计算机信息检索

第4章 信息检索的模型(IR models)

几个问题

- 模型是什么？
- 信息检索模型的内容、分类？
- 信息检索模型的作用？
- 经典的信息检索模型

信息检索模型的由来

信息检索核心问题

检测哪些文献与用户的提问需求相关，哪些不相关，即判断一篇文献是否与用户的查询条件相关?

一系列判定文档是否相关的数学模型

信息检索系统采用检索模型的不同决定了系统的检索性能和检索效果

提纲

1. 信息检索模型的定义和分类
2. 布尔模型*
3. 向量空间模型*
4. 概率模型*
5. 改进的集合论检索模型
6. 改进的代数检索模型
7. 改进的概率检索模型
8. 结构化文本检索模型

1 信息检索模型的定义和分类

- 信息检索模型的定义
- 信息检索模型的分类

信息检索模型的定义

- 模型是采用数学工具，对现实世界某种事物或某种运动的抽象描述。面对相同的输入，模型的输出应能够无限地逼近现实世界的输出，能够透过现象看本质。
- 举例：
  - 天气预测模型
  - 人口增长模型

天气预报模型

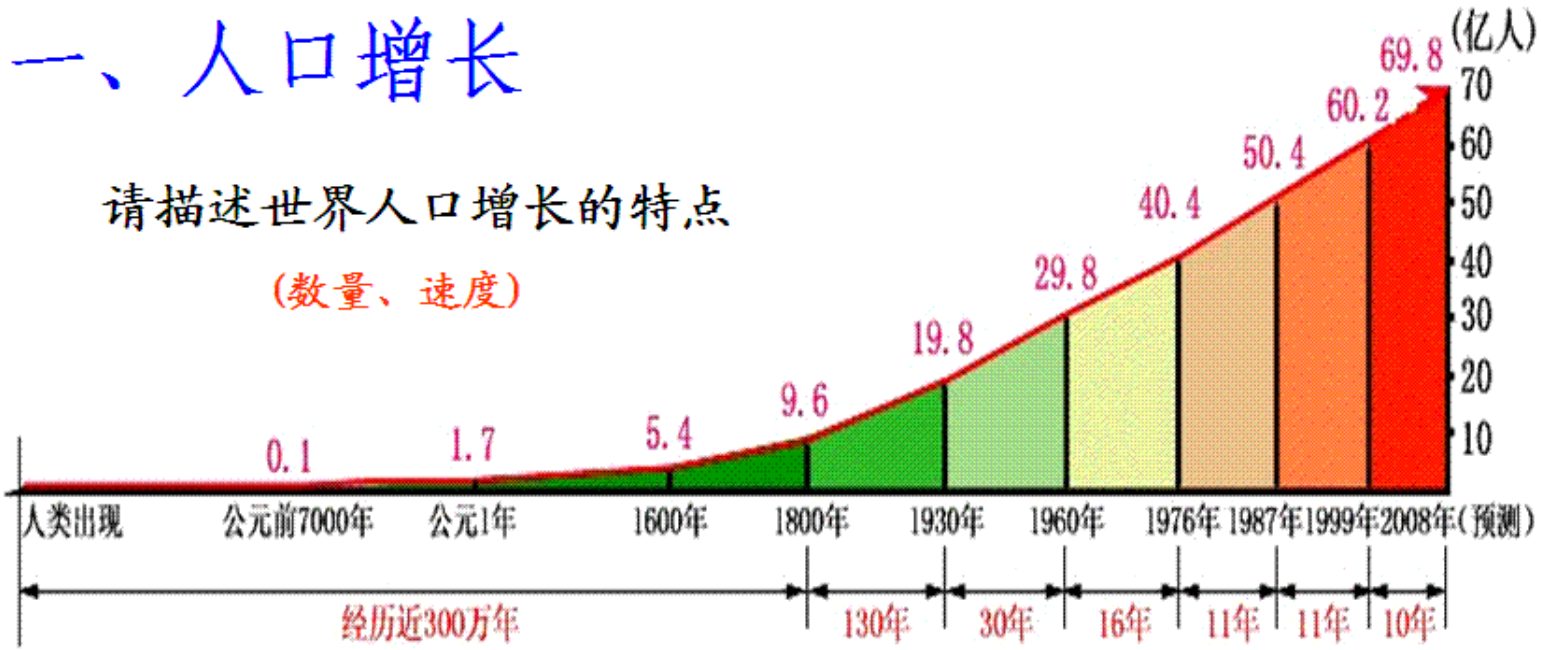
- 天气预报是根据对卫星云图和天气图的分析，结合有关气象资料、地形和季节特点、群众经验等综合研究后作出的。
- 如中国中央气象台的卫星云图，就是中国制造的“风云一号”气象卫星摄取的。利用卫星云图照片进行分析，能提高天气预报的准确率。
- 天气预报就时效的长短通常分为三种：短期天气预报（2—3天）、中期天气预报（4—9天），长期天气预报（10—15天以上）。
- 温度、湿度、气压和风速、风向

人口增长模型

一、人口增长

请描述世界人口增长的特点

（数量、速度）



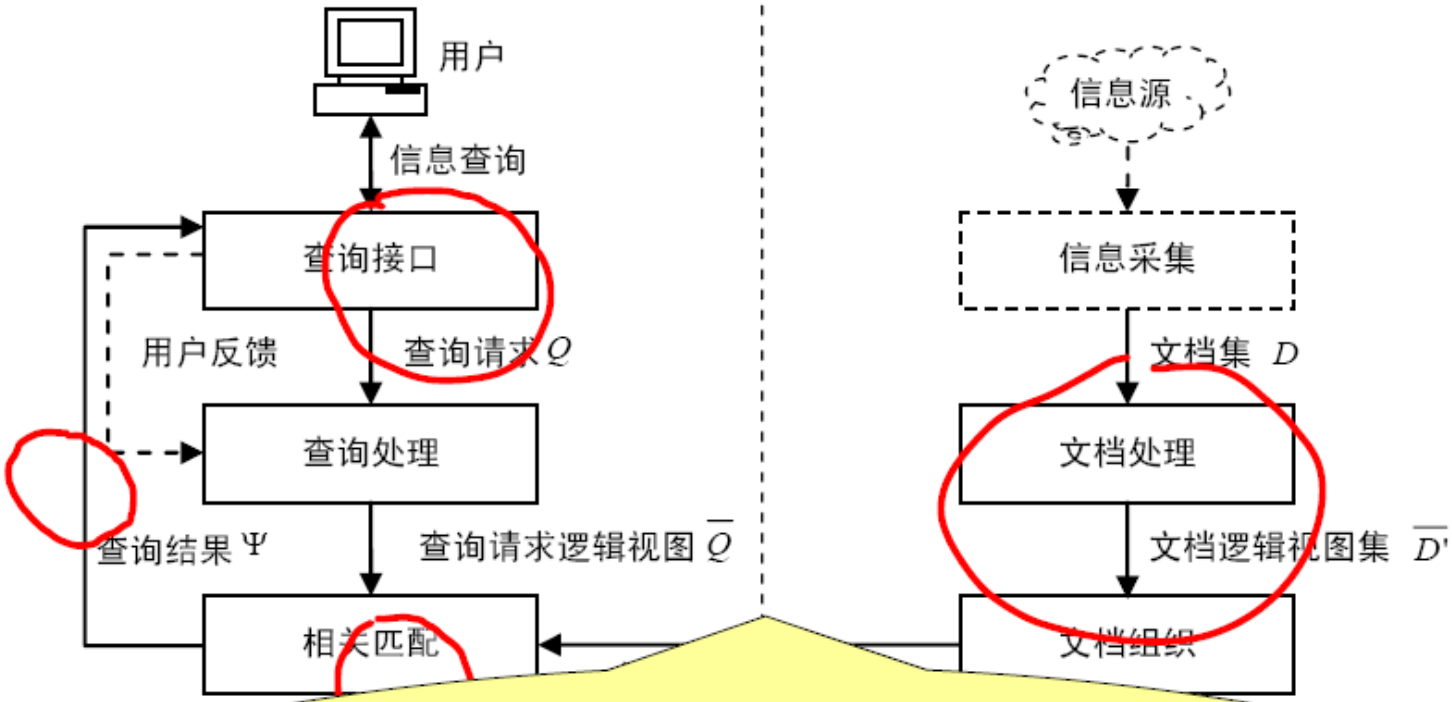
世界人口增长图

人口的自然增长率就是指一定时期内某地人口出生率与死亡率相减的得数。

自然增长率= 出生率-死亡率

= （出生人口-死亡人口）/总人数

信息检索模型的定义



关键问题：

1. 从什么样的视角去看待查询式和文档
2. 基于什么样的理论去看待查询式和文档的关系
3. 如何计算查询式和文档之间的相似度

信息检索模型的定义

信息检索的实施过程

用户通过一系列关键词来阐明自己的信息需求①，信息检索系统则检索与用户查询最为匹配（接近）的文献②，同时借助某种相关性指标对检索出的文献进行排序③

由此可知，信息检索的实施应包含以下几部分：

信息检索模型的定义

①**用户的需求表示：** 包括用户查询信息的获取与表示；

②**文档的表示：** 即文档内容的识别和表示；

③**匹配机制：** 文档和用户需求之间的相关性排序的准则和函数表示，其中相关性排序的准则决定了信息检索系统的基本性能；

④**反馈修正：** 根据检索结果对查询表示进行扩充与参数优化，以提高系统性能。

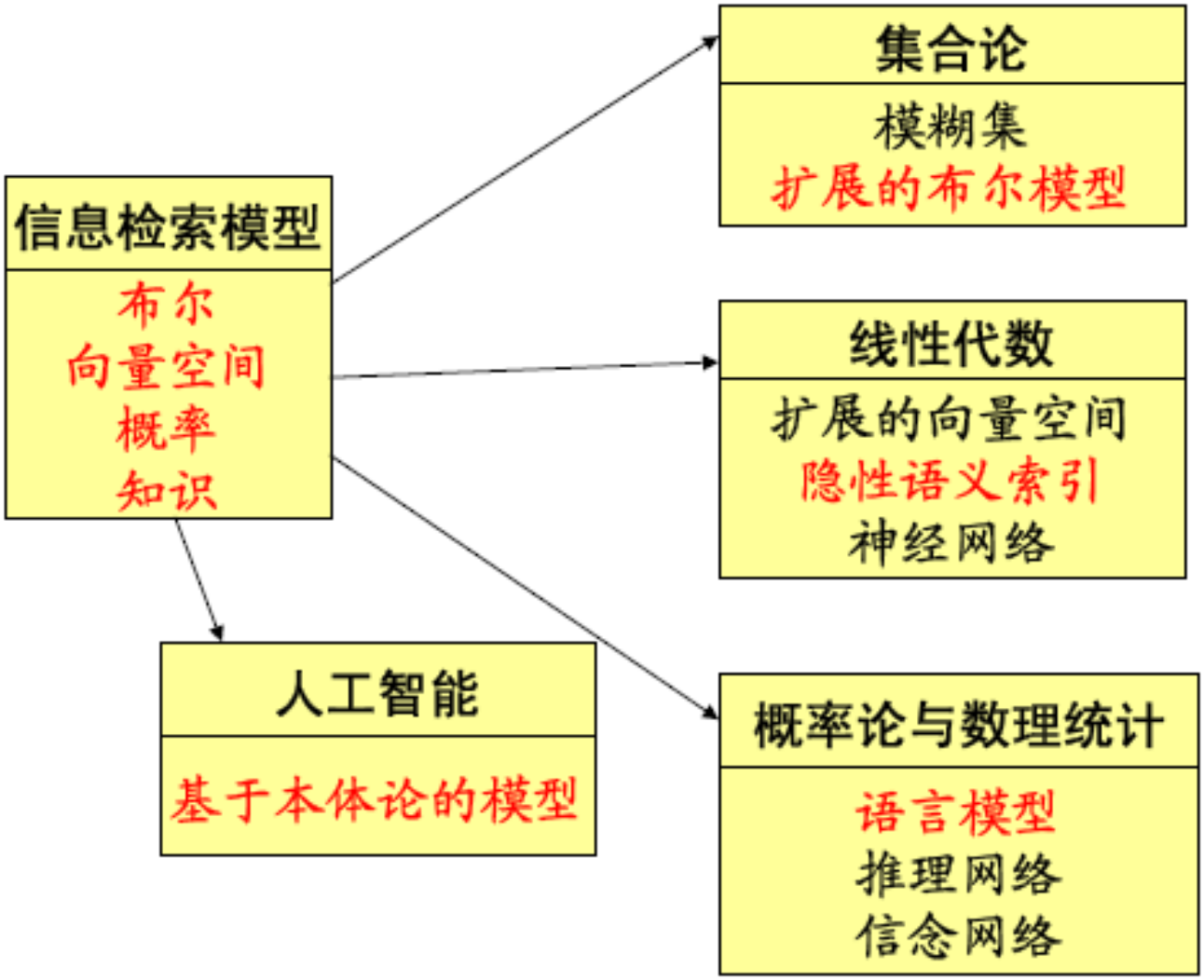
信息检索模型的定义

- **信息检索模型的一般定义：**

  用一个四元组[D, Q, F, R(qi, dj)] 表示，其中：

  - D: 文档集的机内表示（词或字或短语或N元组）
  - Q: 用户需求的机内表示
  - F: D与Q之间的检索匹配框架(Frame)
  - R(qi, dj): 排序函数，计算qi和dj相关度

信息检索模型的分类



信息检索模型的分类

- **布尔模型**：布尔代数 → [0, 1]集合论 → 集合论模型
- **向量模型**：文献和查询都用*t*维空间的向量来表示，亦称代数模型；
- **概率模型**：把检索看作是文献表示和查询之间匹配程度的概率估计问题。

提纲

1. 信息检索模型的定义和分类
2. 布尔模型*
3. 向量空间模型*
4. 概率模型*
5. 改进的集合论检索模型
6. 改进的代数检索模型
7. 改进的概率检索模型
8. 结构化文本检索模型

2. 布尔模型

- 布尔模型的定义
- 布尔模型示例
- 布尔模型应用情况
- 布尔模型优缺点

布尔模型的定义

利用布尔逻辑运算符进行*检索词和文档*的逻辑匹配

布尔模型的定义

布尔逻辑在布尔检索中定义：

A为真：A=1	B为真：B=1
A为假：A=0	B为假：B=0

可以从这两种陈述中产生第三种陈述C，即

$$C = f(A, B)$$

其中，A可看成是提问，B可看成是被检索的文献，C可看作是检索的结果，f 可看作是检索模型

布尔模型的定义

① f为逻辑与（AND）：A与B都为真时，C才为真

② f为逻辑或（OR）：A与B只要一个为真C就为真

③ f为逻辑异或（XOR）：A与B的陈述不同时，C就为真

④ f为逻辑非（NOT）：C=NOT(A AND B)，即：

A=0，B=1
A=1，B=0
A=0，B=0
⟶ C=1

A=1，B=1
⟶ C=0

布尔模型的定义

- **信息检索一般模型[D, Q, F, R(qi, dj)] 解释为：**
  - 文档D表示为标引项的集合，各标引项权值采用二值{0，1}表示。
  - 用户查询Q表示查询项的布尔组合，用“与、或、非”连接起来，并用括弧指示优先次序. 为了便于计算，一般采用析取范式表示。

布尔模型的定义

- **信息检索一般模型[D, Q, F, R(qi, dj)] 解释为：**
  - 检索匹配框架F：
    - 一个文档当且仅当它能够满足布尔查询时，才将其检索出来。
    - 检索策略基于二值判定标准。
  - 排序函数R
    - 根据匹配检索框架F判定文档dj和qi是否二值{0，1}相关。

布尔模型的定义

- 布尔模型是基于集合论和布尔代数的一种简单检索模型，是早期搜索引擎所使用的检索模型。
- 它的特点是查找那些对于某个查询词返回为“真”的文档。
- 在该模型中，一个查询词就是一个布尔表达式，包括关键词以及逻辑运算符。

布尔模型的定义

- **通过布尔表达式，可以表达用户希望文档所具有的特征：**
  - 例如必须包含哪些关键词，不能包含哪些关键词等等。
- **例如我们希望查找那些既含有“南京”又含有“大学”的网页，那么查询词可以写作**
  - “南京 AND 大学”
- **由于文档必须严格符合检索词的要求才能够被检索出来，因此布尔检索模型又被称为“完全匹配检索”(Exact-Match Retrieval)。**

布尔模型示例

- 例：文档集包含两个文档：

  文档1：a b c f g h

  文档2：a f b x y z

  用户查询：文档中出现a或者b，但一定要出现z。

- 检索过程：

a) 将查询表示为布尔表达式 $q = (a \vee b) \wedge z$ ，并转换成析取范式 $q_{DNF} = (1,0,1) \vee (0,1,1) \vee (1,1,1)$

b) 文档1和文档2的三元组对应值分别为(1,1,0)和(1,1,1)

c) 经过匹配 ,将文档2返回

文献号	主题词							
	A	B	C	D	E	F	G	H
1	1	1	0	1	1	0	1	1
2	0	0	1	0	0	1	0	1
3	0	0	0	0	0	1	1	1
4	1	0	0	1	0	0	0	0
5	0	1	0	1	0	0	1	0
6	0	1	1	0	0	0	1	0
7	1	0	0	0	1	1	0	0
8	0	1	0	0	0	1	0	1
9	1	0	0	1	0	0	0	0
10	0	0	1	0	1	0	1	0

这10篇文献通过8个主题词来揭示其内容，即通过8个主题词的不同组合来揭示10篇文献中的不同内容

如果检索的逻辑表达式为：A OR B，则检索结果为：

1，4，5，6，7，8，9

如果检索的逻辑表达式为：A AND B，则检索结果为：

1

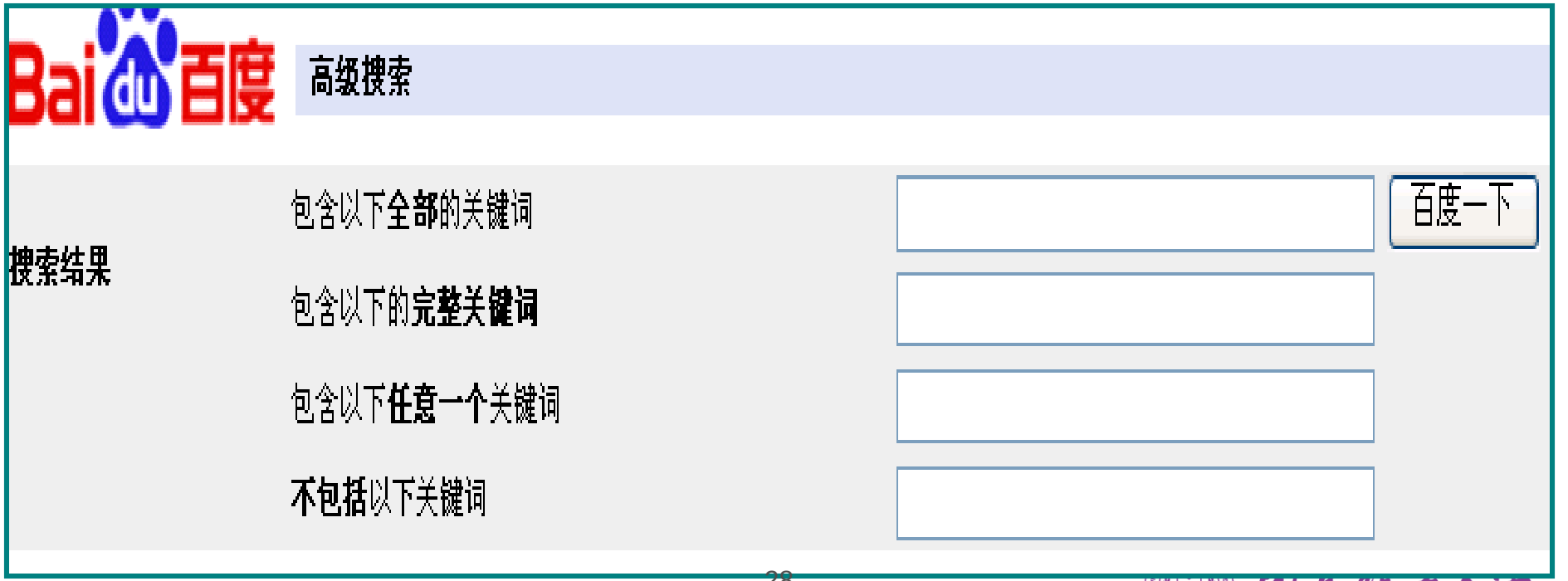
文献号	主题词							
	A	B	C	D	E	F	G	H
1	1	1	0	1	1	0	1	1
2	0	0	1	0	0	1	0	1
3	0	0	0	0	0	1	1	1
4	1	0	0	1	0	0	0	0
5	0	1	0	1	0	0	1	0
6	0	1	1	0	0	0	1	0
7	1	0	0	0	1	1	0	0
8	0	1	0	0	0	1	0	1
9	1	0	0	1	0	0	0	0
10	0	0	1	0	1	0	1	0

这10篇文献通过8个主题词来揭示其内容，即通过8个主题词的不同组合来揭示10篇文献中的不同内容

如果检索的逻辑表达式为：(A OR B) and C，则检索结果为：

如果检索的逻辑表达式为：(A AND B) OR E，则检索结果为：

南京理工大学
NANJING UNIVERSITY OF SCIENCE & TECHNOLOGY



布尔模型的应用情况

- 最早、最简单的IR模型，也是应用最广泛的模型
- 目前仍然应用于商业系统中，如开源检索系统Lucene使用了布尔（Boolean）模型
- 思考：搜索引擎是不是也用到了布尔模型？

布尔检索模型的优缺点（优势）

简单、易理解、易实现和快速检索。

布尔检索模型的优缺点（不足）

- 布尔检索是基于0-1二值判定标准的，判断的结果是：文献之间要么相关，要么不相关

- 检索结果也不能按用户定义的重要性排序输出，用户只能从头到尾浏览输出结果才能知道哪些文献更适合自己的需要

布尔检索模型的优缺点（不足）

- 对于“或”组配的提问，只包含提问式中一个标引词的文献与包含提问式中几个标引词的文献被认为是同样重要；
- 对于“与”组配的提问，包含多个标引词的文献又被认为与不包含任何标引词的文献同样不相关；

最大不足：没有考虑相关性程度

提纲

1. 信息检索模型的定义和分类
2. 布尔模型*
3. 向量空间模型*
4. 概率模型*
5. 改进的集合论检索模型
6. 改进的代数检索模型
7. 改进的概率检索模型
8. 结构化文本检索模型

3. 向量空间模型

- 向量空间模型的定义
- 常见相似度计算方法
- 向量空间模型与布尔模型的比较
- 向量空间模型一般应用步骤
- 应用示例

向量空间模型的定义

向量空间模型基本情况

- 向量空间模型由Salton等在1968年提出
- 这个模型对于查询与文档的相关度有较强的可计算性和可操作性，并且被广泛应用于文本检索、自动文摘、关键词自动提取、文本分类等方面。

向量空间模型的定义

- 该模型采用一组标引项首先将查询和待检索文档集中的文档分别表示查询向量和文档向量；
- 然后通过计算查询向量和文档向量之间的相似度，并根据求得的相似度大小对文档检索结果进行排序；
- 超过一定阈值就作为检索结果加以输出。

向量空间模型的定义

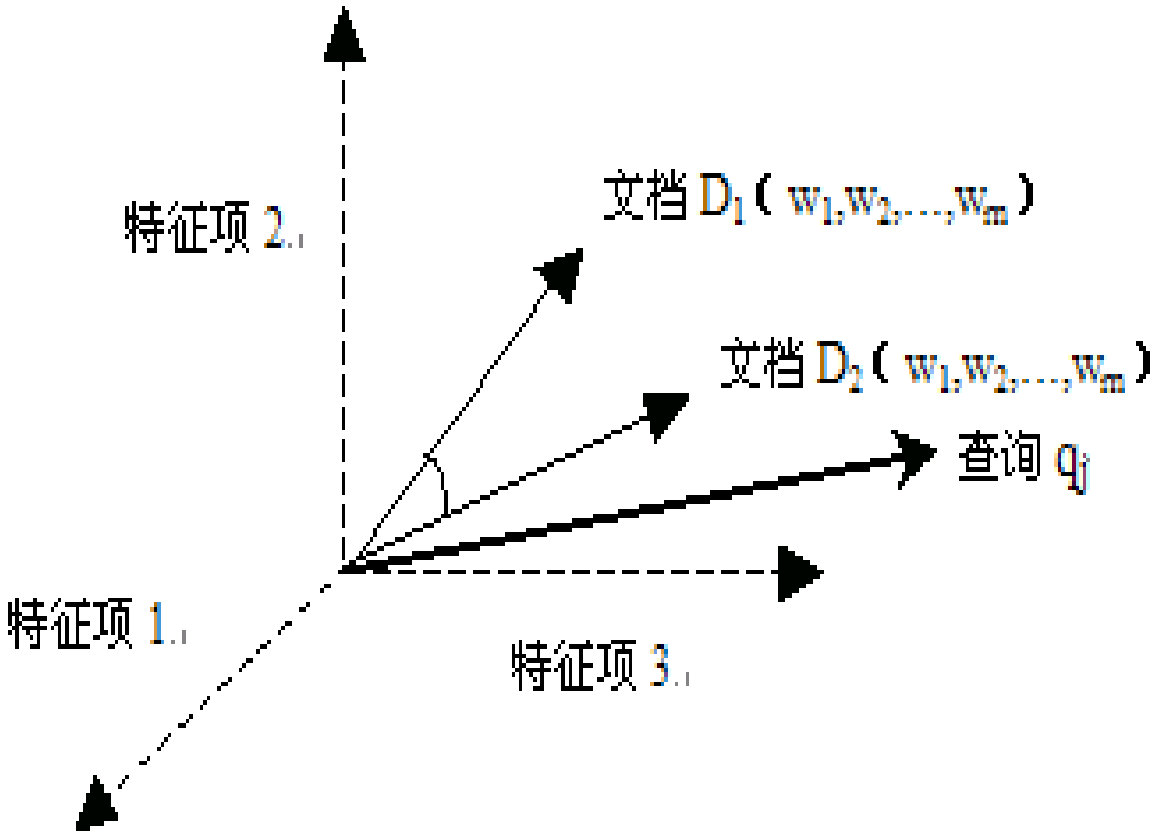
文档向量空间表示

- 对一个文档集而言，通过文档标引，可以将每一篇文档转换成特定的标引项来表示。
- 如果将每一个标引项看作是多维空间中的一维，则由这些标引项组成的集合就定义了一个多维向量空间。
- 文档集合中的任一文档都可以表示成为这一多维空间中的一个向量，这个空间就称为“文档空间”。

文档向量空间表示

1 以特征项作为文档表示的坐标，特征项可以选择字、词和词组等，表示向量中的各个分量

2用向量的形式把文档表示为多维空间中的一个点



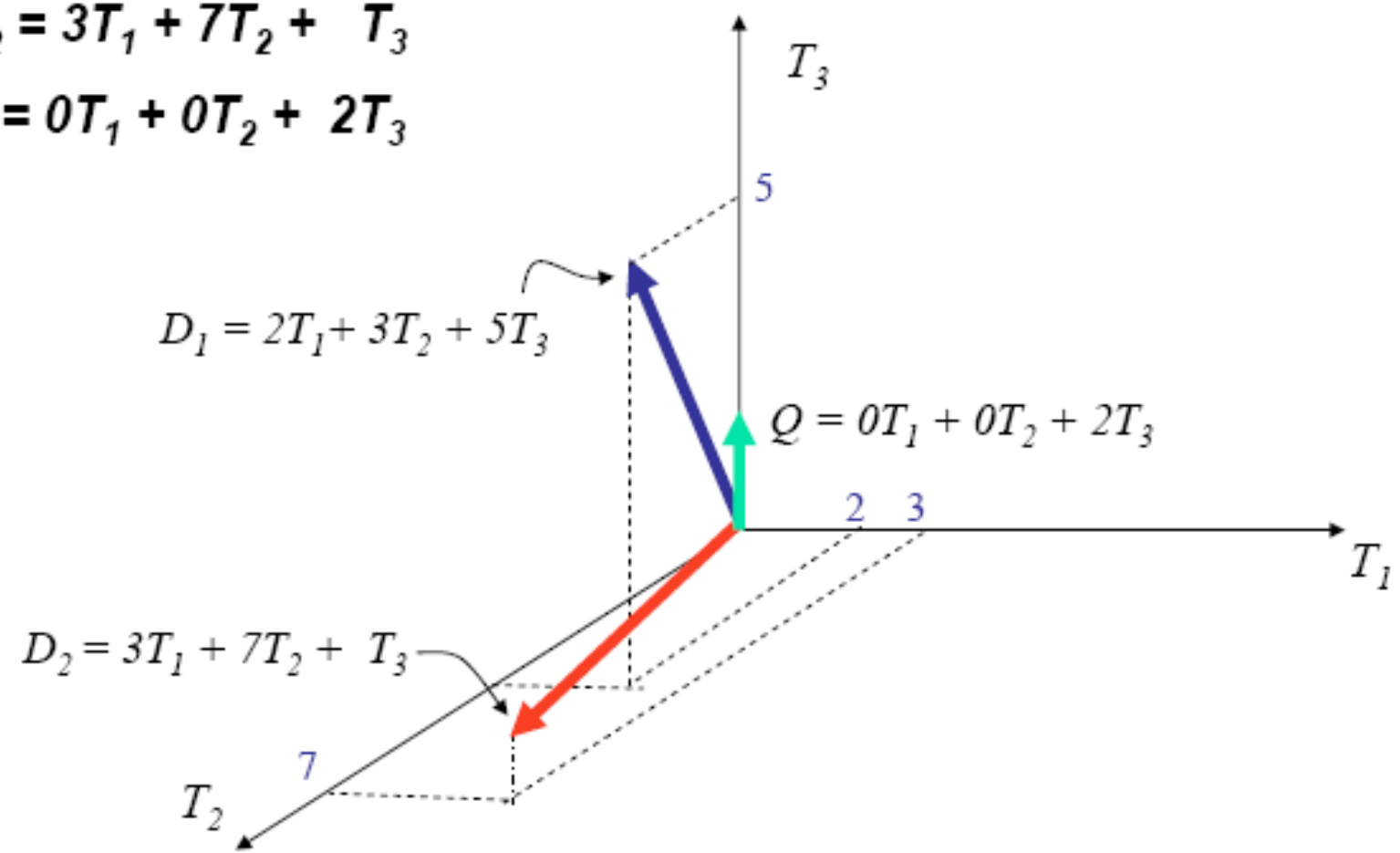
文档向量空间表示

- 文档、查询向量表示举例

$D_1 = 2T_1 + 3T_2 + 5T_3$ —— 系数表示标引项权重

$D_2 = 3T_1 + 7T_2 + T_3$

$Q = 0T_1 + 0T_2 + 2T_3$



文档向量空间表示

二值向量空间模型(向量空间模型的特例)

每篇文档都被表示成如下形式的向量：

$$D_i = (d_{i1}, d_{i2}, \cdots, d_{il}), \quad d_{ik} = 0 \text{ 或 } 1$$

其中d_{ik}则表示特征项T_k在文档D_i中的权值。如果权值为0，表示T_k没有在D_i中出现。

这种仅用0、1表示特征项权值的方法无法体现特征项在文档中的重要程度

文献号	主题词							
	A	B	C	D	E	F	G	H
1	1	1	0	1	1	0	1	1
2	0	0	1	0	0	1	0	1
3	0	0	0	0	0	1	1	1
4	1	0	0	1	0	0	0	0
5	0	1	0	1	0	0	1	0
6	0	1	1	0	0	0	1	0
7	1	0	0	0	1	1	0	0
8	0	1	0	0	0	1	0	1
9	1	0	0	1	0	0	0	0
10	0	0	1	0	1	0	1	0

文档向量空间表示

基于词频的向量空间模型：

使用更精确的词频来代替0、1二值，还是有问题？

当前一般运用TF-IDF值表示权重，即：

tf_{ik}：为特征项T_k在文档D_i中的出现频率，称为项频

df_k：文档集D中出现特征项T_k的文档数量，称为文档频率

$$w_{ik} = tf_{ik} / df_k$$

如果特征项tf_{ik}在文档D_i中的作用较大，必然有较高的项频和相对较低的文档频率，故其权重w_{ik}也较大

TF-IDF解读

- **term frequency**
- **inverse document frequency**
- **是一种统计方法，用以评估一个字词对于一个文件集或一个语料库中的其中一份文献的重要程度：**
  - 字词的重要性随着它在文件中出现的次数成正比增加；
  - 但同时会随着它在语料库中出现的频率成反比下降

TF-IDF解读

$$w = new\_tf \times new\_cf \times norm$$

- **标引项频率因子new_tf（局部权重）**
  - 主要考虑标引项在文档中出现的频次
- **文档频率因子new_cf（全局权重）**
  - 主要考虑文档集中包含该标引项的文档数量
- **规范化因子norm**
  - 通常为了抵消不同文档篇幅长短差异带来的影响

TF-IDF解读

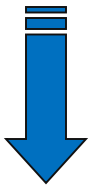
$$w_{ij} = tf_{ij} \times idf_j = tf_{ij} \times \log \frac{N}{n_j}$$

- tf_{ij}：标引项j 在文档i中的频率
- idf_j：包含标引项j 的逆文档频率
- n_j：文档集中包含标引项j 的文档数量
- N: 文档集中文档总数

文档向量空间表示

W_{ij}	K_1	K_2	…	K_n
D_1	0	1	…	0
D_2	1	0.8		0.5
…	…	…	…	…
D_m	0.2	0		1

查询和文档都被表示成了n维向量



衡量文档和查询的相关度问题转化成：计算文档向量和查询向量之间的相似度问题

常见相似度计算方法

- 向量匹配在信息检索系统中非常关键
- 检索结果是根据文档向量与查询向量间相似度匹配的结果来进行排列和输出的。
- 在信息检索系统中，希望系统提供的文档相似度排列结果能完全符合人们关于查询相关性的判断。

常见相似度计算方法

- **常用的匹配方法有：**
  - 1> 内积相似度运算（包括余弦相似度运算）
  - 2> 距离相似度运算
  - 3> 基于项匹配个数的相似度运算
  - 4> 基于概率向量的相似度运算

常见相似度计算方法

- 1> 内积相似度运算

设：$q=(w_{q1},w_{q2},\cdots,w_{qn})$　　$d_j=(w_{j1},w_{j2},\cdots,w_{jn})$

则查询向量q与文档向量d_j的内积相似度为：

$$sim(q,d_j)=q\bullet d_j=\sum_{i=1}^{n}w_{qi}\times w_{ji}$$

由于上述公式对长文档更为有利，为此作归一化处理：

$$sim(q,d_j)=\frac{q\bullet d_j}{|q|\times|d_j|}=\frac{\sum_{i=1}^{n}w_{qi}\times w_{ji}}{\sqrt{\sum_{i=1}^{n}w_{qi}^2}\sqrt{\sum_{i=1}^{n}w_{ji}^2}}$$

此时相似度演变成了q、d_j两个向量间的夹角余弦值，

通常称作余弦相似度

常见相似度计算方法

- 2> 距离相似度运算

$$L_p(q,d_j)=[\sum_{i=1}^{n}|w_{qi}-w_{di}|^p]^{\frac{1}{p}}$$

当p=1时：

$$L_p(q,d_j)=\sum_{i=1}^{n}|w_{qi}-w_{di}|$$

当p=2时：

$$L_p(q,d_j)=[\sum_{i=1}^{n}|w_{qi}-w_{di}|^2]^{\frac{1}{2}}$$

当$p\to\infty$时：

$$L_p(q,d_j)=\max\{|w_{qi}-w_{di}|\}$$

常见相似度计算方法

■ 3> 基于项匹配个数的相似度运算

✓ Dice系数

$$sim(q,d_j)=\frac{2\sum_{i=1}^{n} w_{qi}\times w_{ji}}{\sum_{i=1}^{n} w_{qi}^2+\sum_{i=1}^{n} w_{ji}^2}$$

当标引项权重选用的是二元权重设计方案时，则上式变为：

$$sim(q,d_j)=\frac{2w}{n_q+n_{d_j}}$$

其中：w是向量q与d_j间包含的相同的项个数，n_q是向量q中非零项的个数，n_{dj}是向量d_j中非零项的个数。

常见相似度计算方法

✓ Jaccard系数

$$sim(q,d_j)=\frac{\sum_{i=1}^{n} w_{qi}\times w_{ji}}{\sum_{i=1}^{n} w_{qi}^2+\sum_{i=1}^{n} w_{ji}^2-\sum_{i=1}^{n} w_{qi}\times w_{ji}}$$

当标引项权重选用的是二元权重设计方案时，则上式变为：

$$sim(q,d_j)=\frac{w}{n_q+n_{d_j}-w}$$

其中：w是向量q与d_j间包含的相同的项个数，n_q是向量q中非零项的个数，n_{dj}是向量d_j中非零项的个数。

常见相似度计算方法

■ 4> 基于概率向量的相似度运算

定义1：如果一个 n 维向量 $P=(p_1,p_2,\ldots,p_n)$ 满足

$$\sum_{i=1}^{n} p_i = 1$$

且 $p_i \geq 0$，则向量P称为概率向量。

定义2：给定一个概率向量$P=(p_1,p_2,\ldots,p_n)$，则其信息熵函数定义为：

$$H(P) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

向量空间检索模型的优点

- 根据文档和查询之间的相似度对文献进行排序，有效地提高了检索效果；
- 对标引词的权重进行了改进，从而进一步提高了检索的准确程度；
- 把文档和查询本身简化为标引词及其权重集合的向量表示，把对文档内容和查询要求的处理简化为向量空间中向量的运算；
- 可以实现文档的自动分类。

向量空间检索模型的缺点

- 标引词被认为是相互正交，太理想化，降低了检索的准确性
- 相似度的计算量大，并且当有新文档加入时，则必须重新计算词的权值

向量空间模型一般应用步骤

①对文档集进行分词；

②统计词条频率及相应文档频率，计算词条权重；

③确立标引项；

④将文档集中的文档用向量加以表示；

⑤选择相似度计算方法。

应用示例：基于向量空间模型的图书检索系统

① 需求分析：图书检索系统的流程

② 模型设计：

A. 文档集表示

B. 查询表示

C. 相似度计算

③ 结果反馈

应用示例：基于向量空间模型的图书检索系统

① 需求分析：图书检索系统的流程

A. 输入：关键词（作者、ISBN、出版商等）

B. 处理：向量空间模型

C. 输出：排序好的结果

应用示例：基于向量空间模型的图书检索系统

② 模型设计

A. 文档集表示

B. 查询表示

C. 相似度计算

A. 文档集表示

B. 查询表示

- 特征项t(Term)：书目信息的抽取，指出书目信息中能够代表图书的语言单位，即检索项。

$$D(t_1, t_2, \cdots, t_n)$$

其中，n代表了特征项的数量

- 特征项权重W_k(Weight)：指特征项t_n能够代表图书D能力的大小，体现了特征项在书目信息中的重要程度

图书的特征向量矩阵：

$$A = \begin{bmatrix} W_{11} & W_{12} & \cdots & W_{1n} \\ W_{21} & \cdots & \cdots & \\ \cdots & & & \\ \cdots & & & \\ \cdots & & & \\ W_{m1} & \cdots & \cdots & W_{mn} \end{bmatrix} = (W_{ij})_{m \times n}$$

C. 相似度计算

- **相似度S(Similarity)：指查询内容与文档的相关程度的大小**

① 通过相似度计算公式计算出每个文档向量与查询向量的相似度

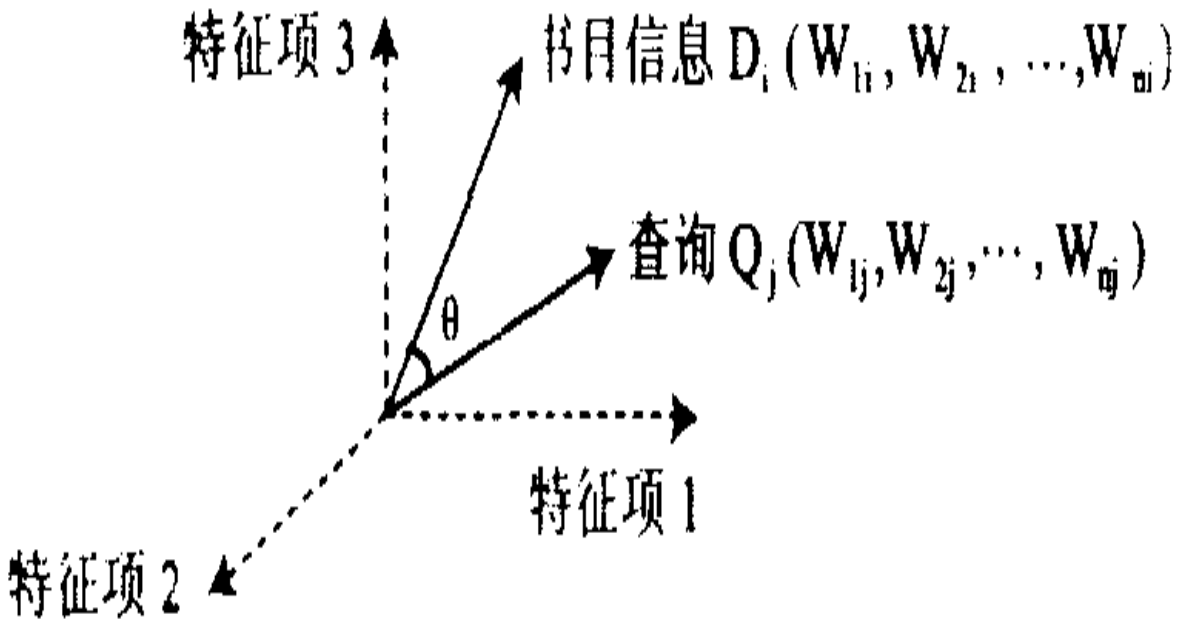
② 排序结果与设立的阈值进行比较，如果大于阈值将输出

$$Sim(D_i, Q_j) = \cos\theta = \frac{\sum_{k=1}^{n} W_{ki} \times W_{kj}}{\sqrt{\left(\sum_{k=1}^{n} W_{ki}^2\right)\left(\sum_{k=1}^{n} W_{kj}^2\right)}}$$

用户查询的向量表示：

$$Q = (q_1 , q_2 , \cdots , q_n)$$

用户查询与书目信息的相关性转化为用户检索向量与书目特征向量矩阵的夹角计算。夹角越小，相似性越大



例如：设定此检索系统中检索项为关键词。

D1 = (0.6, 0.2, 0.3, 0)

D2 = (0.5, 1, 0, 0.4)

D3 = (0.7, 0, 0.9, 0)

阈值设定为0.2。

用户输入的关键词为：“计算机应用 基础”，则：

Q = (0.7, 0.1, 0, 0)

哪些图书返回？图书的排序是什么？与文档的相似度分别是多少

提纲

1. 信息检索模型的定义和分类
2. 布尔模型*
3. 向量空间模型*
4. 概率模型*
5. 改进的集合论检索模型
6. 改进的代数检索模型
7. 改进的概率检索模型
8. 结构化文本检索模型

向量空间模型

- 信息检索的一般模型在向量空间模型中的定义
- 一般步骤
- **TF-IDF**
- 优缺点

4. 概率模型

- 概率模型的定义
- 概率模型的优缺点

概率模型的定义

- “相关性”是没有一个唯一、准确的定义。
- 主要原因是由于文档对于用户查询的相关性具有随机性和不确定性。
- 可以将一篇文档作为相关文档的可能大小是看成是一个随机事件，这样该事件的概率可以用来描述文档的相关程度。

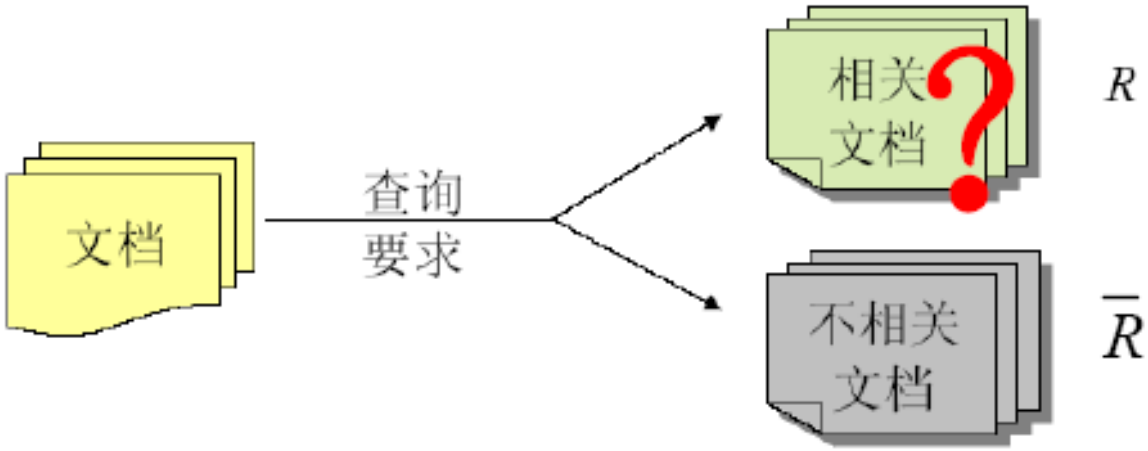
概率检索模型将文档向量与查询向量间的相关度概率化，在概率论的框架下解决信息检索的问题。

概率模型的定义

二值独立概率模型的定义

- 定义基本假设前提
  - 文档对查询的相关性与文档集合中的其它文档无关，这点被称为**概率模型的相关性独立**原则；
  - 文档和查询中的**标引项与标引项之间是相互独立的**；
  - 文档和查询中的标引项权重都是**二值**的，即要么是0，要么是1；
  - 文档相关性是**二值**的，即只有相关和不相关两种，也就是说，一篇文档要么属于理想相关文档集，要么不属于理想相关文档集

二值独立概率模型的定义



$$P(A|B)=\frac{P(A,B)}{P(B)}=\frac{P(B|A)P(A)}{P(B)}$$

对每个Q定义排序(Ranking)函数RSV(Q,D)：

$$\log\frac{P(R=1|D)}{P(R=0|D)}=\log\frac{P(D|R=1)P(R=1)/P(D)}{P(D|R=0)P(R=0)/P(D)}$$

$$\propto\log\frac{P(D|R=1)}{P(D|R=0)}$$

对同一Q是常量，对排序不起作用

BIM模型通过Bayes公式对所求条件概率$P(R=1|Q,D)$开进行计算。BIM是一种生成式(generative)模型

对于同一Q，$P(R=1|Q,D)$可以简记为$P(R=1|D)$

其中，$P(D|R=1)$、$P(D|R=0)$分别表示在相关和不相关情况下生成文档D的概率。Ranking函数显然是随着$P(R=1|D)$的增长而增长。

一个例子

- 查询为：信息　检索　教程
- 所有Term的概率为:

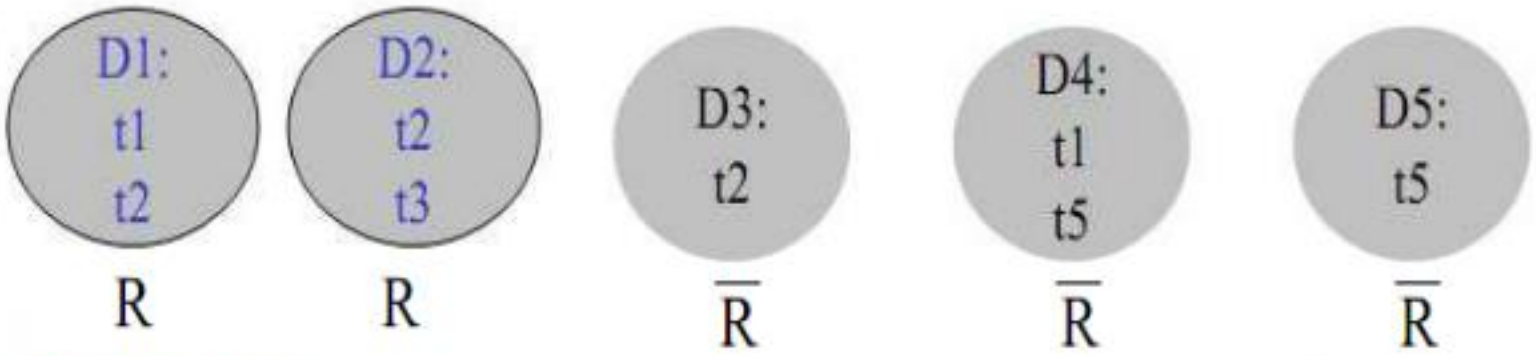
Term	信息	检索	教材	教程	课件
R=1	0.8	0.9	0.3	0.32	0.15
R=0	0.3	0.1	0.35	0.33	0.10

- 文档D1：检索　课件
  - P(D1|R=1)=(1-0.8)*0.9*(1-0.3)*(1-0.32)*0.15
  - P(D1|R=0)= (1-0.3)*0.1*(1-0.35)*(1-0.33)*0.10
  - P(D1|R=1)/P(D1|R=0)=4.216

一个例子

简单词项权重

- 估计给定词项在相关文档中的概率



- 假设D1和D2是相关文档，D3、D4和D5是非相关文档

一个例子

- 查询为： q = 计算机　硬件
- 所有Term的概率为:

Term	计算机	硬件	电脑	软件	笔记本
R=1	0.8	0.9	0.7	0.3	0.6
R=0	0.3	0.2	0.8	0.7	0.4

- 文档D1 = 电脑　软件
  - Sim (D1, q) = ?
- 文档D2 = 笔记本　硬件
  - Sim (D2, q) = ?

概率模型的优缺点

优点：

采用了相关反馈原理克服了不确定性推理的缺点，将文档按相关的概率降序排列

缺点：

①参数估计的难度较大；

②标引词K_i在文档中的权重是二值的(要么出现，要么不出现)，没有考虑标引词K_i在文档中出现的概率；

③文档和查询的表达也比较困难。

文献号	主题词							
	A	B	C	D	E	F	G	H
1	1	1	0	1	1	0	1	1
2	0	0	1	0	0	1	0	1
3	0	0	0	0	0	1	1	1
4	1	0	0	1	0	0	0	0
5	0	1	0	1	0	0	1	0
6	1	1	1	1	0	0	1	0
7	1	0	0	0	1	1	0	0
8	0	1	0	0	0	1	0	1
9	1	0	0	1	0	0	0	0
10	0	1	0	1	1	0	1	0

这10篇文献通过8个主题词来揭示其内容，即通过8个主题词的不同组合来揭示10篇文献中的不同内容

如果检索的逻辑表达式为：(A and B) or D

1、布尔模型的检索结果是什么？

2、向量空间模型的结果是什么？

文献号	主题词							
	A	B	C	D	E	F	G	H
1	10	70	0	60	10	0	20	10
2	0	0	1	0	0	1	0	1
3	0	0	0	0	0	1	1	1
4	10	80	0	30	0	0	0	0
5	0	10	0	10	0	0	10	0
6	20	50	10	10	0	0	10	0
7	10	40	0	10	10	10	0	0
8	0	30	0	0	0	10	0	10
9	30	90	0	40	0	0	0	0
10	0	10	0	0	10	0	10	0

这10篇文献通过8个主题词来揭示其内容，即通过8个主题词的不同组合来揭示10篇文献中的不同内容

如果检索的逻辑表达式为：(A and B) or D

1、关键词在每篇文档中的重要程度怎么表示？

2、计算关键词在每篇文档中的重要程度？

2、向量空间模型的结果是什么？

该检索系统中检索项为关键词，依次分别：信息、信息组织、信息检索、分类。每篇文档通过概率模型表示为：

D1 = (0.6, 0.2, 0.3, 0)

D2 = (0.5, 1, 0, 0.4)

D3 = (0.7, 0, 0.9, 0)

用户需要查询与信息组织相关的内容，请确定检索关键词；查询的表示；结果的排序。

网络环境下检索模型的现状

最简单的检索系统只需要按照查询词之间的逻辑关系返回相应的文档就可以了，但这种做法不能表达结果与查询之间的深层关系。

目前有两大主流技术用于分析结果和查询的相关性：链接分析和基于内容的计算。

为了把最符合用户需求的结果显示在前面，还需要利用各种信息对结果进行重排序。

- **链接分析**：许多研究者发现，WWW上超链结构是个非常丰富和重要的资源，如果能够充分利用的话，可以极大地提高检索结果的质量

Sergey Brin 和Larry Page 在1998年提出了PageRank算法，同年J.Kleinberg 提出了HITS算法，其它一些学者也相继提出了另外的链接分析算法，如SALSA，PHITS，Bayesian等算法。

PageRank示例

Google的网页排序

- 在Google中搜索“体育新闻”

南京理工大学
NANJING UNIVERSITY OF SCIENCE & TECHNOLOGY

Google的网页排序

查询词和文档的相关性

- 在Google中搜索“体育新闻”
  - 搜索引擎工作的简要过程如下
  - 针对查询词“体育新闻”进行分词——》“体育”、“新闻”
  - 根据建立的倒排索引，将同时包含“体育”和“新闻”的文档返回，并根据相关性进行排序
    - 这里的相关性主要是基于内容的相关性
    - 但是会有一些垃圾网页，虽然也包含大量的查询词，但却并非满足用户需要的文档，如下图，一个网页中虽然出现了四次“体育新闻”但却不是用户所需要的
    - 因此，页面本身的重要性在网页排序中也起着很重要的作用

Google的网页排序

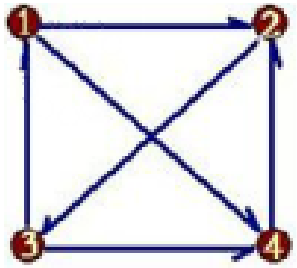
■在Google中搜索“体育新闻”

Tag:	ESPN STAR NBA 官方网站 体育新闻 欧冠直播 英超 意甲直播 西甲 德甲 篮球 比分直播 AC米兰
域名：	nczb.com.cn
标题：	南国早报网首页 南宁新闻,广西新闻,国内新闻,国际新闻,体育新闻,娱乐新闻,财经新闻,房产新闻,汽车新闻,图库,校园新闻,健康新闻
简介：	...
Tag:	南宁新闻 广西新闻 国内新闻 国际新闻 体育新闻 娱乐新闻 财经新闻 房产新闻 汽车新闻 图库 校园新闻 健康新闻
域名：	trade-114.cn
标题：	资讯网
简介：	专业的行业资讯网站，提高各种行业资讯,国内新闻，时事新闻，财经新闻，军事新闻，体育新闻，电子行业咨询等及时资讯网站。...
Tag:	行业资讯 国内新闻 时事新闻 财经新闻 军事新闻 体育新闻 电子行业咨询
域名：	dcrb.cn
标题：	丹东日报——首页
简介：	丹东日报 社会新闻、经济新闻、国内新闻、国际新闻、体育新闻、文娱新闻 ..
Tag:	丹东日报 社会新闻 经济新闻 国内新闻 国际新闻 体育新闻 文娱新闻

Google的网页排序

- 如何度量网页本身的重要性呢？

  互联网上的每一篇html文档除了包含文本、图片、视频等信息外，还包含了大量的链接关系，利用这些链接关系，能够发现某些重要的网页



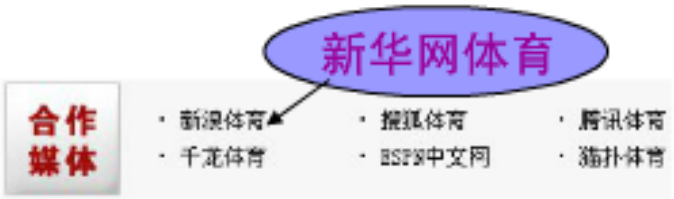
网页是节点，网页间的链接关系是边

- 直观地看，某网页A链向网页B，则可以认为网页A觉得网页B有链接价值，是比较重要的网页。

  某网页被指向的次数越多，则它的重要性越高；越是重要的网页，所链接的网页的重要性也越高。

Google的网页排序

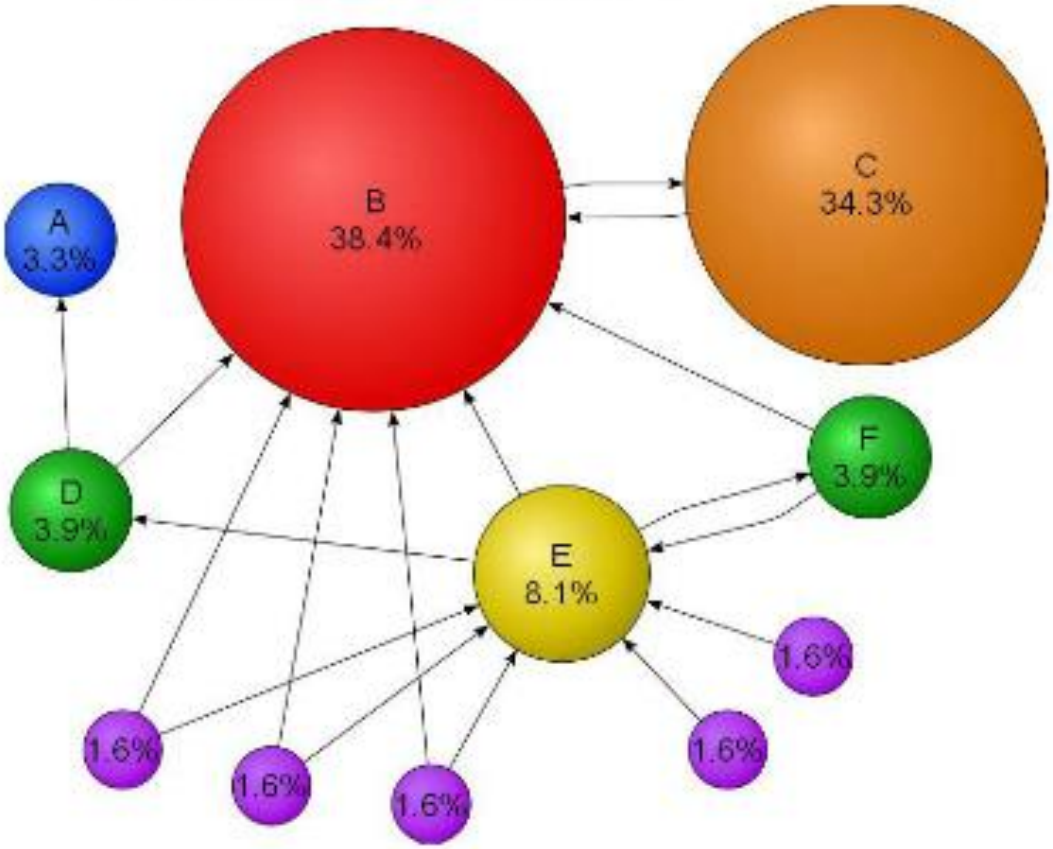
- 如何度量网页本身的重要性呢？
  - 比如，新华网体育在其首页中对新浪体育做了链接，人民网体育同样在其首页中对新浪体育做了链接



  - 可见，新浪体育被链接的次数较多；同时，人民网体育和新华网体育也都是比较“重要”的网页，因此新浪体育也应该是比较“重要”的网页。

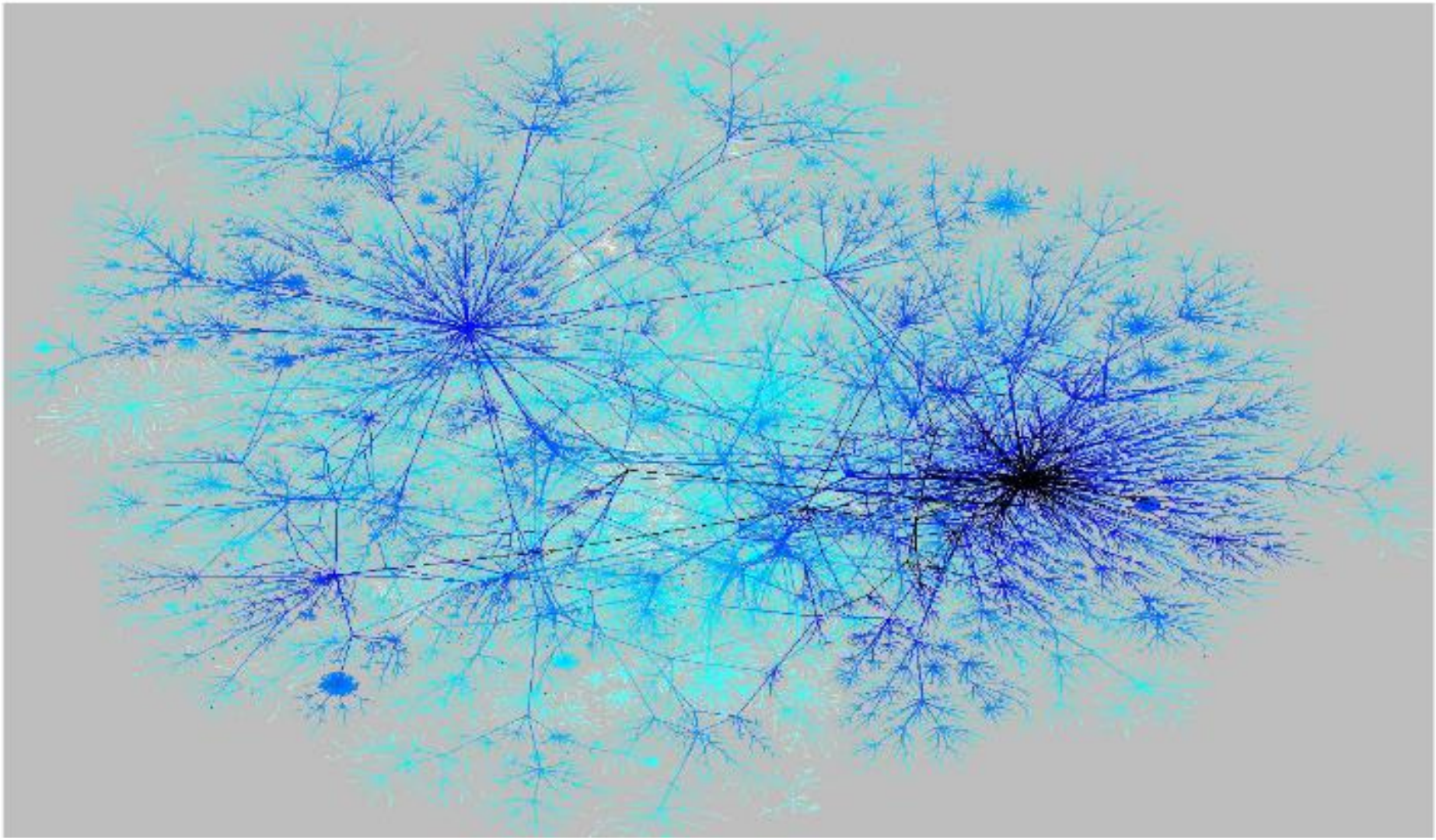
Google的网页排序

- 一个更加形象的图



链向网页E的链接远远多于链向网页C的链接，但是网页C的重要性却大于网页E。这是因为因为网页C被网页B所链接，而网页B有很高的重要性。

Http网页链接示意图



什么是PageRank

PageRank是一种在搜索引擎中根据网页之间相互的链接关系计算网页排名的技术。

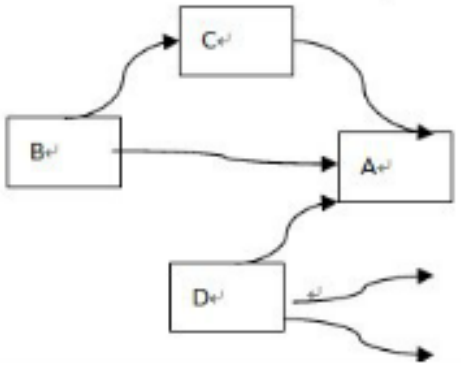
PageRank是Google用来标识网页的等级或重要性的一种方法。其级别从1到10级，PR值越高说明该网页越受欢迎（越重要）。

PageRank近似于一个用户，是指在Internet上随机地单击链接将会到达特定网页的可能性。通常，能够从更多地方到达的网页更为重要，因此具有更高的PageRank。

如果要查看此站点PageRank值，请安装GOOGLE工具条并启用PageRank特性，或者在firefox安装SearchStatus插件。

Pagerank算法原理：

PageRank简单计算：



- 假设一个由只有**4**个页面组成的集合：**A**，**B**，**C**和**D**。如果所有页面都链向**A**，那么**A**的**PR**（**PageRank**）值将是**B**，**C**及**D**的和。

$$PR(A) = PR(B) + PR(C) + PR(D)$$

- 继续假设**B**也有链接到**C**，并且**D**也有链接到包括**A**的**3**个页面。一个页面不能投票**2**次。所以**B**给每个页面半票。以同样的逻辑，**D**投出的票只有三分之一算到了**A**的**PageRank**上。

$$PR(A) = \frac{PR(B)}{2} + \frac{PR(C)}{1} + \frac{PR(D)}{3}$$

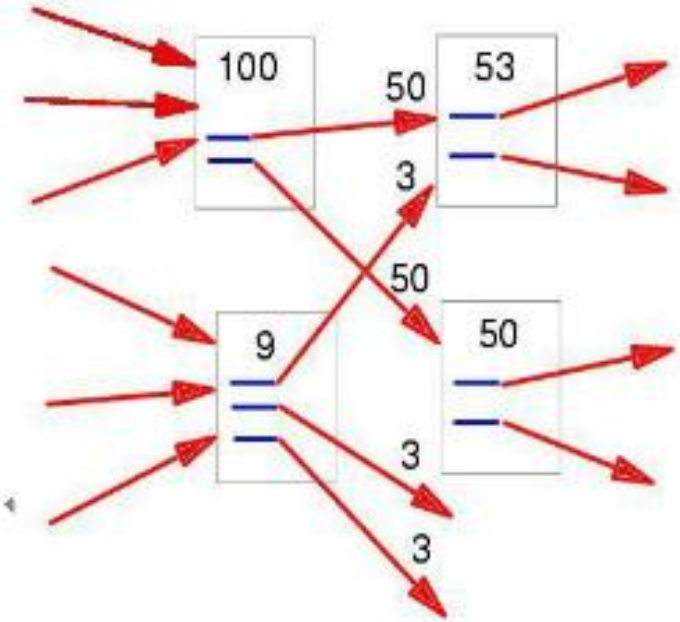
- 换句话说，根据链出总数平分一个页面的**PR**值。

$$PR(A) = \frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)}$$

PageRank的简化模型

可以把互联网上的各网页之间的链接关系看成一个有向图。假设冲浪者浏览的下一个网页链接来自于当前网页。建立简化模型：对于任意网页Pi，它的PageRank值可表示为如下：其中Bi为所有链接到网页i的网页集合，Lj为网页j的对外链接数（出度）。

$$PR_i = \sum_{j \in B_i} \frac{PR_j}{L_j}$$



PR_i：网页i的pagerank值

PR_j：网页j的pagerak值

L_j：网页j链出的连接数

B_i：链接到网页i的网页集合

基于内容的计算

沿用传统的文本分类方法，多采用**向量空间模型**、**概率模型**等方法来逐一计算用户查询和结果的**相似度**（相关性）

比较

- 链接分析充分利用了Web上丰富的链接结构信息，但它很少考虑网页本身的内容；
- 基于内容的计算则较为深入地揭示了查询和结果之间的语义关系，但忽略了不同网页之间的指向关系

因此现在很多系统尝试把两者结合起来，以达到更好的效果

提纲

1. 信息检索模型的定义和分类
2. 布尔模型*
3. 向量空间模型*
4. 概率模型*
5. 改进的集合论检索模型
6. 改进的代数检索模型
7. 改进的概率检索模型
8. 结构化文本检索模型

5. 改进的集合论检索模型

- 扩展布尔检索模型
- 模糊检索模型

扩展布尔检索模型

扩展布尔检索模型是由Salton于1983年提出，它**将向量检索模型与布尔检索模型融为一体**，并克服了传统布尔模型的一些缺陷。

扩展布尔检索模型

（1）原理

设文本集中每篇文本仅由两个标引词t_1和t_2标引，并且t_1、t_2允许赋以权值，其权值范围为[0,1]，权值越接近1，说明该词越能反映文本的内容，反之，越不能反映文本的内容。

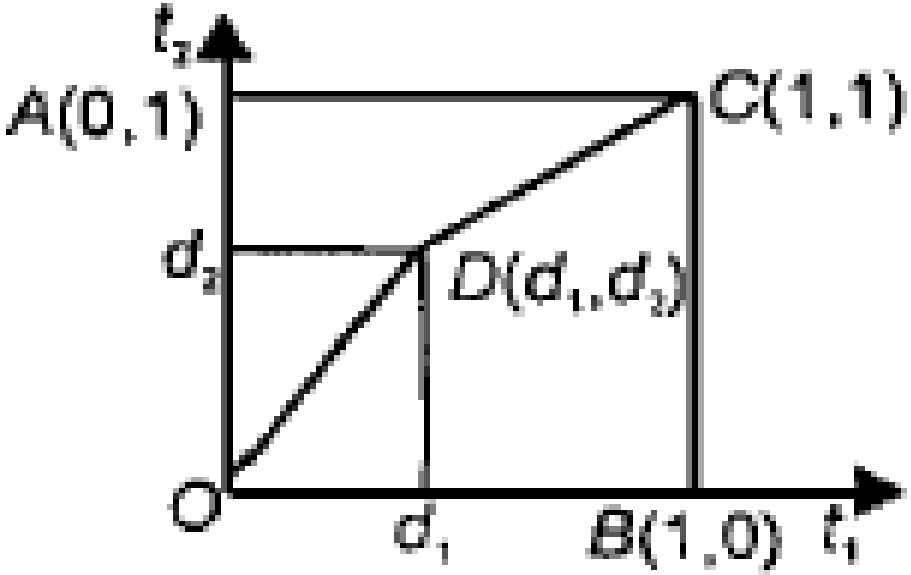


图 扩展布尔逻辑的矢量表示

扩展布尔检索模型

- 对于由t_1和t_2构成的检索式$q = t_1 \vee t_2$：

对于某一文本D来说，当D点离A、B、C 3点越接近时说明相似度越大。因而D与O的距离可以作为衡量一文本与查询q相关程度的一个尺度：

$$|DO| = \sqrt{(d_1 - 0)^2 + (d_2 - 0)^2} = \sqrt{d_1^2 + d_2^2}$$

扩展布尔检索模型

显然 $0 \leqslant |DO| \leqslant \sqrt{2}$ 为了使相似度控制在0 与1之间，将相似度定义为：

$$\sin(D, Q(t_1 \vee t_2)) = \sqrt{\frac{d_1^2 + d_2^2}{2}}$$

扩展布尔检索模型

- 对于由t_1 和t_2 构成的查询$q = t_1 \wedge t_2$:

  只有C点才是最理想的文本，用D与C的距离作为我们衡量一文本与查询q的相关程度的一个尺度

$$|DC| = \sqrt{(1 - d_1)^2 + (1 - d_2)^2}$$

于是，可把相似度定义为:

$$\sin(D, Q(t_1 \wedge t_2)) = 1 - \sqrt{\frac{(1 - d_1)^2 + (1 - d_2)^2}{2}}$$

扩展布尔检索模型

- 推广到对检索标引词进行加权的情形

  设检索标引词t_1、t_2的权值分别为a、b，且$0 \le a, b \le 1$，则上式可进一步推广为：

$$\sin(d, Q(t_1, a) \vee (t_2, b)) = \sqrt{\frac{a^2 d_1^2 + b^2 d_2^2}{a^2 + b^2}}$$

$$\sin(d, Q(t_1, a) \wedge (t_2, b)) = 1 - \sqrt{\frac{a^2 (1 - d_1)^2 + b^2 (1 - d_2)^2}{a^2 + b^2}}$$

扩展布尔检索模型

n 个标引词时的相似度计算公式

设$\boldsymbol{d} = (d_1, d_2, \cdots, d_n)$，其中$d_i$表示第$i$个标引词$t_i$的权值，$0 \le d_i \le 1$。由布尔运算符“$\vee$”及“$\wedge$”所确定的检索式分别为：

$$Q_{\vee(p)} = (t_1, a_1) \vee (t_2, a_2) \vee \cdots \vee (t_n, a_n)$$

$$Q_{\wedge(p)} = (t_1, a_1) \wedge (t_2, a_2) \wedge \cdots \wedge (t_n, a_n)$$

扩展布尔检索模型

- 在n个标引词生成的n维欧氏空间中应用L_P矢量模公式进行欧氏模的计算，将文本和查询的相似度定义为：

$$\mathrm{sim}(d, Q_{\vee(p)}) = \left[\frac{a_1^p d_1^p + a_2^p d_2^p + \cdots + a_n^p d_n^p}{a_1^p + a_2^p + \cdots + a_n^p}\right]^{\frac{1}{p}},$$

$$\mathrm{sim}(d, Q_{\wedge(p)}) = 1 - \left[\frac{a_1^p (1 - d_1)^p + a_2^p (1 - d_2)^p + \cdots + a_n^p (1 - d_n)^p}{a_1^p + a_2^p + \cdots + a_n^p}\right]$$

模糊检索模型

(1) 马尔可夫链在图书情报管理预测中的应用

□马尔可夫链的含义

在给定当前知识或信息的情况下，过去（即当前以前的历史状态）对于预测将来（即当前以后的未来状态）是无关的。

矩阵$(P_{ij})_{m\times m}$称为转移概率矩阵

马尔可夫链

例：天气预报问题

如果明天是否有雨仅与今日是否有雨有关，而与过去的天气无关. 并设今日下雨，明日有雨概率为**0.7**，今日无雨明日有雨的概率为**0.4**，并把有雨称为**0**状态，无雨称为**1**状态。则问：今日有雨且第**5**日仍有雨的概率为多少？

马尔可夫链

解：设状态**0**代表有雨，状态**1**代表无雨，
则一步转移矩阵为：

$$P=\begin{pmatrix} P_{00} & P_{01} \\ P_{10} & P_{11} \end{pmatrix}=\begin{pmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{pmatrix}$$

$$P^{(4)}=P^4=\begin{pmatrix} P_{00} & P_{01} \\ P_{10} & P_{11} \end{pmatrix}^4=\begin{pmatrix} 0.5749 & 0.4251 \\ 0.5668 & 0.4332 \end{pmatrix}=\begin{pmatrix} P_{00}^{(4)} & P_{01}^{(4)} \\ P_{10}^{(4)} & P_{11}^{(4)} \end{pmatrix}$$

所以今天有雨，第**5**天有雨的概率为：

$$P_{00}^{(4)}=0.5749$$

模糊检索模型

在实际问题中，转移概率是通过统计t_1，t_2，…，t_n时刻状态转移频率来加以确定的，即：

$$P_{ij}=f_{ij}/f_{i.}$$

其中，P_{ij}为转移概率矩阵元素，f_{ij}为频数矩阵元素，$f_{i.}$为频数矩阵行元素之和。

例：我院图书资料室的阅览室连续20个晚上开放时间内接待读者人数的统计数据如下：

45, 46, 55, 50, 72, 103, 79, 93, 51, 44, 47, 49, 51, 63, 70, 68, 58, 50, 53, 78

第1天　第2天　第20天

到阅览室人数为44~103人，将这20个数据分成以下3种状态，试推算第21天的工作状态：

读者人数	工作状态
X<50	较闲(S_1)
50≤x≤78	一般(S_2)
X>78	较忙(S_3)

模糊检索模型

记f_{ij}(i，j=1，2，3)表示从状态S_i转移到S_j的次数(频数)，则有：

$$f_{11}=3，f_{12}=2，f_{13}=0$$
$$f_{21}=1，f_{22}=9，f_{23}=1$$
$$f_{31}=0，f_{32}=1，f_{33}=2$$

由上式可计算转移概率矩阵为：P=？

$$P=\begin{bmatrix} 0.6 & 0.4 & 0 \\ 0.09 & 0.82 & 0.09 \\ 0 & 0.33 & 0.67 \end{bmatrix}$$

模糊检索模型

根据目前(第20天)的读者人数x=78，属于S_2状态，可以推知：第21天的工作人员工作状态为？

第21天的读者人数保持在50≤x≤78可能性最大（82%），转移至S_1或S_3状态的可能性极小

存在问题：

如果第20天的读者人数为79人，属S_3状态，其它条件均无变化，则有转移概率矩阵P为：

$$P=\begin{bmatrix} 0.6 & 0.40 & 0 \\ 0.09 & 0.73 & 0.18 \\ 0 & 0.33 & 0.67 \end{bmatrix}$$

由于第20天的读者人数为79人，属S_3状态，可推测：

第21天的读者人数在x>78，即S_3状态的可能性较大，于是可见：

模糊检索模型

上述预测对第20天的值依赖性很大，一个微小的扰动而导致预测的结果大不相同，显然是不符合实际情况的。不合理的原因：

S_1、S_2、S_3的划分太清晰，即：

x=49，则x∈S_1，x=50，则x∈S_2，

x=78，则x∈S_2，x=79，则x∈S_3。

可用模糊理论来解决这个问题

模糊检索模型

(2) 模糊集合理论的含义

“模糊集合(Fuzzy Set)”是美国自动控制专家扎得(L. A. Zadeh)提出来的，其出发点是用“隶属函数”的概念来描述差异的中间过渡

模糊集合理论处理的是边界不明确的集合的表示，其中心思想是把集合中的元素和隶属函数结合在一起。隶属函数的取值在[0,1]上，0表示元素不隶属于该集合，1表示完全隶属于该集合

定义：给定论域U，U的模糊子集A可以定义为U到闭区间[0,1] 上的一个映射：

μ_A：$U \rightarrow [0,1]$，μ_A为A的隶属度。

模糊检索模型

(3) 对上述图书管理预测情况的解决方法

设上述马尔柯夫链的随机变量的取值范围为U，建立U上的模糊状态集

$$S_1、S_2、\cdots，S_E,$$

对任意的u值，u∈U，有

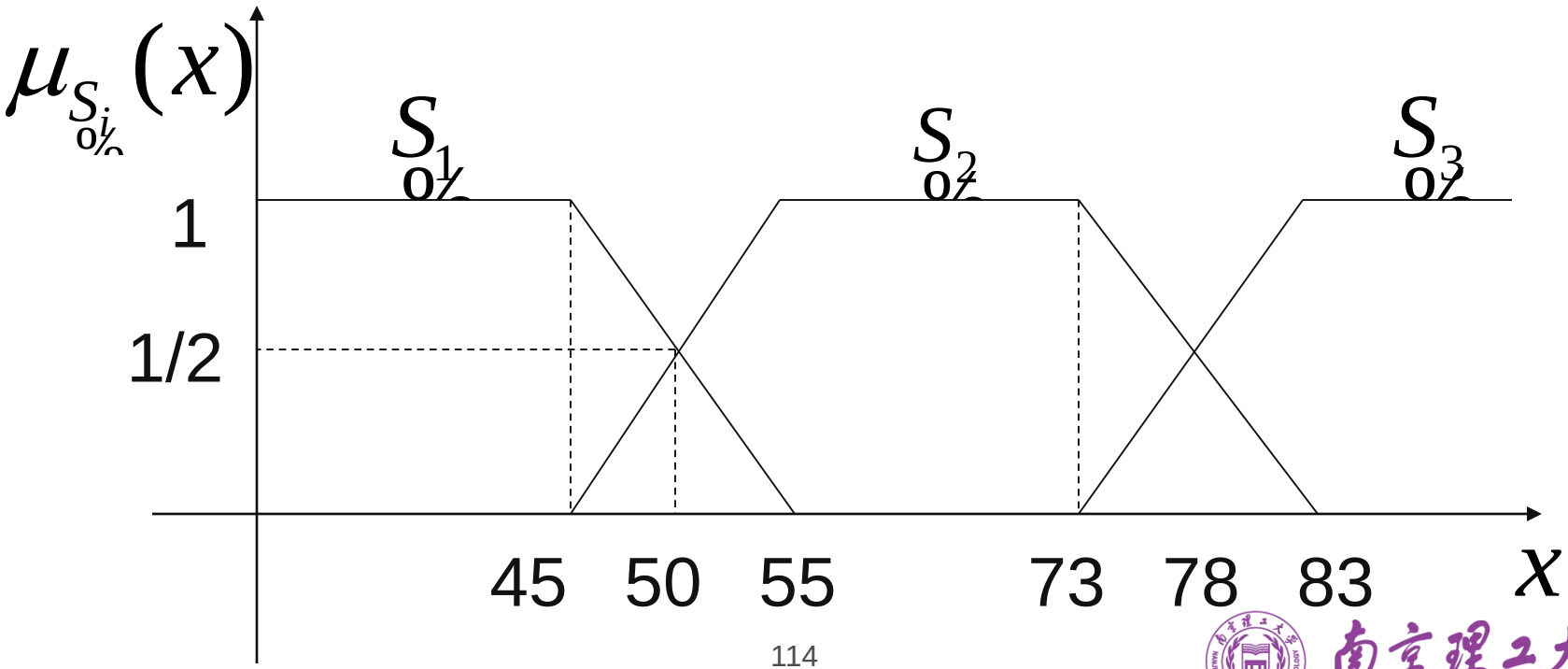
$$\sum_{e=1}^{E} \mu_{S_e}(u) = 1$$

则称 $\mu_{S_e}(u)$ 为u 对模糊状态集S_e的隶属度。

模糊检索模型

- 对上例，在实数域R上建立模糊状态 S_1，S_2，S_3 确定随机变量对各模糊状态的隶属度的分布情况

  假设用梯形图来表示隶属度分布：



$$\mu_1(x)=\begin{cases}1 & x<45\\ (55-x)/10 & 45\le x\le 55\\ 0 & x>55\end{cases}$$

$$\mu_2(x)=\begin{cases}(x-45)/10 & 45\le x\le 55\\ 1 & 55<x\le 73\\ (83-x)/10 & 73<x\le 83\\ 0 & \text{其余}\end{cases}$$

$$\mu_3(x)=\begin{cases}0 & x\le 73\\ (x-73)/10 & 73<x\le 83\\ 1 & x>83\end{cases}$$

各模糊状态分配系数

日期	读者	μ_1	μ_2	μ_3	日期	读者	μ_1	μ_2	μ_3
1	45	1.0	0	0	11	47	0.8	0.2	0
2	46	0.9	0.1	0	12	49	0.6	0.4	0
3	55	0.0	1	0	13	51	0.4	0.6	0
4	50	0.5	0.5	0	14	63	0	1.0	0
5	72	0	1	0	15	70	0	1	0
6	103	0	0	1	16	68	0	1	0
7	79	0	0.4	0.6	17	58	0	1	0
8	93	0	0	1	18	50	0.5	0.5	0
9	51	0.4	0.6	0	19	53	0.2	0.8	0
10	44	1	0	0	20	78	0	0.5	0.5

模糊检索模型

- 构建模糊概率矩阵

定义 $\mu_{S_i}[\xi(t_k)]g\mu_{S_j}[\xi(t_{k+1})]$ 为时刻t_{k+1}时状态S_i到S_j的模糊状态转移矩阵系数，且称：

$$F_{ij}=\sum_{k=1}^{n-1}\mu_{S_i}[\xi(t_k)]g\mu_{S_j}[\xi(t_{k+1})]$$

为状态S_i到状态S_j的模糊转移频数。

$$P_{ij}=F_{ij}/F_{i\cdot}$$

为状态S_i到状态S_j的模糊转移概率。

$$P=(P_{ij})_{m\times m}$$

为模糊矩阵。 对上例，计算得：

模糊检索模型

$$F_{11} = 2.92, F_{12} = 3.28, F_{13} = 0.1$$

$$F_{21} = 1.98, F_{22} = 6.32, F_{23} = 1.8$$

$$F_{31} = 0.4, F_{32} = 1, F_{33} = 1.2$$

得模糊转移矩阵：

$$P = \begin{bmatrix} 0.46 & 0.52 & 0.02 \\ 0.19 & 0.63 & 0.18 \\ 0.15 & 0.39 & 0.46 \end{bmatrix}$$

模糊检索模型

预测

因为在t_n时刻$\xi(t_n)$ 的取值分属于各个模糊集合，故对t_{n+1}时刻的状态的预测要通过隶属度模糊向量和转移概率矩阵的乘积来实现。

第20天的读者人数x=78，属于S_2的隶属度为0.5，属于S_3的隶属度0.5，得模糊向量（0，0.5，0.5），因此，第21天读者人数属于3种状态的概率为：

（0, 0.5, 0.5）·$P_{o/}$ =(0.17, 0.51, 0.32)

第2种状态的可能性最大。

模糊检索模型

如果第20天人数79，其余条件不变，利用类似的方法，可得第21天预测结果为

（0.166，0.486，0.348）

所以还是属于S_2的可能性最大

模糊检索模型

(4) 模糊检索

① 模糊检索与传统信息检索的比较

- **传统信息检索**

模型简明，便于理解

在关系模型中，数据信息被组织成若干张二维表格的结构

模糊检索模型

例如，为了描述某单位职工的基本工资情况，制成关系表格

职工号	姓名	性别	年龄	基本工资
101	曾华	男	25	410.6
102	匡明华	男	32	450.8
103	王丽美	女	65	685.0
104	李林	男	21	380.0
105	王芳	女	56	605.0

模糊检索模型

如果要检索“年龄不大于45岁，基本工资在550元以下的男职工情况”，用SQL语言描述检索要求，则有：

```
SELECT * FROM 职工基本工资情况表
WHERE 性别=“男” AND 年龄≤45 AND 基本工资＜550.0
```

如要检索：

“年龄有点老，基本工资在600元左右的职工情况”

模糊词汇

不能简单用基于布尔检索的关系数据库模型来实现

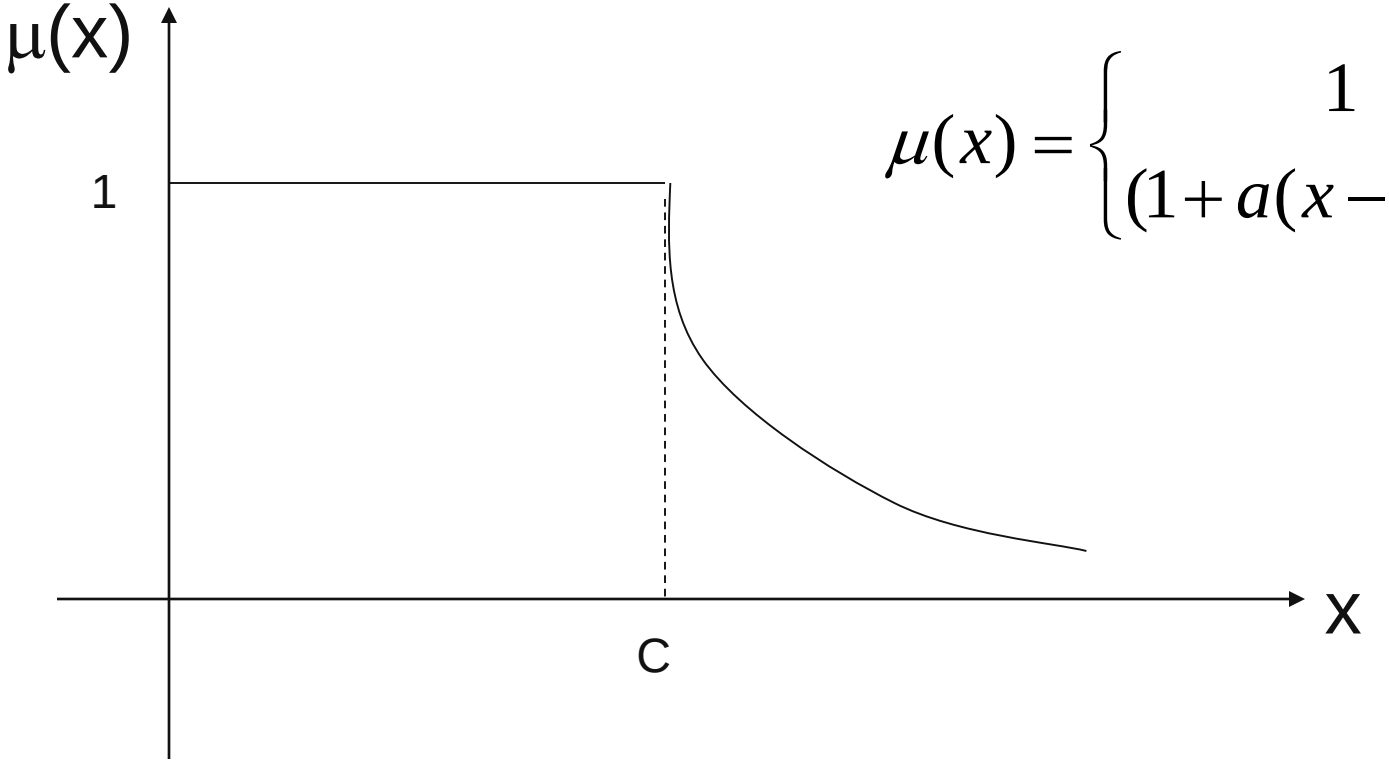
模糊检索模型

模糊检索

在日常生活中人们常常用模糊词汇“年轻”、“中年”、“年老”来描述一个人的年龄情况，假设在充分考虑年龄段间衔接的前提下，分别利用下降型、中间型、上升型分布函数给出描述年龄的各模糊词汇的隶属函数表示

几种主要的描述模糊词汇的分布函数

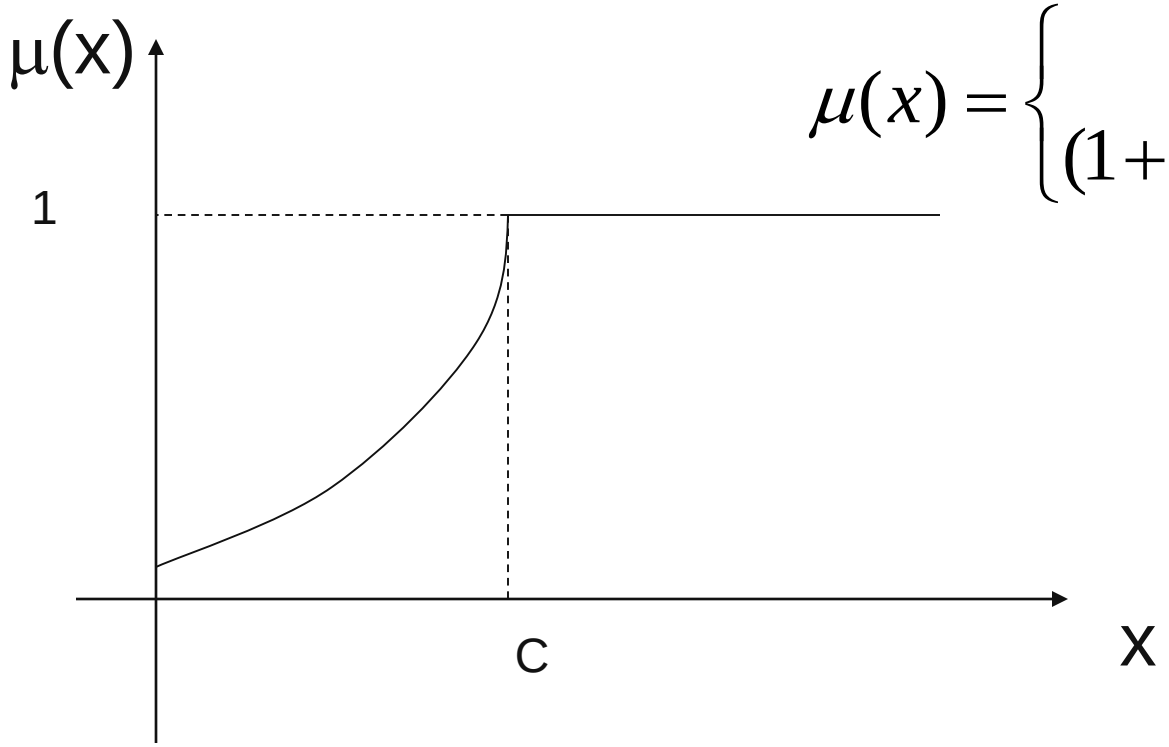
左大右小的偏小型下降函数



$$\mu(x)=\begin{cases}1 & x\le c\\ (1+a(x-c)^{b})^{-1} & x>c\end{cases}$$

模糊检索模型

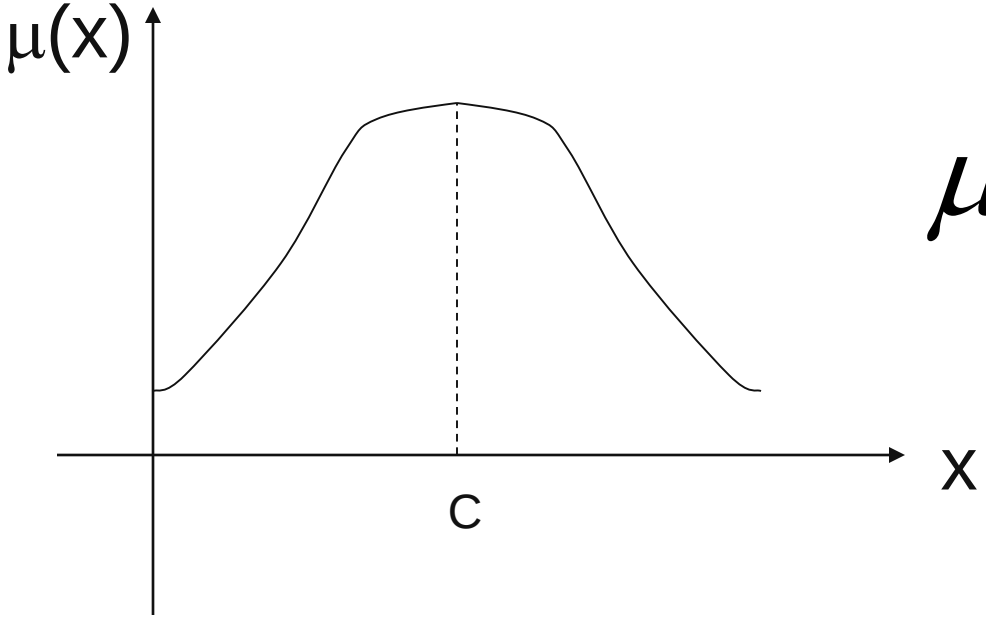
右大左小的偏大型上升函数



$$\mu(x)=\begin{cases}1 & x\geq c\\ (1+a(x-c)^{b})^{-1} & x<c\end{cases}$$

模糊检索模型

对称中间型正态凸函数



$$\mu(x) = e^{-k(x-c)^2}$$

模糊检索模型

假设年轻、中年、年老的隶属函数为：

$$\mu_{年轻}(x)=\begin{cases}1 & x\le 25\\ [1+(\frac{1-25}{10})^2]^{-1} & x>25\end{cases}$$

$$\mu_{中年}(x)=e^{-\frac{(x-45)^2}{118}}$$

$$\mu_{年老}(x)=\begin{cases}1 & x\ge 65\\ [1+(\frac{x-65}{10})^2]^{-1} & x<65\end{cases}$$

模糊检索模型

模糊信息检索方式的构造

SELECT * FROM 职工基本工资情况表

WHERE 性别=“男” AND 年龄=“稍微有点年老”

AND 基本工资=“约600元”

具体步骤为：

模糊检索模型

- **先将年龄字段值代入“稍微有点老”的隶属函数中求出相应的隶属度值；**
- 再将基本工资字段代入“基本工资600元左右”的隶属函数中求出隶属度值；
- 总隶属度=年龄隶属度∩基本工资隶属度；
- 将总隶属度≥阈值λ且满足其它检索条件的职工情况显示出来
- 若令阈值λ为0.5，则经过转换后的SQL检索语句为:

```
SELECT * FROM 职工基本工资情况表
WHERE 性别=“男” AND λ≥0.5
```

提纲

1. 信息检索模型的定义和分类
2. 布尔模型*
3. 向量空间模型*
4. 概率模型*
5. 改进的集合论检索模型
6. 改进的代数检索模型
7. 改进的概率检索模型
8. 结构化文本检索模型

6 改进的代数检索模型

- 广义向量空间模型
- 潜在语义索引
- 神经网络模型(*)

广义向量空间模型

(1) 引入原因

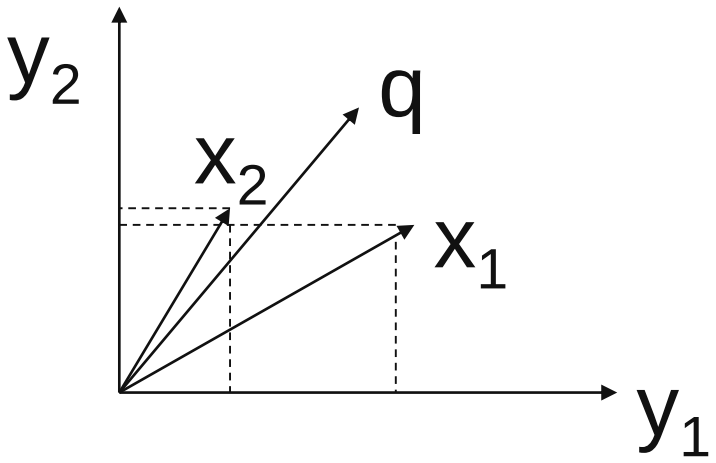
在向量空间模型中，文献向量和提问向量的表示中都假定标引词向量两两正交，然而在实际情况中，标引词之间总存在着一定的相互关系，即：不是两两正交的，一个词的出现可能会引起另外一个相关词的出现。于是**Wong、Ziarko在1985年提出：**

标引词向量不是两两正交的——广义向量空间模型的基本思想

广义向量空间模型

(2) 计算

在广义向量空间模型中，标引词向量由一组更小分量所组成的正交基向量来表示，词与词之间的关系可直接由基向量表示给出较为精确的计算。



$$x_1 = a_1^{(1)} y_1 + a_2^{(1)} y_2$$

$$x_2 = a_1^{(2)} y_1 + a_2^{(2)} y_2$$

$$q = b_1 y_1 + b_2 y_2$$

$$D_i = d_1^{(i)} x_1 + d_2^{(i)} x_2 = (d_1^{(i)}, d_2^{(i)}) \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

$$= (d_1^{(i)}, d_2^{(i)}) \begin{pmatrix} a_1^{(1)} & a_2^{(1)} \\ a_1^{(2)} & a_2^{(2)} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$$

广义向量空间模型

最后，利用标准余弦函数来计算文献向量D_i和提问向量q之间的相似度，并将文献按相似度的大小以递减顺序排列输出。

潜在语义索引(Latent Semantic Indexing，简称LSI)

一义多词：计算机、电子计算机、电脑
一词多义：病毒—医学、计算机

(1) 引入原因

用标引词来概括文献和提问文本的内容，由于受到同义词、多义词等的影响，可能导致检索效果低下，其准确性、完整性也不够理想

许多不相关的文献也有可能包括在结果集合中，没有用任何关键词进行标引的文献有可能被遗漏掉

潜语义标引模型

(2) 潜语义标引模型的含义

是一种将检索词和文件表示成矩阵(称为检索词——文档矩阵)的向量空间模型。

所谓“以语义为基础的检索”是指：

语义相同

被检索到的有关信息与使用者的查询不一定具有共同使用的术语，通过对检索词——文件矩阵降秩，可以去掉矩阵表示的数据库中的无关信息和噪声。

提纲

1. 信息检索模型的定义和分类
2. 布尔模型*
3. 向量空间模型*
4. 概率模型*
5. 改进的集合论检索模型
6. 改进的代数检索模型
7. 改进的概率检索模型
8. 结构化文本检索模型

7 改进的概率检索模型

传统的概率模型相关度排序函数的定义虽然比较直观，但是并没有具体给出相关文档集的定义，不仅在理论上存在很大的模糊性，而且实际操作起来也比较繁琐。

7 改进的概率检索模型

于是人们对该模型进行了推广，各种基于贝叶斯网络的信息检索模型引起了人们的广泛兴趣

- 推理网络检索模型(*)
- 信任度网络检索模型(*)

7 改进的概率检索模型

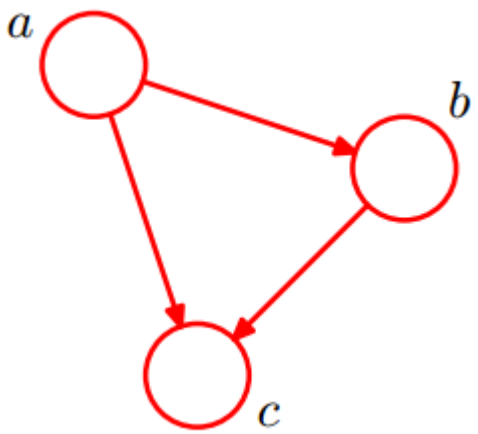
什么是贝叶斯网络？

是用来表示变量间连接概率的图形模式。它提供了一种自然的表示因果信息的方法，用来发现数据间的潜在关系。

贝叶斯网络

状态之间的因果关系可以有一个量化的可信度，用贝叶斯公式来表示，贝叶斯网络因此而得名

节点表示状态变量，有向边表示变量间的依赖关系

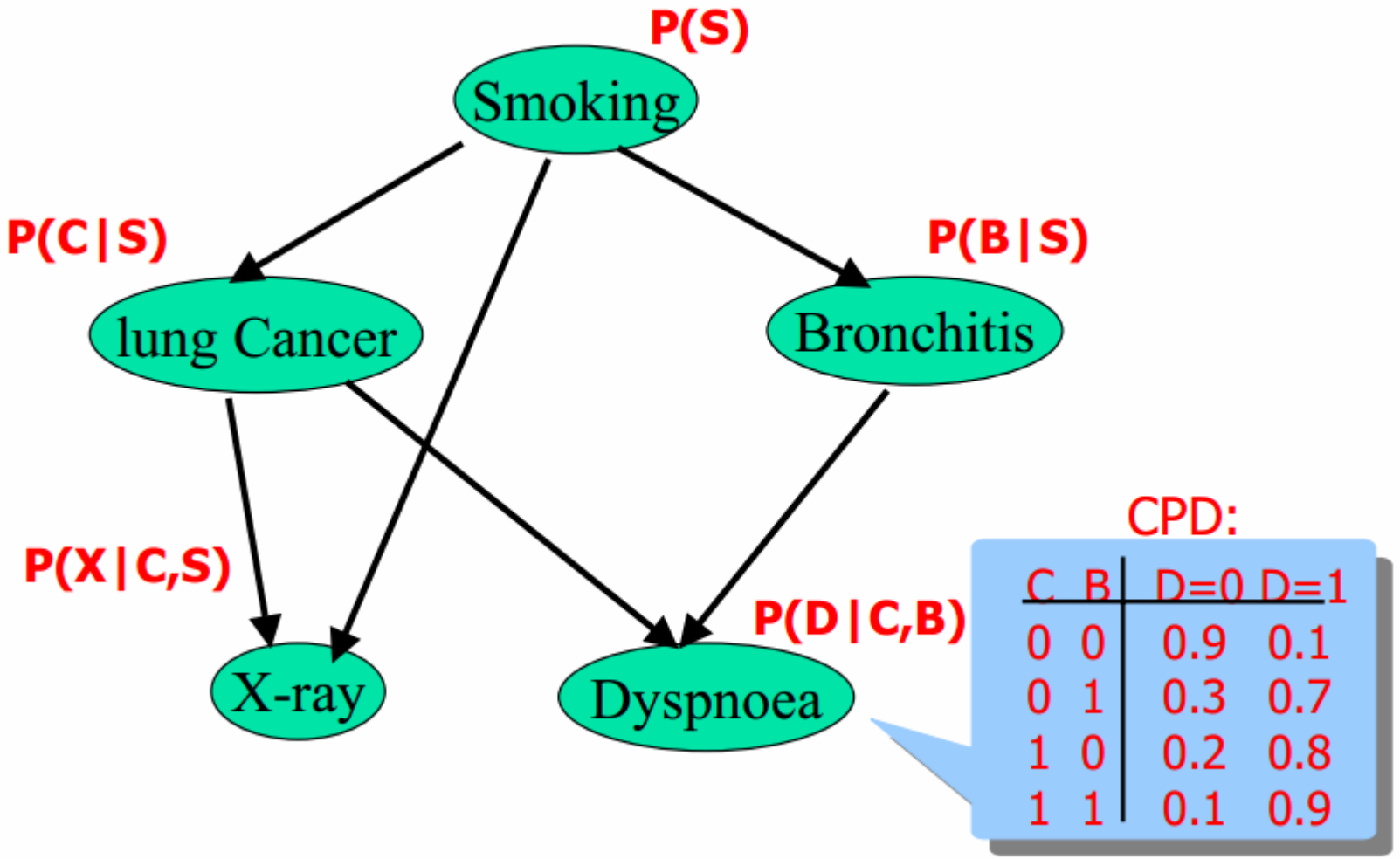


连接两个节点的箭头代表此两个随机变量是具有因果关系，或非条件独立。

应用：拼写检查

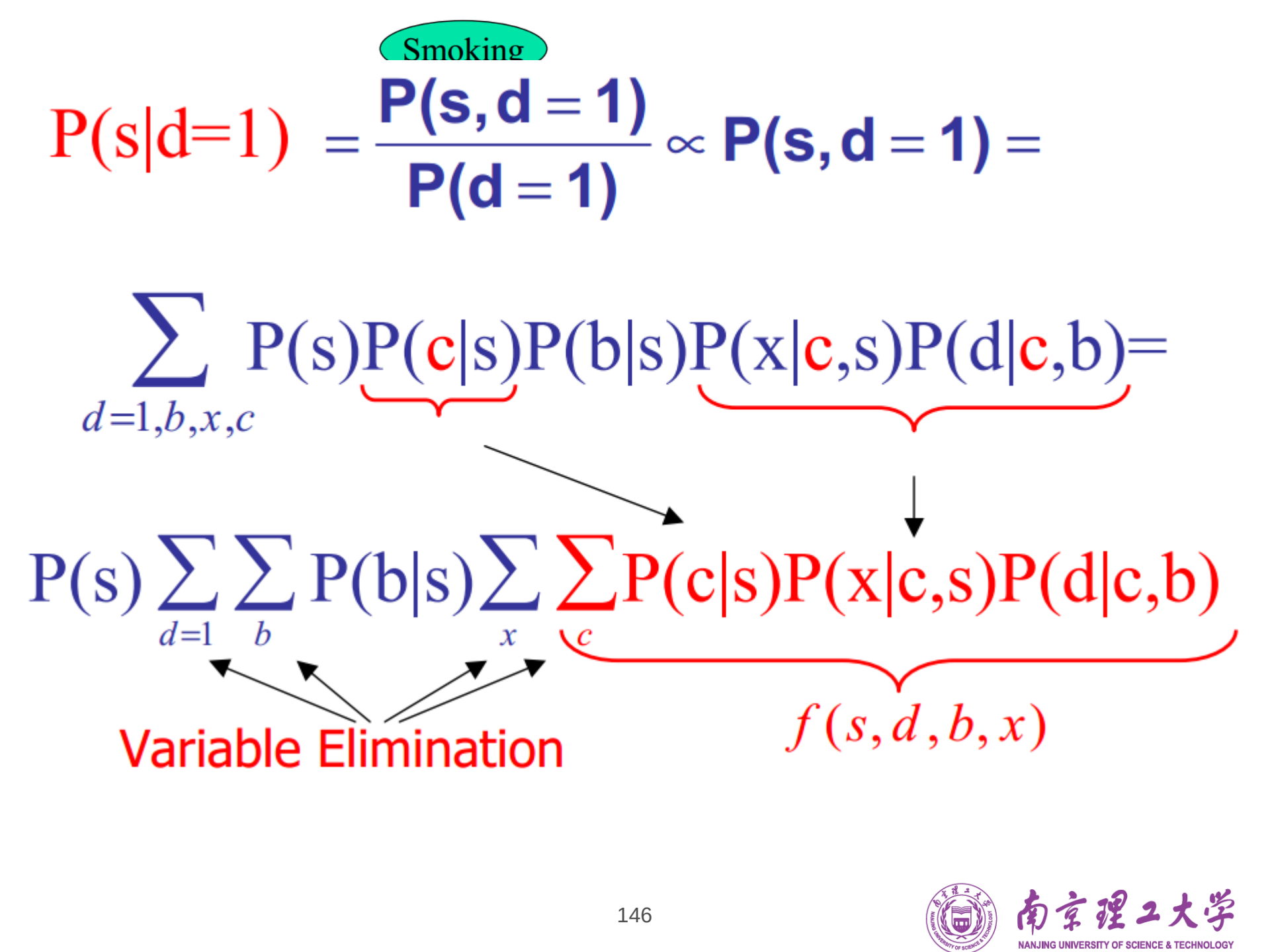
当你不小心输入一个不存在的单词时，搜索引擎会提示你是不是要输入某一个正确的单词，比如当你在Google中输入“Julw”时，系统会猜测你的意图：是不是要搜索“July”，如下图所示：

Google的拼写检查基于贝叶斯方法



其中，各个单词、表达式表示的含义如下：

- Smoking表示吸烟，其概率用P(S)表示，lung Cancer表示的肺癌，一个人在吸烟的情况下得肺癌的概率用P(C|S)表示，X-ray表示需要照医学上的X光，肺癌可能会导致需要照X光，吸烟也有可能会导致需要照X光（所以smoking也是X-ray的一个因），所以，因吸烟且得肺癌而需要照X光的概率用P(X|C,S)表示。
- Bronchitis表示支气管炎，一个人在吸烟的情况下得支气管炎的概率用P(B|S)，Dyspnoea表示呼吸困难，支气管炎可能会导致呼吸困难，肺癌也有可能会导致呼吸困难（所以lung Cancer也是dyspnoea的一个因），因吸烟且得了支气管炎导致呼吸困难的概率用P(D|C,B)表示。
- lung Cancer简记为C，Bronchitis简记为B，dyspnoea简记为D，且C = 0表示lung Cancer不发生的概率，C = 1表示lung Cancer发生的概率，B等于0（B不发生）或1（B发生）也类似于C，同样的，D=1表示D发生的概率，D=0表示D不发生的概率，便可得到dyspnoea的一张概率表，如上图的最右下角所示。



提纲

1. 信息检索模型的定义和分类
2. 布尔模型*
3. 向量空间模型*
4. 概率模型*
5. 改进的集合论检索模型
6. 改进的代数检索模型
7. 改进的概率检索模型
8. 结构化文本检索模型

8 结构化文本检索模型

- 结构化文本的含义
- 结构化文本检索的含义
- 标记语言来结构化文本的检索方法
- 非重叠链表的检索方法
- 邻近节点的检索方法

结构化文本的含义

结构化文本：和表达的思想内容相对应，在物理形式上有明显的组织结构和层次关系的文本

一般在文本信息中按照元素的包含关系加入文本的结构信息（如作者、标题、摘要、章的标题等）

结构化文本检索的含义

假定用户回忆起所感兴趣的文献中包含这样一页，该页中有一行：“原子大爆炸”在文本中以斜体的形式出现在一个图表的周围，图表的标识中含有单词“地球”。

传统的信息检索模型中，这个查询可以表示为：

“原子大爆炸” AND “地球”

同时包含这两个字符串的文献都被检索出来，显然用户会得到超过所希望的文献数量

结构化文本检索的含义

假定用户通过一个含义更丰富的表达式来表达他的查询：

同页 (相邻 (“原子大爆炸” 图 (标识 (“地球”))))

这个式子反映了他视觉回忆中的具体细节，能够更精确地表达用户的查询要求

结构化文本检索的含义

将①*文本中的内容信息*与②*文献结构信息*相结合的检索模型称之为结构化文本检索模型

标记语言结构化文本的检索方法

主要思想

用语言来定义和说明文本的结构。

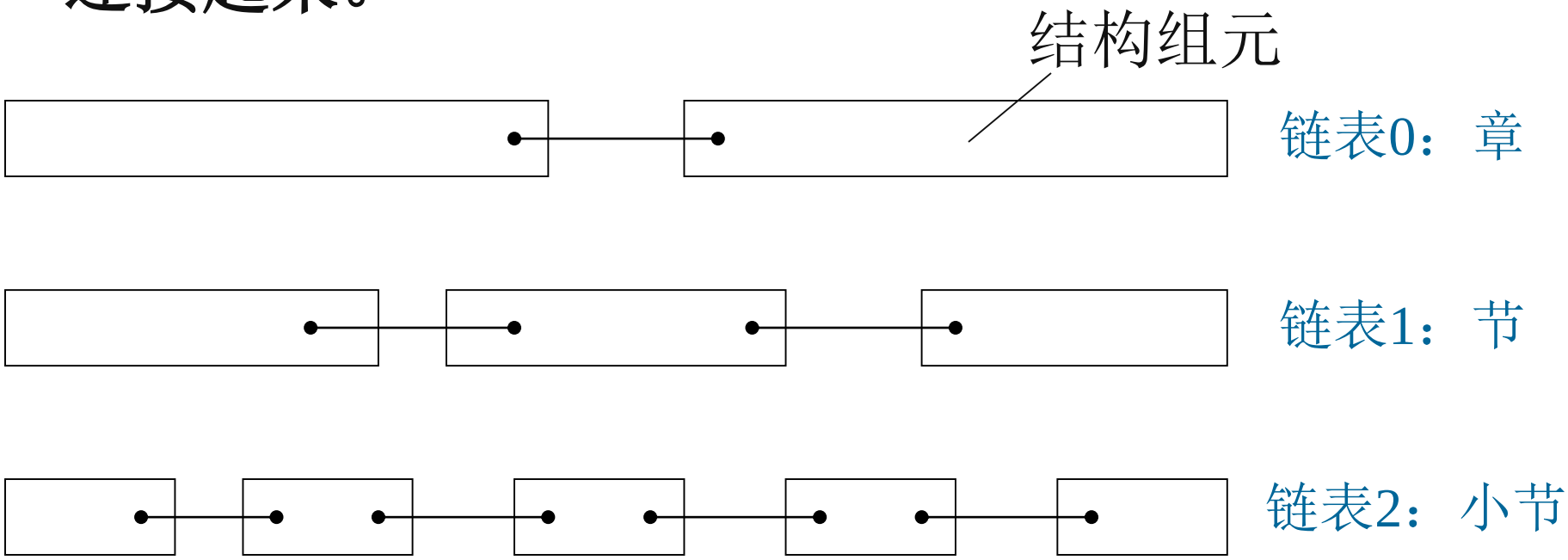
HTML、XML等一些格式化文本的标记语言可以结构化各种复杂的文本，甚至非文本的一些文档

主要是使用一些附加的、具有文本语法的标记将文本格式化：描述文本的格式、结构、语义和属性

非重叠链表结构化文本的方法

(1) 主要思想

把整个文本划分为非重叠的文本区域，并用链表连接起来。



通过3个独立的索引链表来表示文本中的结构

非重叠链表结构化文本的方法

- 为每一个链表建立一个独立的倒排文件
- 在倒排文件中，每个结构组元作为索引中的一项，与每一项相关的是一个文本区域的链表

非重叠链表结构化文本的方法

倒排文件(或称倒排文档索引)

是一种面向单词的标引机制，它为文本集合建立标引，以加快检索任务的速度。由两种表格组成：词汇表和事件表

1	6	9 11	17 19…	55	60

This is a text. A text has many words. Words are made from letters.

词汇表

Letters
Made
Many
text
words

事件表

60…
50 …
28 …
11,19 …
33,40…

倒排索引

非重叠链表结构化文本的方法

(2) 优缺点

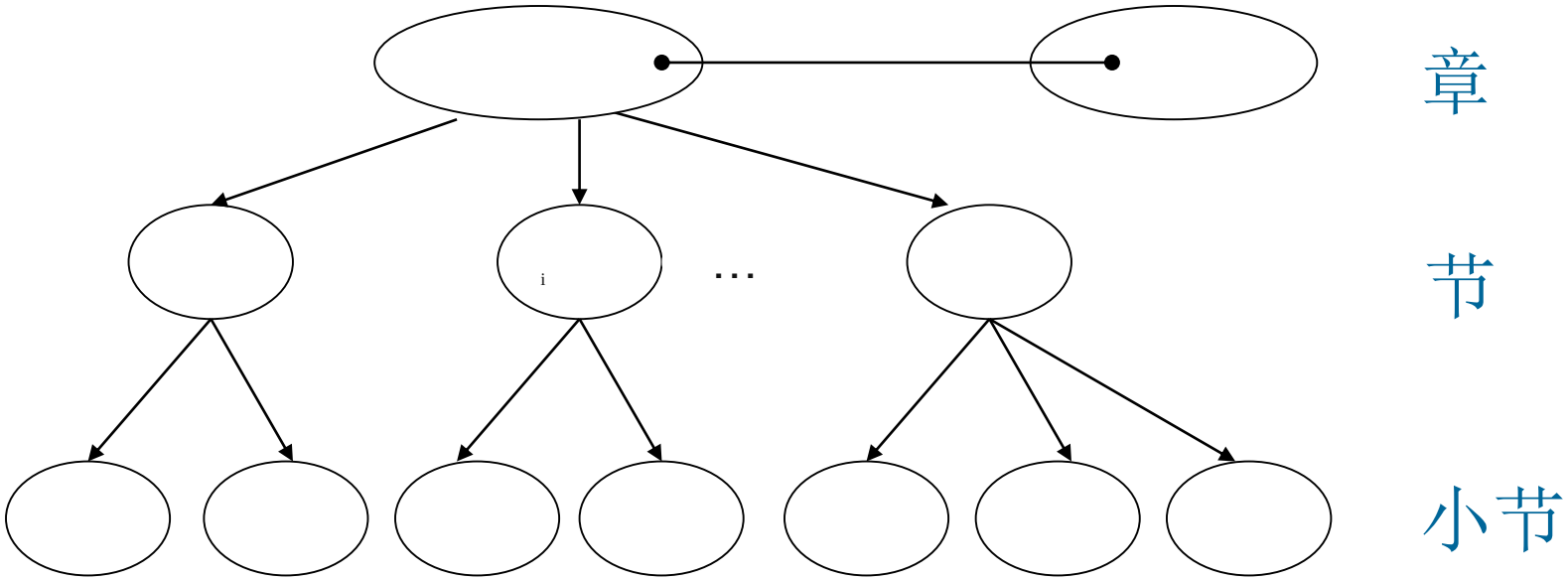
优点：非重叠链表结构化文本方法从文本的自身结构出发将文本格式结构化

缺点：对于一些非文字文本没有严格的描述方式，链表的结构也比较复杂，难于理解

基于邻接节点的方法

(1) 实现思路

是一种允许在相同文本中定义独立的分层索引结构的方法。



结构化单元的层次索引图

基于邻接节点的方法

每个索引结构是一个严格的层次结构，由章、节、小节、段和行组成，其中每个结构组元（章、节、小节、段和行）称为节点

基于邻接节点的方法

(2) 优缺点

这种方法将文本层次化结构分解，查询速度快，但文本区域出现重复，查询结果是嵌套的文本区域。

The End