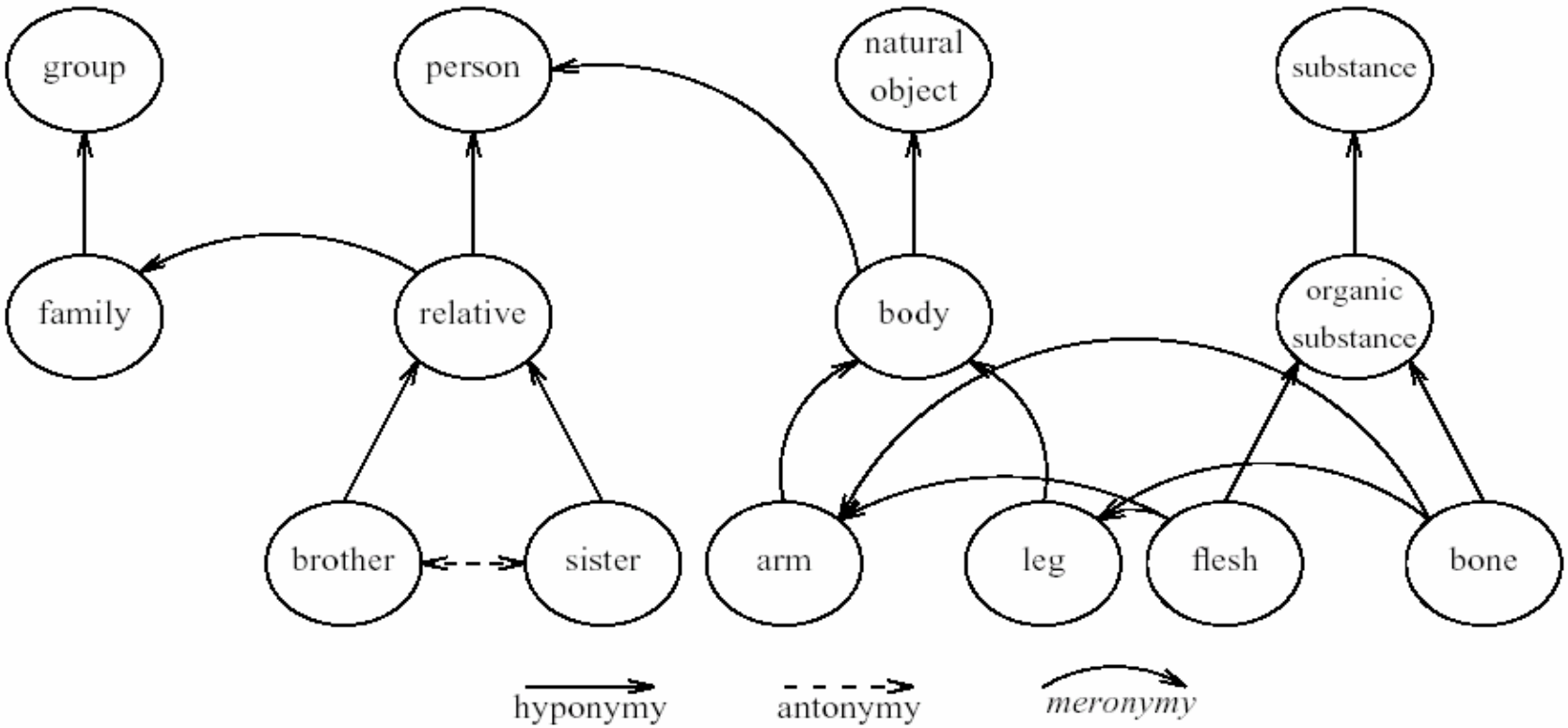
- 同义词集 Synset
  - 用一组同义词的集合Synset来表示一个概念
  - 每一个概念有一段描述性的说明
- 关系
  - 上下位关系（hyponymy，troponymy）
  - 同义反义关系（synonymy，antonymy）
  - 部分整体关系（entailment，meronymy）
  - ……

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

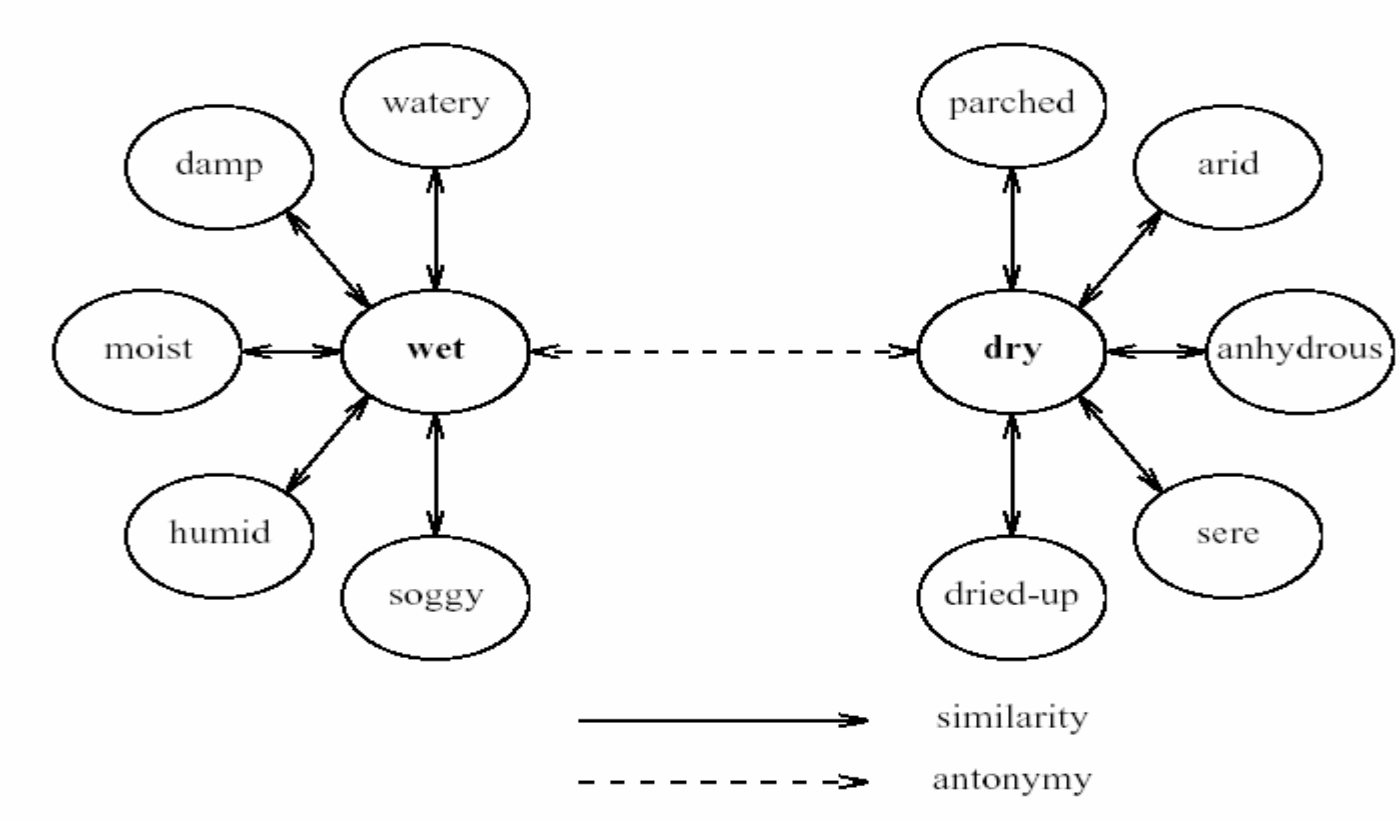
WordNet (4)

名词概念的组织：

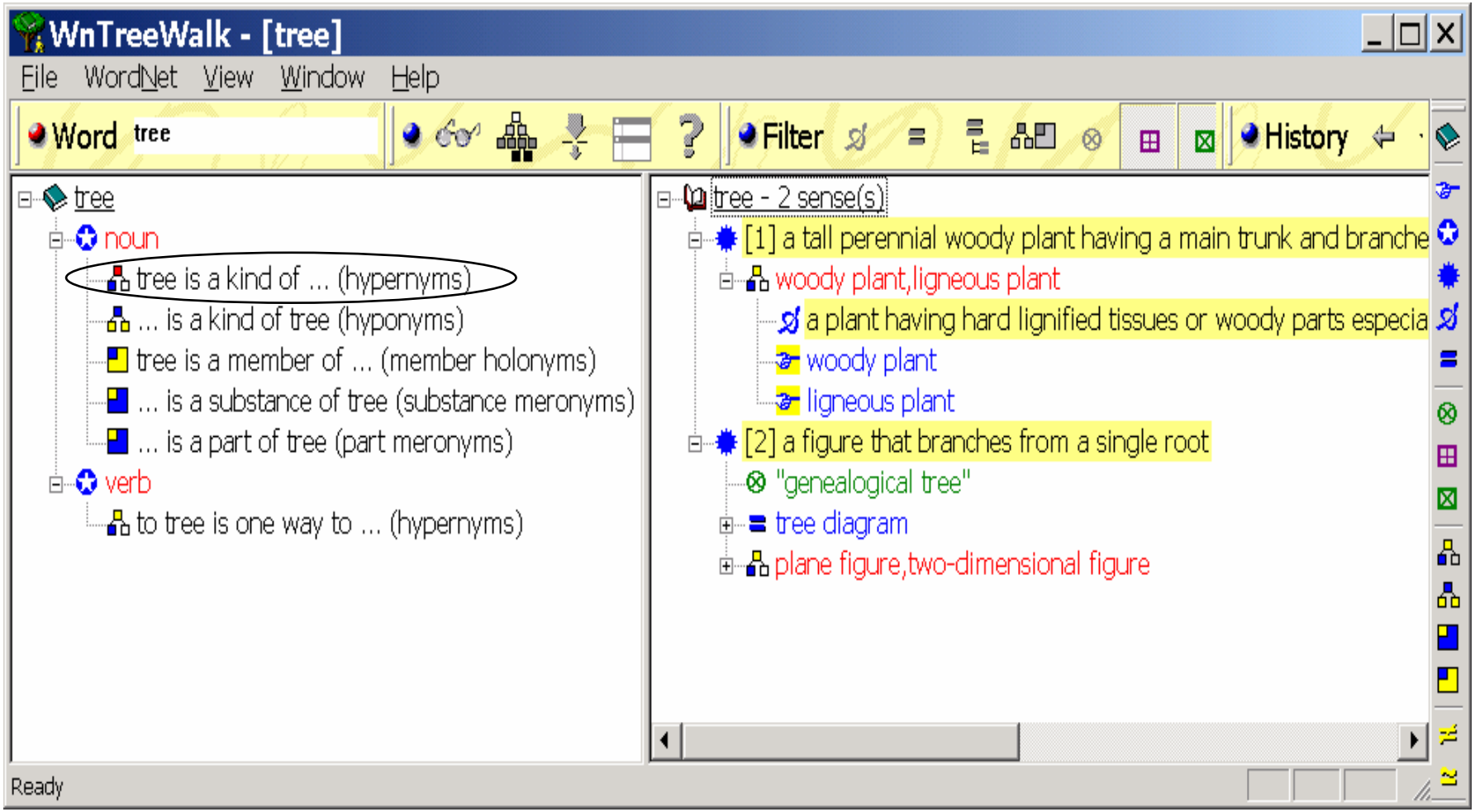


WordNet (5)

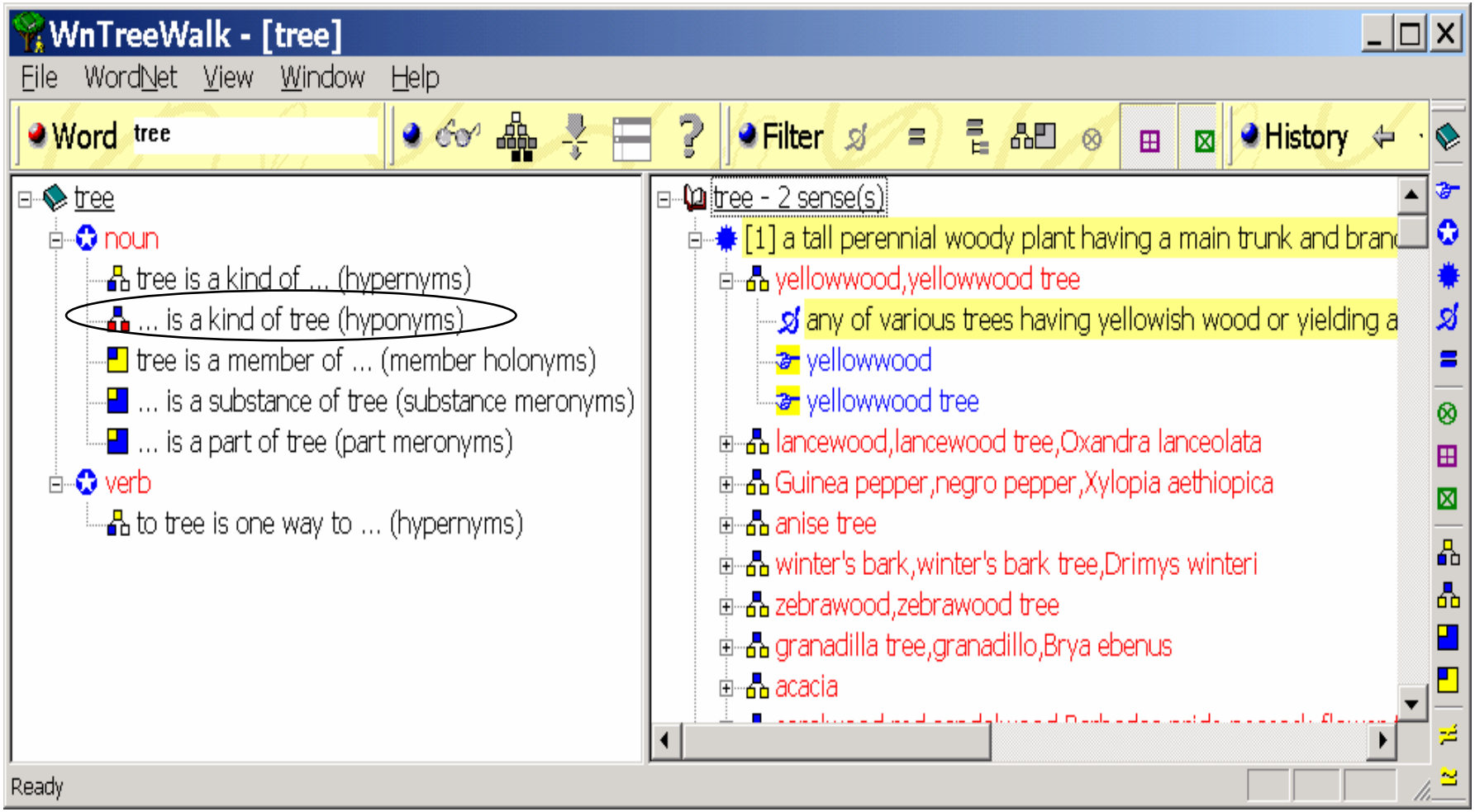
形容词概念的组织：



WordNet (6)


WnTreeWalk - [tree]
File WordNet View Window Help
Word tree
Filter
History
tree
noun
tree is a kind of ... (hypernyms)
... is a kind of tree (hyponyms)
tree is a member of ... (member holonyms)
... is a substance of tree (substance meronyms)
... is a part of tree (part meronyms)
verb
to tree is one way to ... (hypernyms)
tree - 2 sense(s)
[1] a tall perennial woody plant having a main trunk and branche
woody plant,ligneous plant
a plant having hard lignified tissues or woody parts especia
woody plant
ligneous plant
[2] a figure that branches from a single root
"genealogical tree"
tree diagram
plane figure,two-dimensional figure
Ready

WordNet (7)


WnTreeWalk - [tree]
File WordNet View Window Help
Word tree
Filter
History
tree
noun
tree is a kind of ... (hypernyms)
... is a kind of tree (hyponyms)
tree is a member of ... (member holonyms)
... is a substance of tree (substance meronyms)
... is a part of tree (part meronyms)
verb
to tree is one way to ... (hypernyms)
tree - 2 sense(s)
[1] a tall perennial woody plant having a main trunk and branc
yellowwood,yellowwood tree
any of various trees having yellowish wood or yielding a
yellowwood
yellowwood tree
lancewood,lancewood tree,Oxandra lanceolata
Guinea pepper,negro pepper,Xylopia aethiopica
anise tree
winter's bark,winter's bark tree,Drimys winteri
zebrawood,zebrawood tree
granadilla tree,granadillo,Brya ebenus
acacia
Ready

内容提要：词典

- 词典与词典编纂
- 人读词典与机读词典
- 词典检索算法
- 应用之一：汉语新词语检测
- 应用之二：词汇语义相似度计算

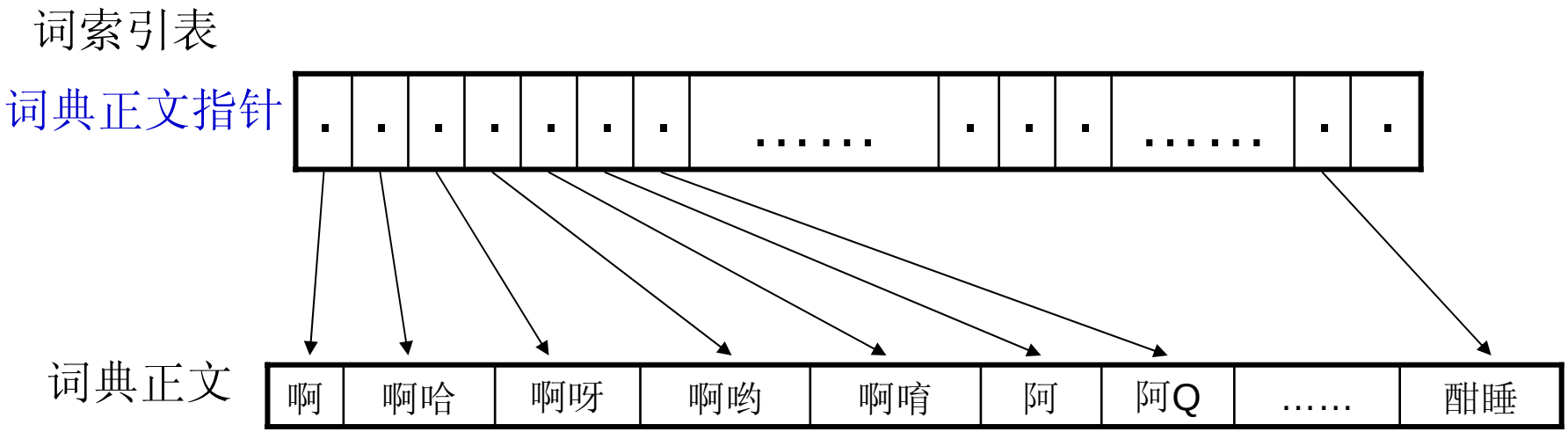
词典检索算法 (1)

- 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
    - 精确匹配检索
    - 前缀匹配检索
      - 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
    - 模糊匹配检索
    - ……
  - 增量式索引：词典增加或修改时不用重建全部索引

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

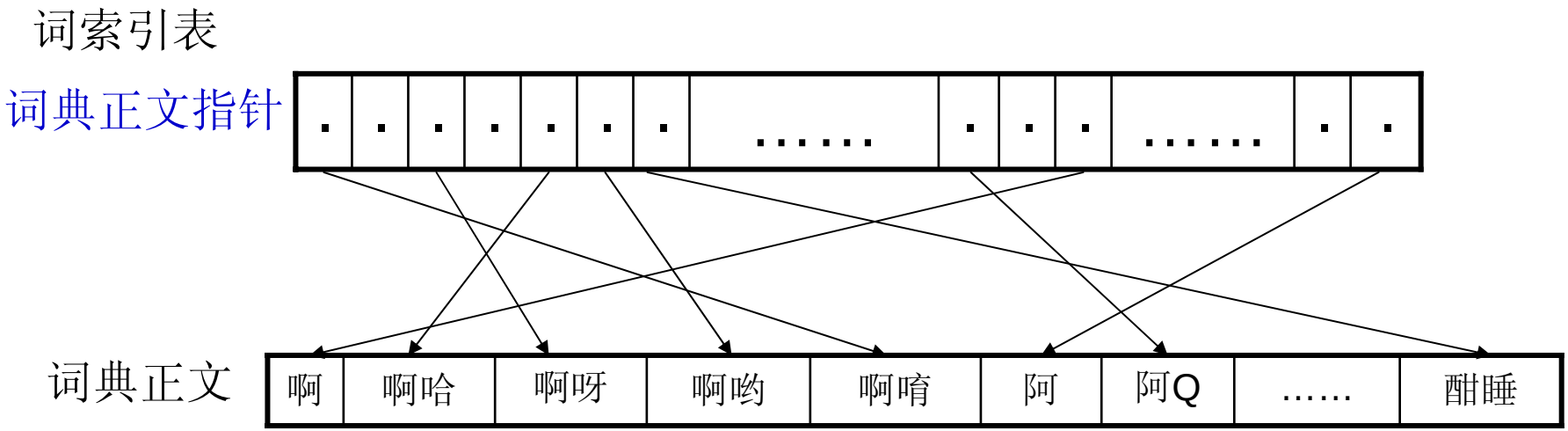


- 索引结构简单，占用空间小
- 不能实现增量式索引：每增加一个词需重新排序

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：查找时精确匹配
- 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：查找时匹配前缀，多分枝查找
  - 例子：
    - 查询词：阿拉丁
    - 词典项：阿 阿Q 阿爸 阿拉 阿拉伯 阿拉伯人 阿拉丁 阿拉斯加

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

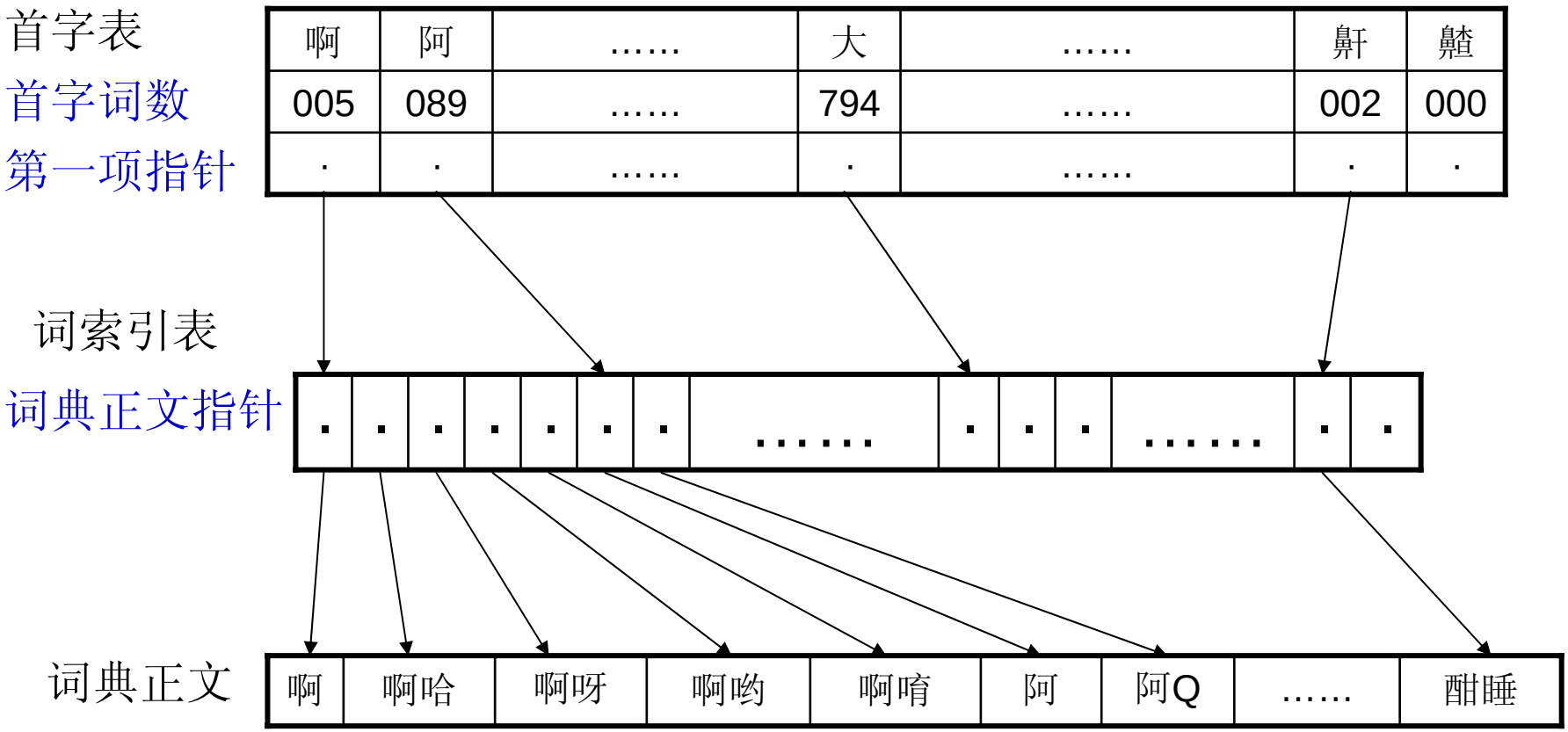
词典散列索引的检索算法

- 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：逐字散列，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

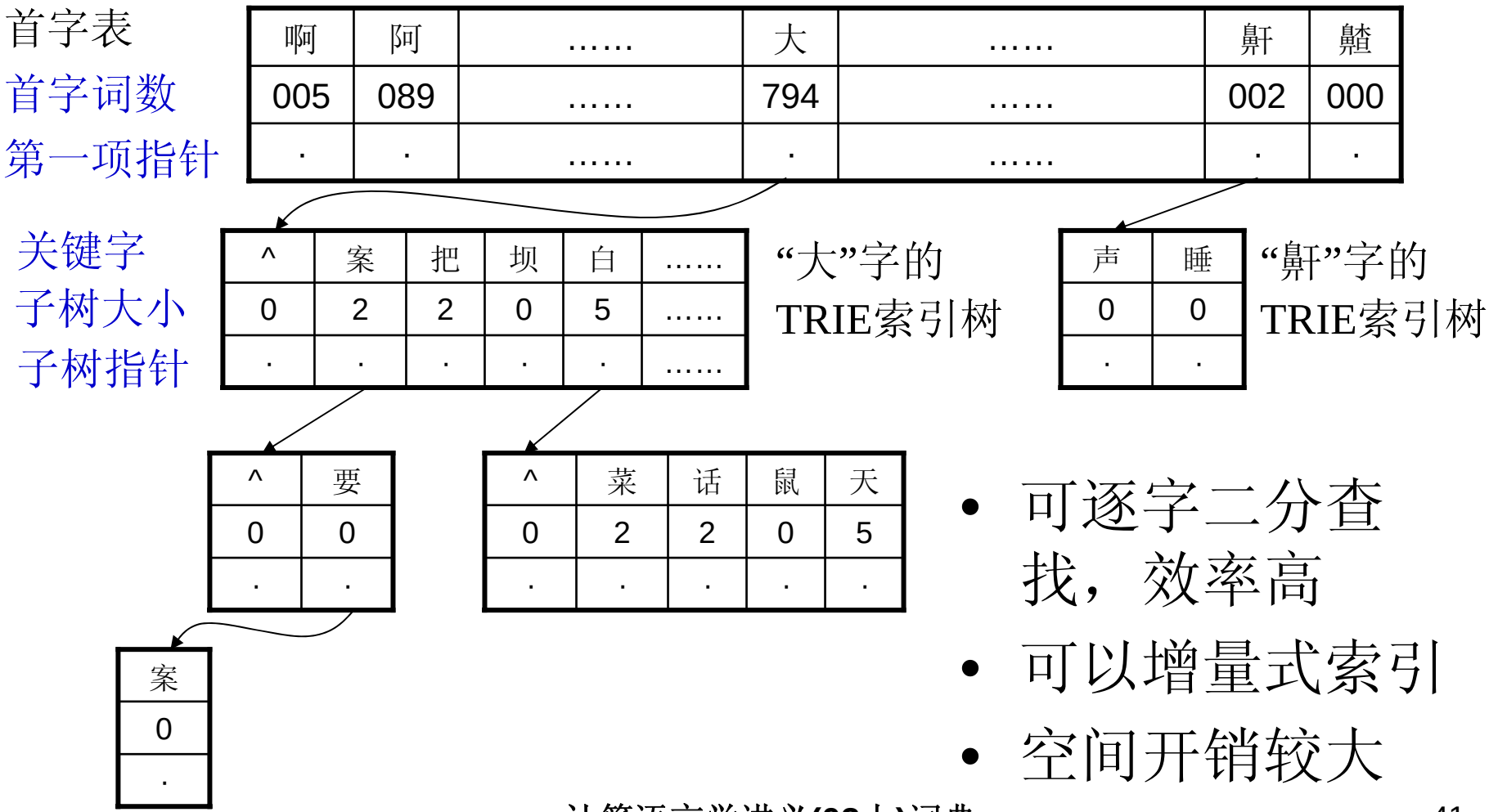
汉语词典按首字顺序索引



首字二分检索

- 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



- 可逐字二分查找，效率高
- 可以增量式索引
- 空间开销较大

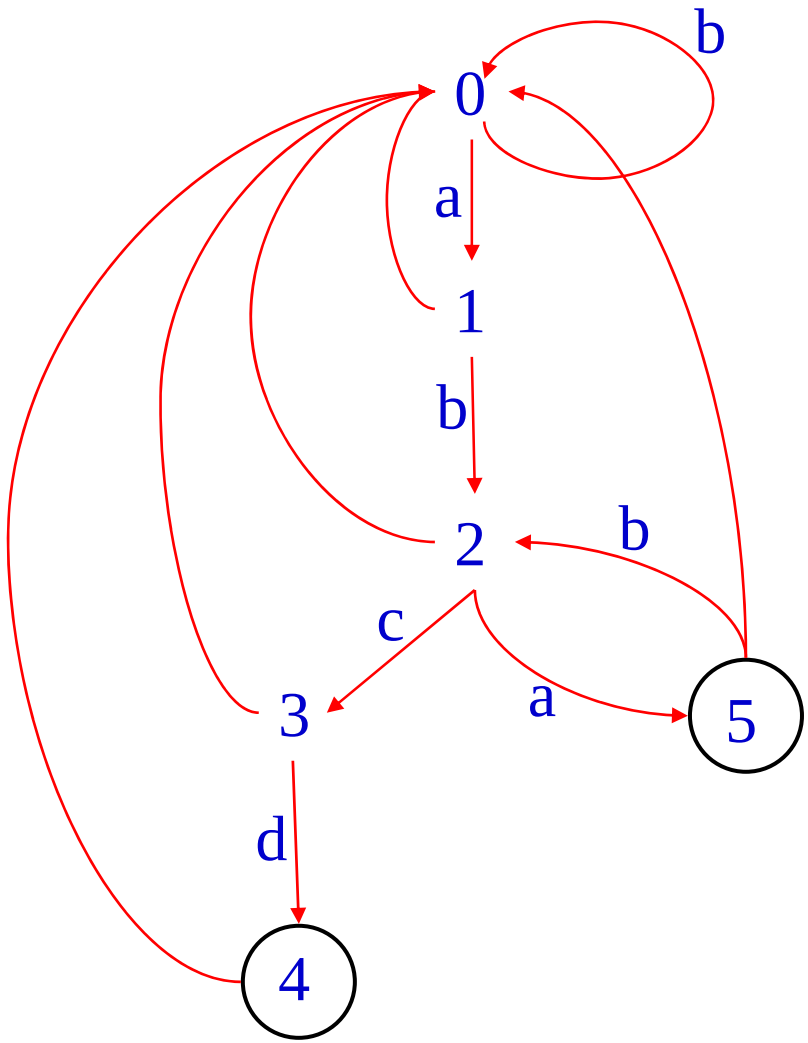
AC算法 (1)

- 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)

aba

abcd

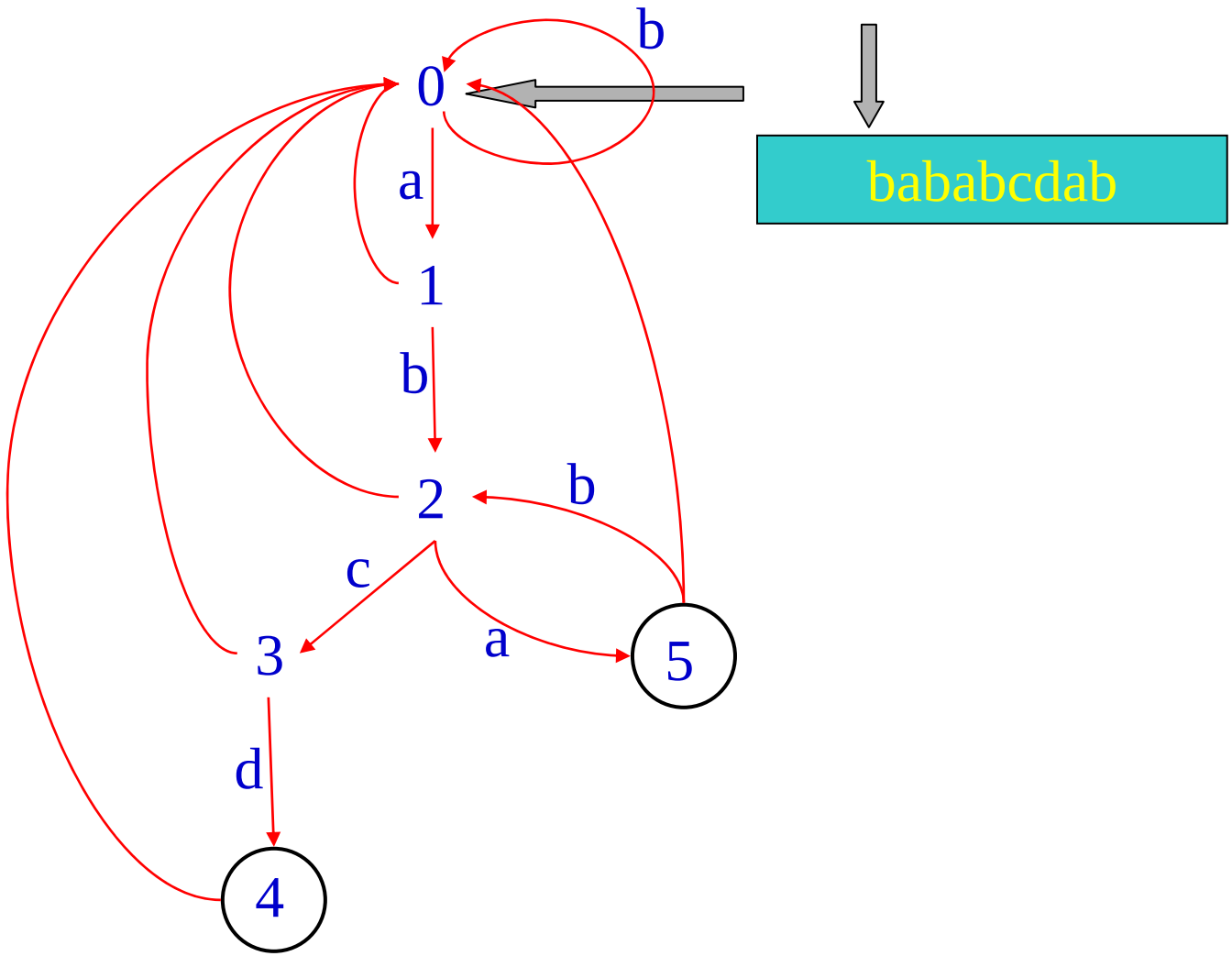


若下一字符没有对应的边，则自动沿空边回到起始状态

AC算法 (3)

aba

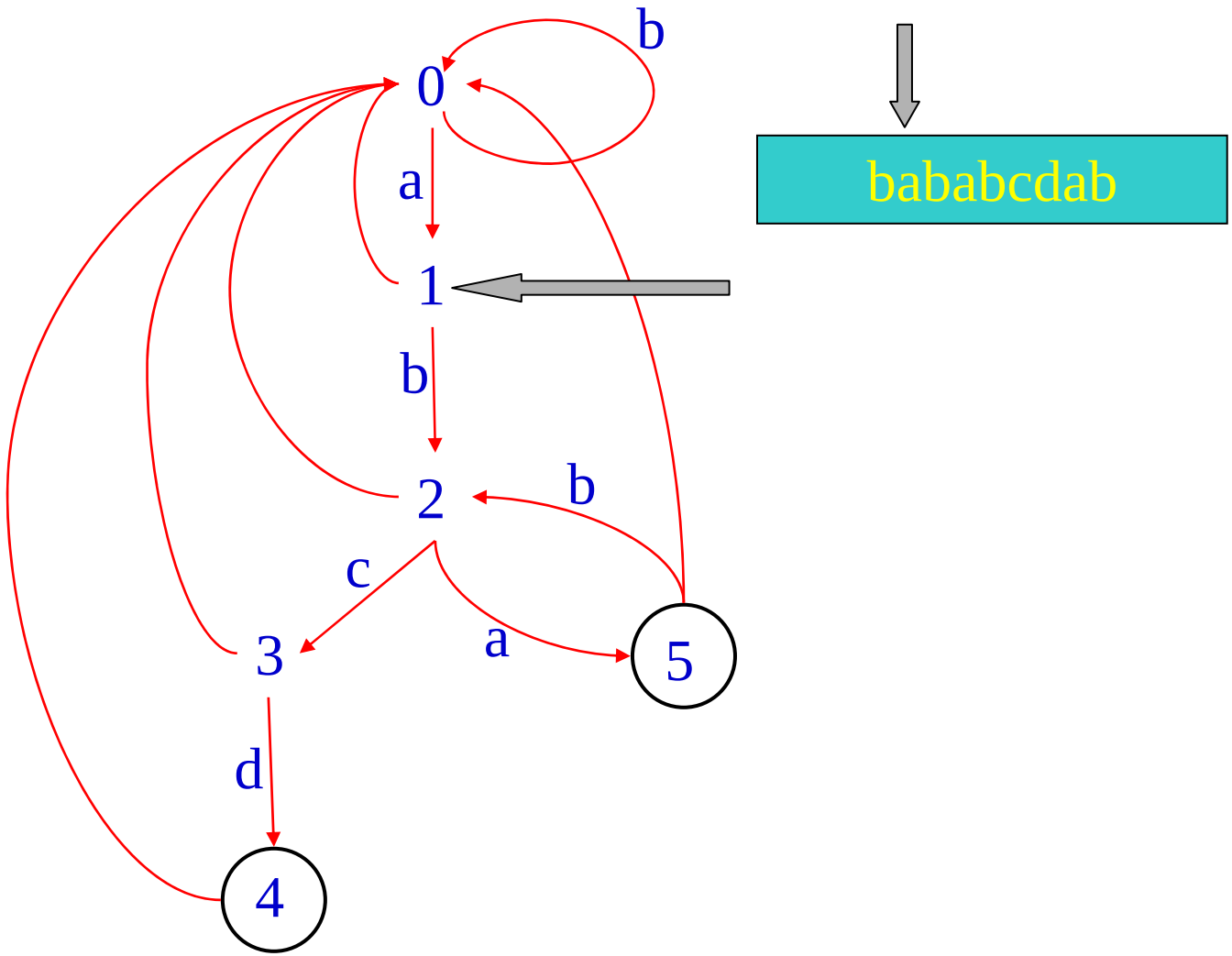
abcd



AC算法 (4)

aba

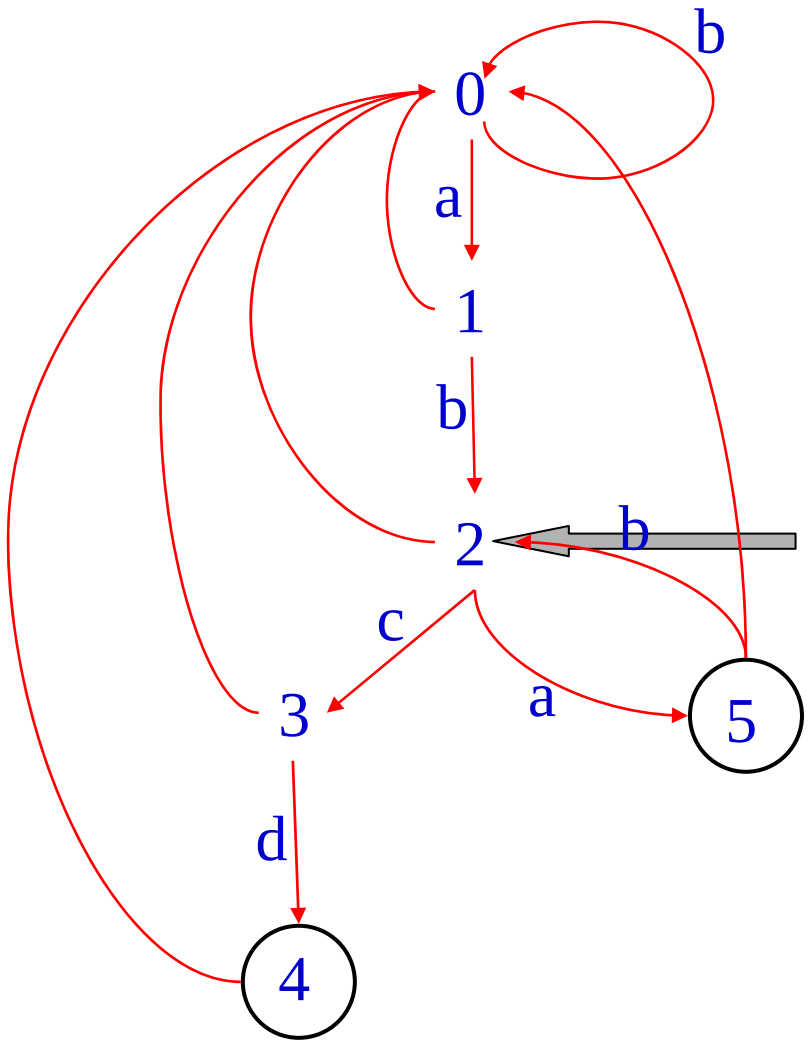
abcd



AC算法 (5)

aba

abcd

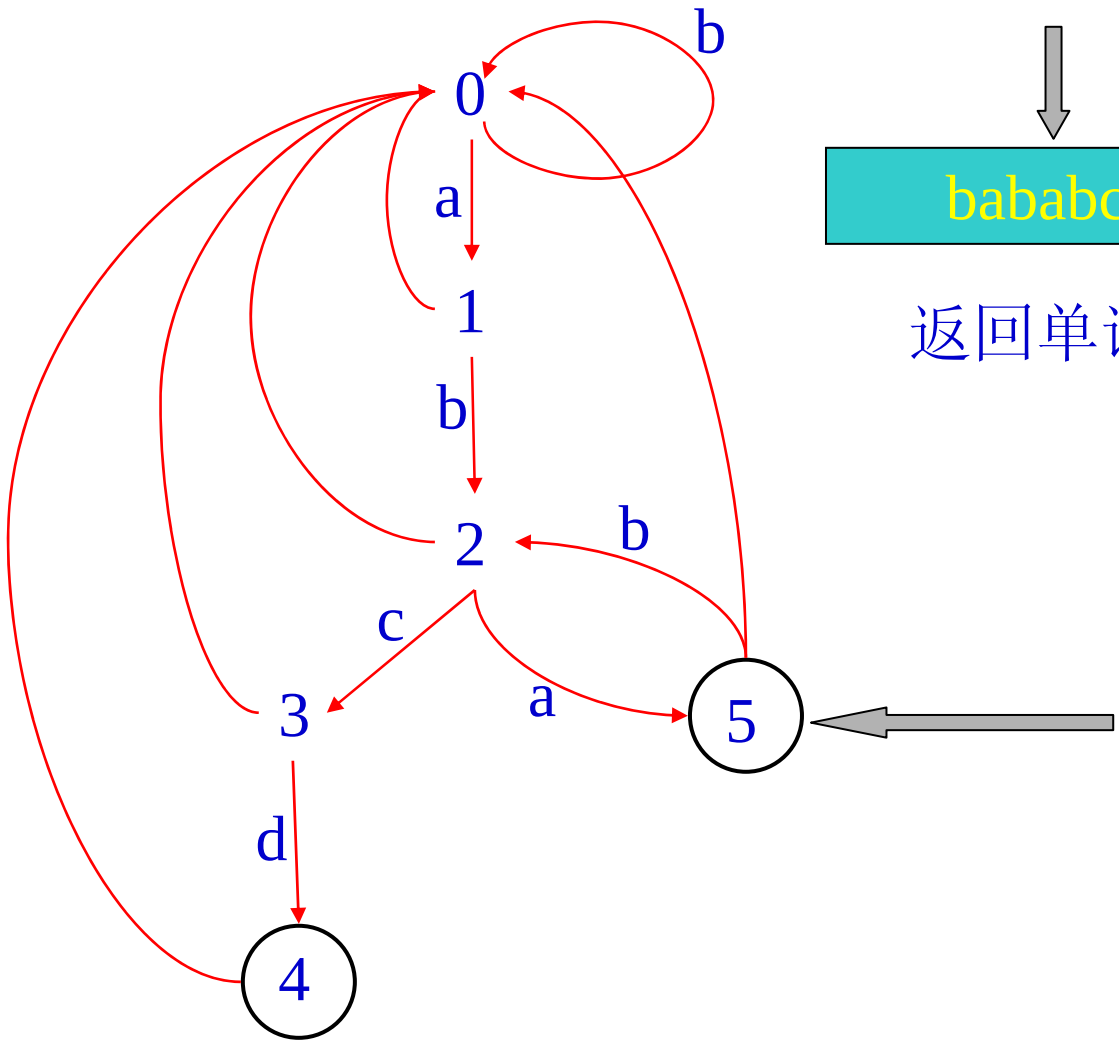


bababcdab

AC算法 (6)

aba

abcd



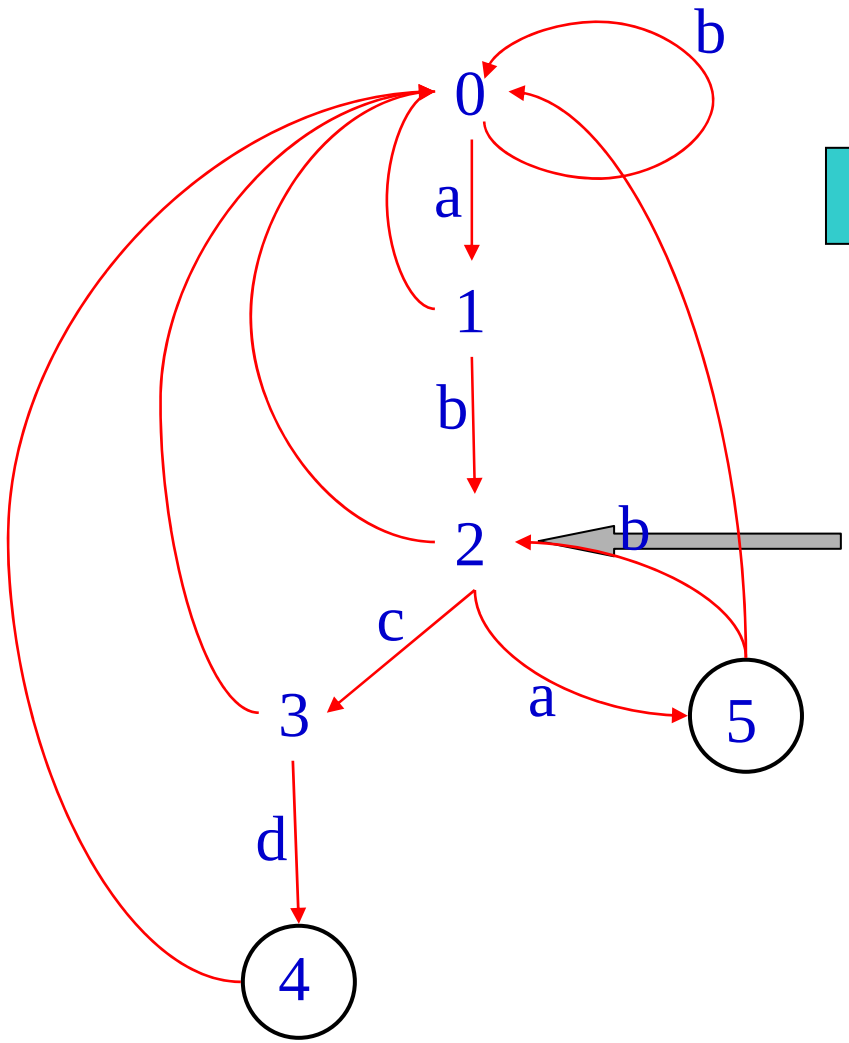
bababcdab

返回单词aba

AC算法 (7)

aba

abcd

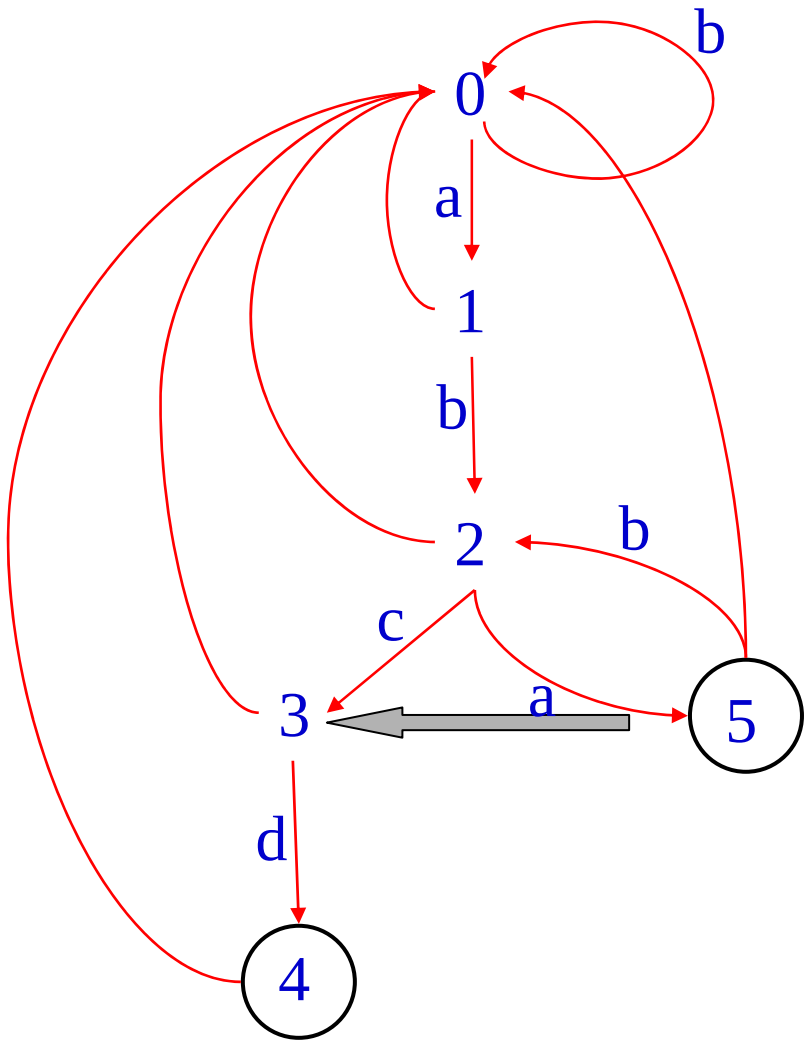


bababcdab

AC算法 (8)

aba

abcd

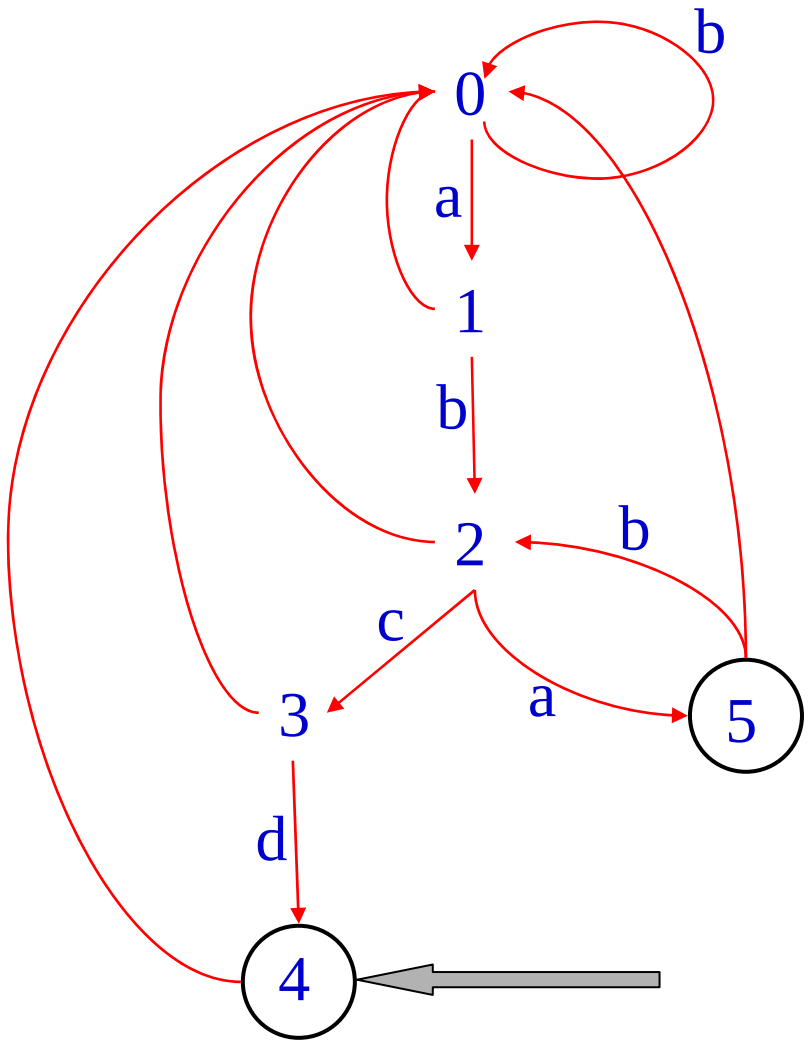


bababcdab

AC算法 (9)

aba

abcd



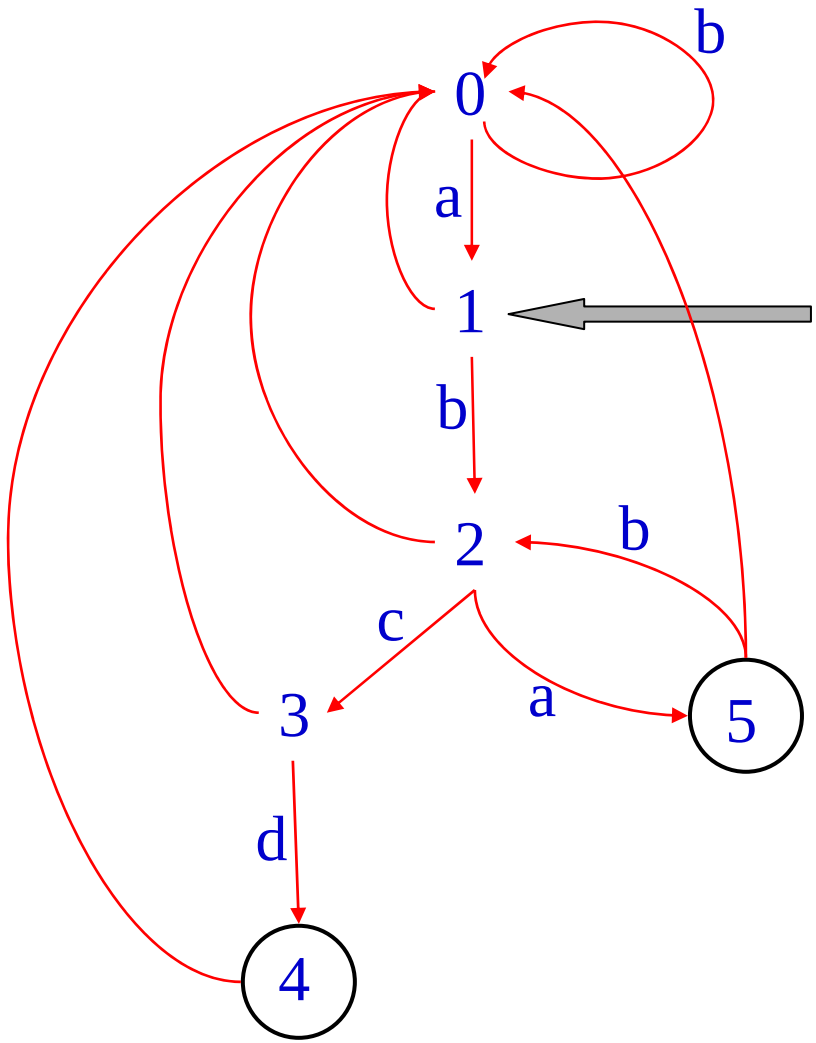
bababcdab

返回单词abcd

AC算法 (10)

aba

abcd

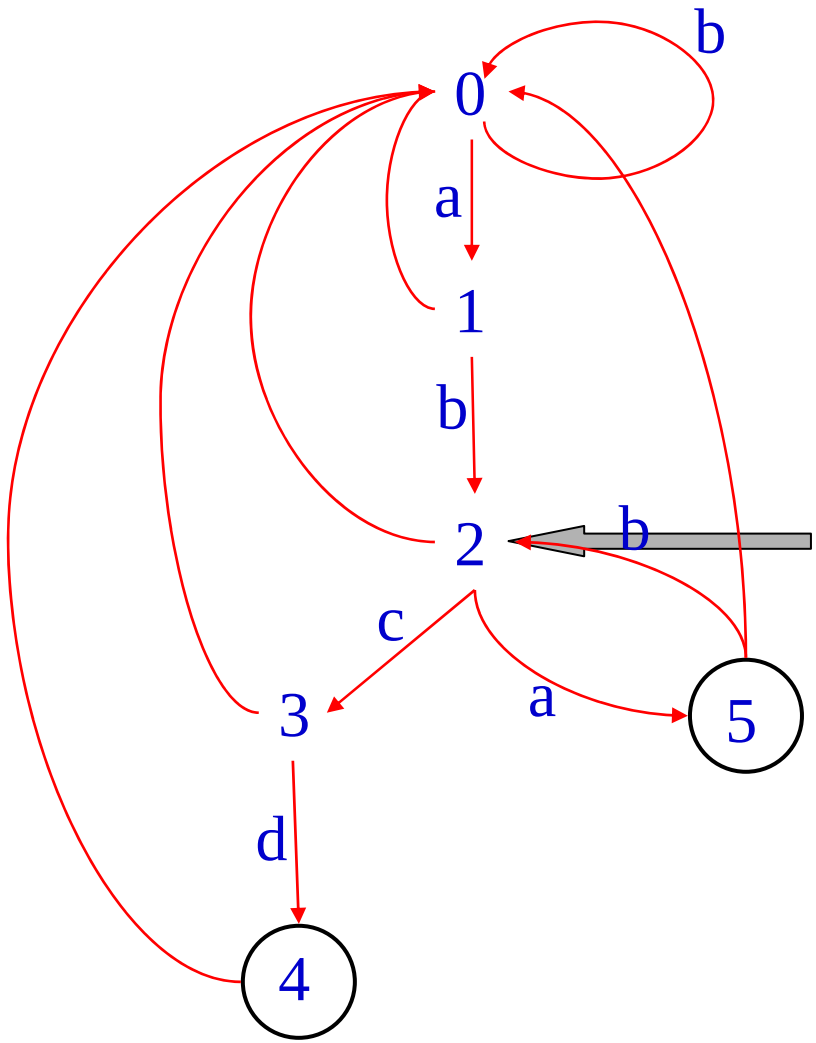


babababcdab

AC算法 (11)

aba

abcd



bababcdab

内容提要：词典

- 词典与词典编纂
- 人读词典与机读词典
- 词典检索算法
- 应用之一：汉语新词语检测
- 应用之二：词汇语义相似度计算

汉语新词语

- 随着经济、社会的飞速发展和对外交流的日渐频繁，自然语言中新词的不断涌现
- 在汉语这种词与词之间没有明显界限的语言中，新词的识别问题尤为严重。除了命名实体和字母词因为有明显的构成特征而相对比较容易被识别外，其他新词很难识别。

汉语新词语分类

- 新词语的种类
  - 命名实体：包括人名、地名、商品名、公司字号、机构名等；
  - 缩略语：如“非典”、“计生委”等；
  - 方言词：如“靓”、“买单”等；
  - 新造词：如“伊妹儿”、“美眉”等；
  - 专业术语：如“非典型肺炎”、“蓝光光盘”等；
  - 音译词：如“酷”、“秀”、“克隆”等；
  - 外来字母词：如WTO、APEC、SARS等。
- 以上划分是按新词语产生的来源进行分类的，在计算机处理时难以找到明显的规律
(命名实体和字母词除外)

汉语新词语的出现规律

- 虽然从词语的构成规律上看不出新词的规律，但是从一个整体的角度看，新词具有下述的两个特征：
  - 新词语在文本中重复出现
  - 新词语出现的时间有规律
- 随着Internet的蓬勃发展以及网页采集技术的成熟，已经可以轻易的从网上采集大规模的网页，同时目前概率词法分析器也达到了比较成熟的阶段，因此这两个条件给我们利用新词语的这两个特征做好了准备。

基于重复串的新词语识别

1. 语料采集：从互联网上定点收集大量文本语料
2. 语料切分：对这些文本语料进行词语切分
3. 重复串检测：从切分的结果中识别出所有的词典中没有出现的重复串
4. 新词语判别：把检测得到的重复串作为候选，判别其是否是真正的新词语
5. 人工校对：采用人机互助的方法从候选新词中找出真正的新词

重复串检测

目标：检测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。

重复串检测算法

- Makoto Nagao的算法
  - Makoto Nagao, Shinsuke Mori. A new method of N-gram statistics for large number of n and automatic extraction of words and phrases from large text data of Japanese. Proceedings of ACL-1994[J], 1994
  - 吕学强对Nagao算法的改进：吕学强，面向机器翻译的E-Chunk获取与应用研究，第三章 单语Chunk获取方法，东北大学博士论文，2003年
  - 复杂度：O(n*log(n))，n是文本长度，算法限制较多
- 后缀树算法
  - 后缀树算法比较成熟，在很多领域都有应用
  - 后缀树构造时间与字符表大小有关，字符表较小时最小可到O(n)，但对于大字符表的情况时间复杂度有所提高
- 邹纲的方法（系统演示）
  - 邹纲，中文新词语自动检测研究，第四章 重复串查找，中国科学院计算技术研究所硕士论文，2004
  - 复杂度：O(k*n)，n是文本长度，k是最长重复串长度，算法限制较少

新词语的判别

- 垃圾过滤：过滤掉重复串中的垃圾
- 常见方法：
  - 反例：词典过滤
  - 规则：过滤规则
  - 内部统计信息：互信息(MI)、DICE系数、x^2 等等
  - 外部统计信息：边界信息熵

新词语识别（实验结果）

- 对于《人民日报》2002年和2001年语料分别进行重复子串识别（单纯依靠重复串检测，没有后期判别）
- 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**

新词语识别（实验结果示例）

词语：抗击非典

出现总次数：3081

出现的文章数：847

1. 报纸名：中国汽车报　日期：2003-06-10

   网址：http://www.people.com.cn/GB/paper1668/9424/872549.html

   例句:抗击非典的斗争已经进入到扫尾阶段。

2. 报纸名：中国汽车报　日期：2003-06-03

   网址：http://www.people.com.cn/GB/paper1668/9366/868060.html

   例句:此外，公司决定立即采购一批专用的清洁机、高压水枪、高效除垢剂以及杀菌消毒喷雾机等投入使用，全力以赴，抗击非典。

3. 报纸名：中国汽车报　日期：2003-05-20

   网址：http://www.people.com.cn/GB/paper1668/9365/867876.html

   例句:日前，郑州宇通客车股份有限公司捐资100万元用于抗击非典

内容提要：词典

- 词典与词典编纂
- 人读词典与机读词典
- 词典检索算法
- 应用之一：汉语新词语检测
- 应用之二：词汇语义相似度计算

什么是词语相似度

- 与具体应用密切相关
  - 实例机器翻译：文本中词语的可替换程度
  - 信息检索：查询扩展的有效程度
  - ……
- 取值范围：[0 , 1] 之间
  - 1: 相似度最高，词语和其本身相似度为1;
  - 0: 相似度最低
- 影响词语相似度的因素：
  - 形态、句法、语义、语用；语义起主要作用

词语相似度与词语距离

- 词语距离取值范围：[0，∞)
- 词语相似度和词语距离的关系
  - 词语距离为 0 ⇔ 词语相似度为1
  - 词语距离为∞ ⇔ 词语相似度为0
  - 词语距离越大，相似度越小（单调递减）
- 转换公式：（示例）

$$Sim\ (W_1, W_2) = \frac{\alpha}{Dis\ (W_1, W\ ) + \alpha}$$

α：可调节参数，相似度0.5时的距离值

词语相似度与词语相关性(1)

- 词语相关性
  - 两个词语互相关联的程度
  - 可以用共现的频率来衡量
  - 取值范围：[0,1]之间
- 相似度和相关性反映词语的不同特点
  - 相关性高未必相似性高：医生，疾病
  - 相似性高未必相关性高：耶和华，玉皇大帝

词语相似度与词语相关性(2)

- 相关性与相似度反映词语的不同特点
  - 相关性反映的是词语的组合特性
  - 相似性反映的是词语的聚合特性
- 相关性和相似度有着密切的联系
  - 相似度高的两个词，它们与其他词的相关性特点也相似
  - 可以利用相关性特点来计算词的相似性
    李涓子(1999)

词语相似度计算方法(1)

- 基于世界知识的词语相似度计算方法
  - 利用同义词词典（Thesaurus）
  - 计算两个概念在树状概念层次体系中的距离
  - 影响概念距离的其他因素（深度、密度）
- 基于语料库的词语相似度计算方法
  - 利用固定宽度的文本窗口得到词语共现向量根据共现向量夹角余弦得到相似度
  - 利用其他统计信息（如互信息）

词语相似度计算方法(2)

- 基于世界知识的词语相似度计算方法
  - 简单有效，直观，易于理解
  - 主观性较强，有时不够客观
  - 主要反映语义特征，忽略句法语用等特征
- 基于语料库的词语相似度计算方法
  - 客观，综合反映形态句法语义语用等特点
  - 计算复杂，性能依赖于语料库
  - 数据稀疏严重，噪声干扰大

基于《知网》的词汇语义相似度计算 (1)

刘群，李素建，基于《知网》的词汇语义相似度计算，第三届汉语词汇语义学研讨会，台北，2002年5月，修改后被收录于：Computational Linguistics and Chinese Language Processing，Vol.7, No.2, August 2002, pp.59-76

基于《知网》的词汇语义相似度计算 (2)

- 基于知识的词语相似度计算方法
- 与通常的同义词词典不同，在《知网》中，词语概念不是表现为一个概念层级体系中的一个结点，而是用一专用的知识描述语言来表示
- 举例：试计算以下几个概念之间的相似度：
  - 男人：human|人,family|家,male|男
  - 爱好者：human|人,*FondOf|喜欢,#WhileAway|消闲
  - 儿童基金会：part|部件,%institution|机构,politics|政,#young|幼,#fund|资金,(institution|机构=UN|联合国)

《知网》的知识描述语言(1)

词	概念编号	描述语言
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)

注：这些例子取自知网1999版，与前面介绍的的2005版的例子有所不同

《知网》的知识描述语言(2)

- 虚词描述格式：“{句法义原}”、“{关系义原}”；
- 实词描述格式：由一系列用逗号隔开的“语义描述式”组成，这些“语义描述式”有三种形式：
  - 独立义原描述式：“基本义原”、“(具体词)”；
  - 关系义原描述式：“关系义原=基本义原”、“关系义原=(具体词)”、“(关系义原=具体词)”
  - 符号义原描述式：“关系符号 基本义原”、“关系符号(具体词)”
- 在实词的描述中，第一个描述式总是一个基本义原，描述了该实词的最基本的语义特征

基于《知网》的词语相似度计算

- 困难：知识描述语言的复杂语法
- 方法一：
  - 只计算第一独立义原的相似度
  - 优点：简单
  - 缺点：过于粗疏
- 方法二：Li Sujian, et al. (2002)
  - 综合利用《知网》和《同义词词林》
  - 利用了《知网》义原之间除上下位以外的其他关系
  - 综合考虑相似度和相关度（未必合理）

词语的相似度计算

对于两个汉语词语W_1和W_2，如果W_1有n个义项（概念）：S_{11}，S_{12}，……，S_{1n}，W_2有m个义项（概念）：S_{21}，S_{22}，……，S_{2m}，我们规定，W_1和W_2的相似度各个概念的相似度之最大值，也就是说：

$$Sim(W_1, W_2) = \max_{i=1..n, j=1...m} Sim(S_{1i}, S_{2j})$$

注：在实际的文本中最好先排岐。

义原的相似度计算

- 义原之间的语义距离：

$$Sim(p_1, p_2) = \frac{\alpha}{d + \alpha}$$

- 其中p1和p2表示两个义原（primitive），d是p1和p2在义原层次体系中的路径长度，是一个正整数。α 是一个可调节的参数。
- 具体词与义原的相似度一律处理为一个小常数（γ）；
- 具体词和具体词的相似度，如果两个词相同，则为1，否则为0。
- 将任何义原（或具体词）与空值的相似度定义为一个小常数（δ）；

虚词概念的相似度计算

- 由于虚词概念总是用“{句法义原}”或“{关系义原}”这两种方式进行描述，所以，虚词概念的相似度计算非常简单，只需要计算其对应的句法义原或关系义原之间的相似度即可。

实词概念的相似度计算(1)

- 基本原则：
  - 整体相似要建立在部分相似的基础上。
  - 把一个复杂的整体分解成部分，通过计算部分之间的相似度得到整体的相似度。
  - 先在二者的各个部分之间建立一一对应关系（组合配对），分别计算各个组合配对的相似度；
  - 整体相似度等于各个组合配对的相似度的加权评价；
  - 古代的战场的两军对垒：兵对兵、将对将，捉对厮杀 。

实词概念的相似度计算(2)

- 将实词概念的语义表达式分成四个部分：
  - 第一独立义原描述式：$Sim_1(S_1,S_2)$；
  - 其他独立义原描述式：$Sim_2(S_1,S_2)$；
  - 关系义原描述式：$Sim_3(S_1,S_2)$
  - 符号义原描述式：$Sim_4(S_1,S_2)$
- 实词概念整体相似度计算公式：

$$Sim(S_1,S_2)=\sum_{i=1}^{4}\beta_i Sim_i(S_1,S_2)$$

$$\beta_1+\beta_2+\beta_3+\beta_4=1$$

$$\beta_1\geqslant\beta_2\geqslant\beta_3\geqslant\beta_4$$

实词概念的相似度计算(3)

- 发现的问题：如果Sim_1非常小，但Sim_3或者Sim_4比较大，将导致整体的相似度仍然比较大的不合理现象
- 改进的公式：

$$Sim(S_1,S_2)=\sum_{i=1}^{4}\beta_i\prod_{j=1}^{i}Sim_j(S_1,S_2)$$

- 改进的意义：主要部分的相似度值对于次要部分的相似度值起到制约作用，也就是说，如果主要部分相似度比较低，那么次要部分的相似度对于整体相似度所起到的作用也要降低。

实词概念的相似度计算(4)

- 第一独立义原描述式：
  - ❖直接计算两个义原的相似度
- 其他独立义原描述式：
  - ❖两个义原集合的相似度：配对困难
  - – 先计算出所有可能的配对的义原相似度
  - – 取相似度最大的一对，并将它们归为一组
  - – 在剩下的独立义原的配对相似度中，取最大的一对，并归为一组，如此反复，直到所有独立义原都完成分组

实词概念的相似度计算(5)

- 关系义原描述式：
  - ❖把关系义原相同的描述式分为一组，并计算其相似度
- 符号义原描述式：
  - ❖把关系符号相同的描述式分为一组，并计算其相似度
- 计算以上各部分的相似度时，权值都取等值

实验设计(1)

- 词语相似度结果评价
  - 放到实际的系统中（如基于实例的机器翻译系统），观察不同的相似度计算方法对实际系统的性能的影响
  - 人工判别：我们采用的办法
- 实验一
  - 采用本文中提出的词语相似度计算方法
  - 计算一个词和另外选取的一组词的相似度，判断是否符合人的直觉

实验设计(2)

- 实验二
  - ❖三种方法对比
  - – 方法一：仅使用《知网》语义表达式中第一独立义原来计算词语相似度
  - – 方法二：Li Sujian et al. (2002) 中使用的词语语义相似度计算方法
  - – 方法三：本文中介绍的语义相似度计算方法
- ➢ *参数选择：* $\alpha = 1.6$， $\gamma = 0.2$， $\delta = 0.2$
- ➢ *参数选择：* $\beta_1 = 0.5$, $\beta_2 = 0.2$， $\beta_3 = 0.17$， $\beta_4 = 0.13$

实验结果

词语1	词语2	词语2的语义	方法1	方法2	方法3
男人	女人	人,家,女	1.000	0.668	0.833
男人	父亲	人,家,男	1.000	1.000	1.000
男人	母亲	人,家,女	1.000	0.668	0.833
男人	和尚	人,宗教,男	1.000	0.668	0.833
男人	经理	人,#职位,官,商	1.000	0.351	0.657
男人	高兴	属性值,境况,福,良	0.016	0.024	0.013
男人	收音机	机器,*传播	0.186	0.008	0.164
男人	鲤鱼	鱼	0.347	0.009	0.208
男人	苹果	水果	0.285	0.004	0.166
男人	工作	事务,$担任	0.186	0.035	0.164
男人	责任	责任	0.016	0.005	0.010

实验结果分析

- 实验一：考察方法3的结果
  - 与人的直觉比较符合
- 实验二：比较三种方法的结果
  - 方法1的结果比较粗糙，只要是人，相似度都为1，显然不够合理
  - 方法2的结果比方法1更细腻一些，能够区分不同人之间的相似度
  - 方法2有些相似度的结果也不太合理，比如“男人”和“工作”的相似度比“男人”和“鲤鱼”的相似度更高
  - 方法2的结果中，“男人”和“和尚”的相似度比“男人”和“经理”的相似度高出近一倍，不如方法3结果好

复习思考题

- 如果有一部人读的双语词典，你如何将它加工成机读词典？
- 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。