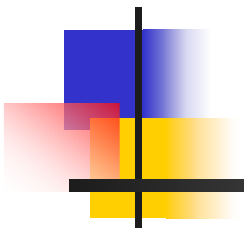
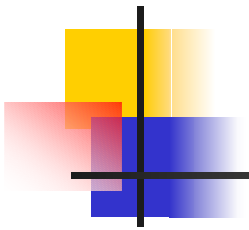


第十章　信息检索

詹卫东

2002.7

http://icl.pku.edu.cn/doubtfire/course/CL/2001_2002_2.htm



提纲

- 1 什么是信息检索（information retrieval）
- 2 信息检索的基本方法
  - 2.1 布尔模型
  - 2.2 向量空间模型
  - 2.3 概率模型
- 3 信息检索系统的评测
- 4 小结：信息检索的困难

1 什么是信息检索（IR）

太阳有多大
我想看关于儿童心理的文章
有不辣的川菜吗
……

Document Retrieval is defined as the matching of some stated user query against useful parts of free-text records.

Donna Harman et al. , 1996, *Document Retrieval* , in Survey of the State of the Art in Human Language Technology

IR的不同情形

信息源情况不同 —— 查询方式不同 —— 查询需求不同

Formatted Data	e.g. relational database
Unformatted Data	e.g. free text, web page, etc.

Query based on regular expression	e.g. SQL
Query with Natural Language	e.g. “汉语拼音的历史情况”

Expert oriented IR system

Common user oriented IR system

IR的需求发展

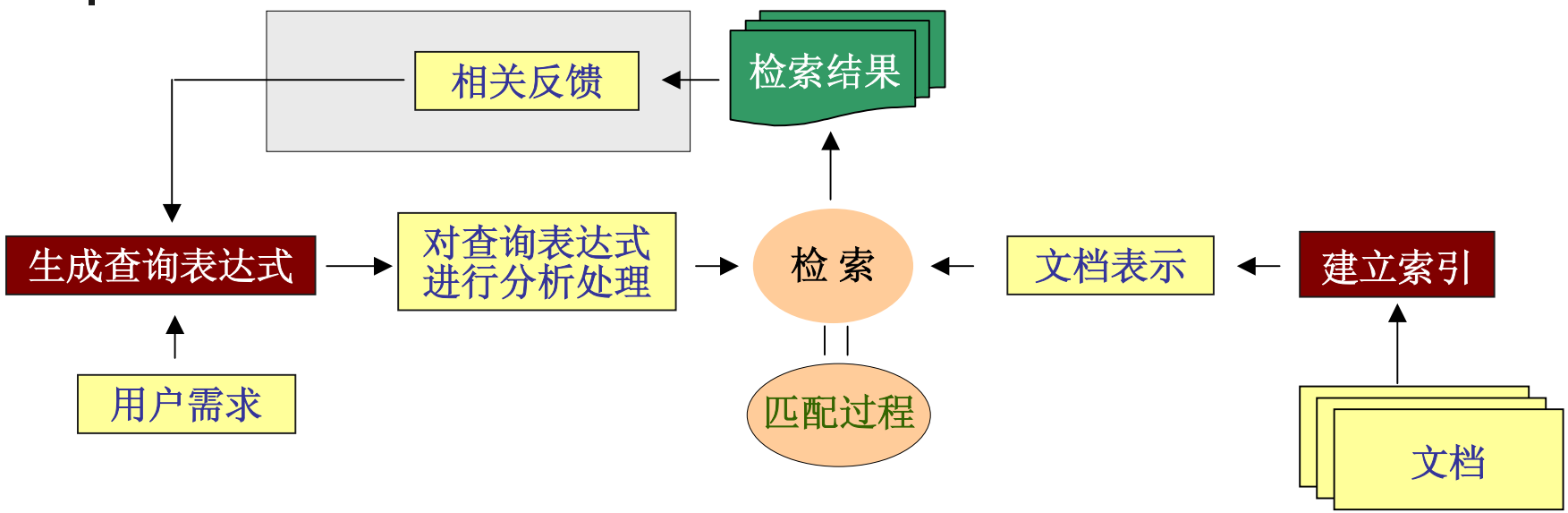
– Birth of World Wide Web 1990
– 50 million pages in November 1995
– 320 million pages in December 1997
– 800 million pages in February 1999
– 1 billion pages in 2000
– and growing every day

Internet & World Wide Web History
http://www.elsop.com/wrc/h_web.htm

media of information : from Hardcopy to electronic device

online data -- online information service

IR系统的一般模式



1 文档索引

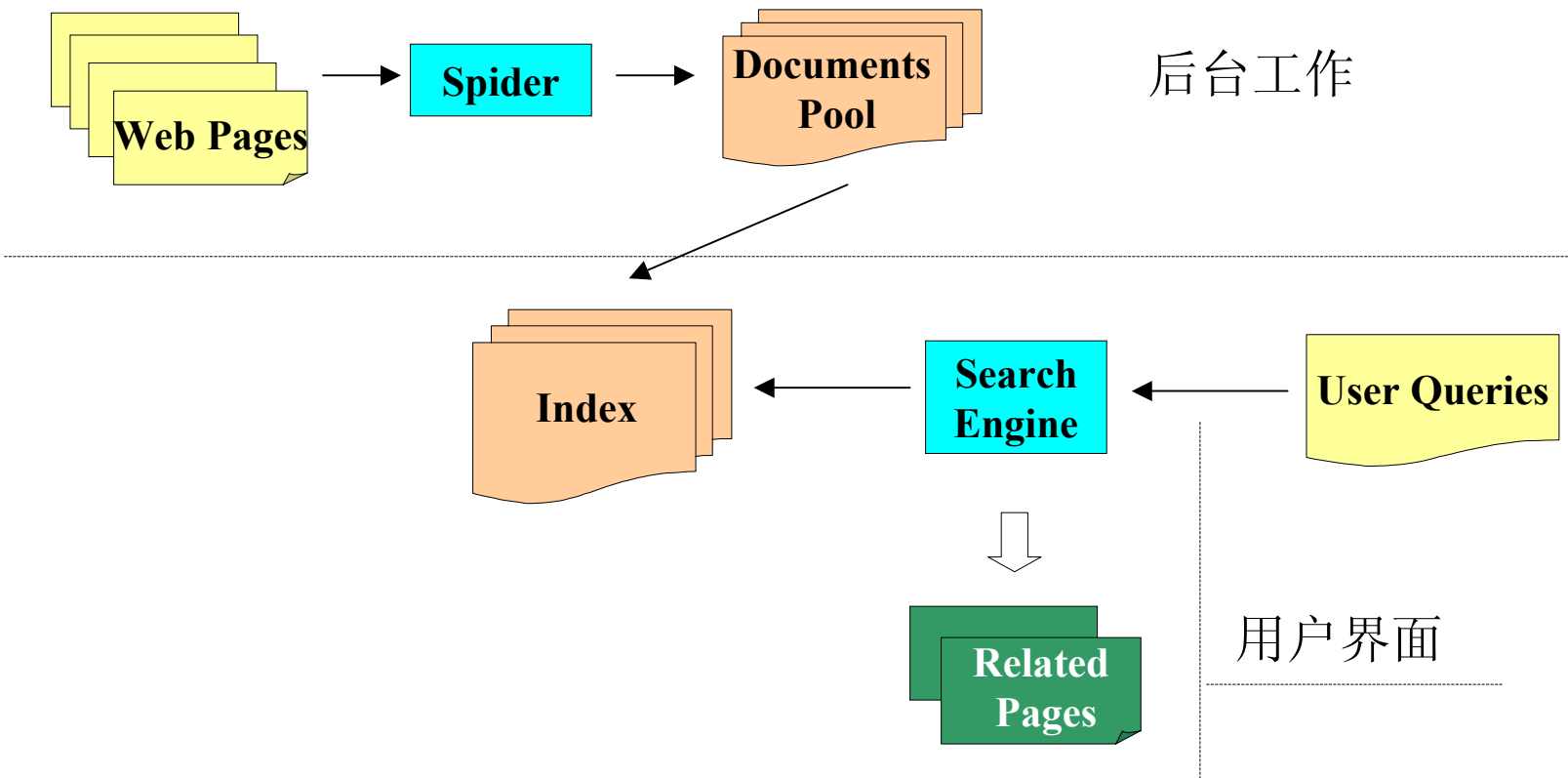
2 文档表示

3 查询表示

4 匹配/检索

5 相关性反馈
relevance feedback

Web search的一般模式



2 信息检索的基本方法

- 布尔模型 （Boolean Model）
- 向量空间模型（Vector Space Model）
- 概率模型 （Probabilistic Model）

2.1 布尔模型

查询表达式：由逻辑算子AND, OR, NOT连接若干
“项目”（term）构成

e.g. 1) “飞碟”

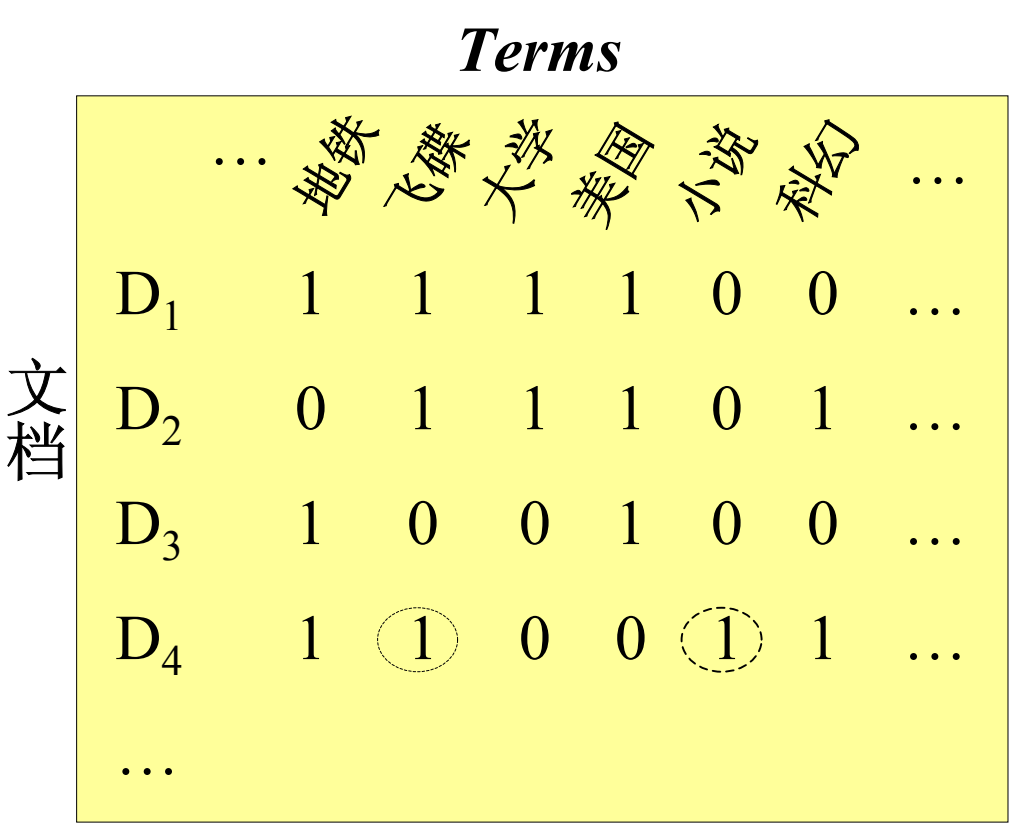
2) “飞碟” AND “美国”

3) “飞碟” AND (“中国” OR (NOT “科幻小说”))

检索/匹配：返回1，文档符合Query要求
返回0，文档不符合Query要求

Exact Matching

布尔检索示例



Query: “飞碟” AND “小说”

Retrieval/ Matching

Result: D_4

真值表（truth table）

P	*Q*	*NOT P*	*P AND Q*	*P OR Q*
0	0	TRUE	FALSE	FALSE
0	1	TRUE	FALSE	TRUE
1	0	FALSE	FALSE	TRUE
1	1	FALSE	TRUE	TRUE

布尔检索的优缺点

优点：	缺点：
1）简单、速度快	1）不够精确，不能反映不同“项目”对一个文档的重要程度的差异
2）查询表达式易于掌握	2）检索结果地位平等，无法排序

“飞碟” AND “小说”：只能检索出D_4，无法显现D_1,D_2,D_3的差异

“飞碟” OR “小说”：可以检出D_1,D_2,D_4，但无法显现它们的差异

扩展的布尔检索（extended Boolean Model）

基本思想：将非此即彼的匹配方式改为计算相似度similarity

e.g. 对于Term_1 OR Term_2 形式的Query，相似度公式为:

$$sim(q_{or}, d_j) = \sqrt{\frac{(x^2 + y^2)}{2}}$$

x表示Term_1在文档d_j中的重要程度 $\in (0,1)$
y表示Term_2在文档d_j中的重要程度 $\in (0,1)$

对于Term_1 AND Term_2 形式的Query，相似度公式为:

$$sim(q_{and}, d_j) = 1 - \sqrt{\frac{(1-x)^2 + (1-y)^2}{2}}$$

扩展的布尔检索相似度计算示例

Query / Doc Sim()	“飞碟” AND “小说”	“飞碟” OR “小说”
D_1	0.293	0.707
D_2	0.293	0.707
D_3	0	0
D_4	1	1

$x, y = 1$ if a term exists in d_j

$x, y = 0$ otherwise

从“一刀切”到“合理拉开差距”

P-norm模型

将上述只包含两个项目的查询式的相似度计算进一步拓展为包含 m 个项目的查询式的相似度计算

$$sim(q_{or}, d) = \left[\frac{x_1^p + x_2^p + \ldots + x_m^p}{m} \right]^{\frac{1}{p}}$$

$$sim(q_{and}, d) = 1 - \left[\frac{(1-x_1)^p + (1-x_2)^p + \ldots + (1-x_m)^p}{m} \right]^{\frac{1}{p}}$$

x_m 表示第 m 个项目在文档d中的权重

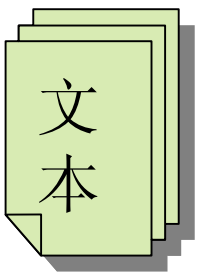
$1 \le p \le \infty$　p 表示项目间逻辑关系严格的程度(degree of strictness)，取值为1最松，取值为无穷大最严

2.2 向量空间模型

要点：

- 文档D和查询Q（不妨统称为文本）都可用向量表示
- 检索过程就是计算文档向量与查询向量之间的相似度
- 可以根据相似度值的不同，对检索结果进行排序
- 可以根据检索结果，进一步做相关检索（relevance feedback）

从文本到向量空间（vector space）



Vocabulary

Index Term$_1$

Index Term$_2$

…

Index Term$_n$

Vector space

	T_1	T_2	…	T_n
D_1	$w_{1,1}$	$w_{1,2}$	…	$w_{1,n}$
D_2	$w_{2,1}$	$w_{2,2}$	…	$w_{2,n}$
…				
D_m	$w_{m,1}$	$w_{m,2}$	…	$w_{m,n}$

若有 n 个项目（term），文本 D_i 就可以表示为一个 n 维向量；$W_{i,j}$ 表示文本 D_i 的第 j 维的权值，即项目权值（term weight）

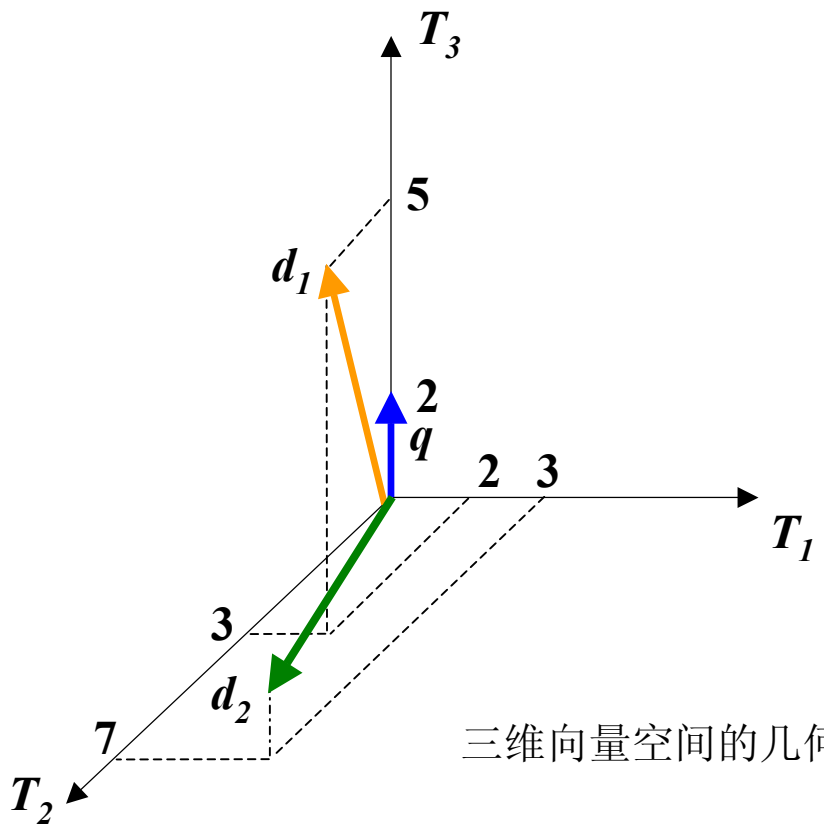
文档的向量表示示例

- 假定有三个项目：

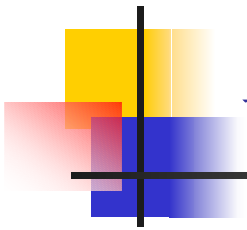
  “葡萄”，“美酒”，“夜光杯”

- 假定以项目在文本中的出现次数为项目的权值

	葡萄 T_1	美酒 T_2	夜光杯 T_3
d_1	2	3	5
d_2	3	7	2
q	0	0	2



三维向量空间的几何图示



计算向量之间的相似程度

向量间相似程度的不同度量方法

- Inner product
- Dice coefficient
- Cosine coefficient
- Jaccard coefficient

在上面的例子中，如何度量 q 跟 d_1 相似还是跟 d_2 相似？

夹角余弦：相似程度的度量方法之一

设有查询向量

$$\vec{q} = (q_1, q_2, ..., q_n)$$

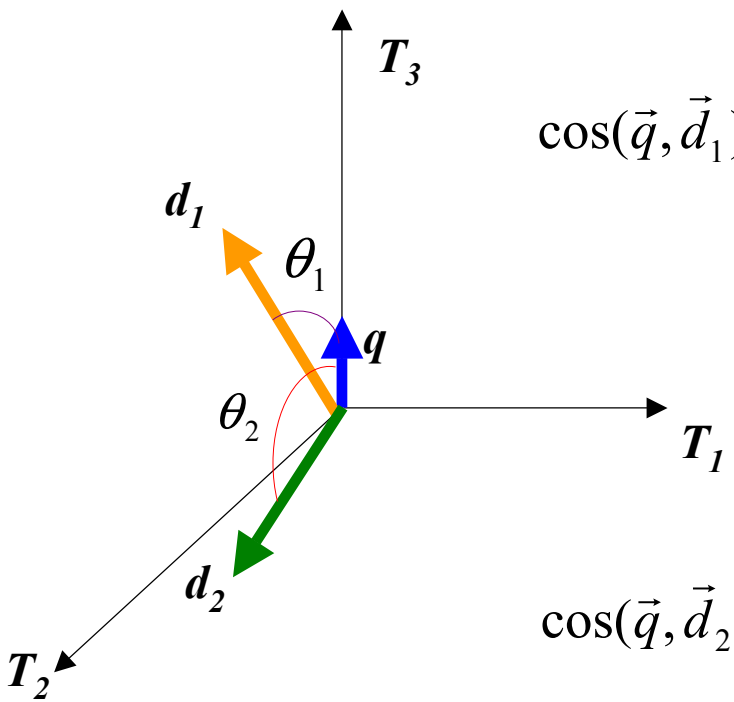
文档向量

$$\vec{d} = (d_1, d_2, ...d_n)$$

$$CosSim\ (\vec{q}, \vec{d}) = \frac{\sum_{i=1}^{n} q_i \times d_i}{\sqrt{\sum_{i=1}^{n} {q_i}^2} \times \sqrt{\sum_{i=1}^{n} {d_i}^2}}$$

夹角余弦计算示例

d_1= <2,3,5>　　d_2 = <3,7,1>　　q = <0,0,2>



$$\cos(\vec{q},\vec{d}_1)=\cos(\theta_1)=\frac{2\times0+3\times0+5\times2}{\sqrt{2^2}\times\sqrt{2^2+3^2+5^2}}=\frac{5}{\sqrt{38}}=0.81$$

q 与 d_1更相似

$$\cos(\vec{q},\vec{d}_2)=\cos(\theta_2)=\frac{3\times0+7\times0+1\times2}{\sqrt{2^2}\times\sqrt{3^2+7^2+1^2}}=\frac{1}{\sqrt{59}}=0.13$$

索引项权值的计算（term weight）

权值的直观含义：一个项目对于一个文本的重要程度

即一个项目在多大程度上可以将这个文档与其他文档区别开

计算权值的两种简单方式：

（1）项目－ 出现/不出现：1或0

（2）项目－ 出现的次数：0，1，2，…

需要更好的加权方法

（3） *tf .idf* 加权法 （term frequency • inverse document frequency）

项频率 逆向文档频率

tf.idf 加权

Term frequency： $term_i$ 在文档 d_j 中的出现次数，记做 $tf_{i,,j}$

$tf_{i,,j}$ 越高，意味着 $term_i$ 对于文档 d_j 就越重要
比如：一篇谈论乔丹的文章，可以预期“乔丹”、“飞人”的 tf 值会比较高

Document frequency： 含有 $term_i$ 的文档的数量，记做 df_i

df_i 越高，意味着 $term_i$ 在衡量文档之间相似性方面作用越低，
比如“的”的df值肯定非常高，因此不具有区别性，这类词称为“非焦点词”

Inverse document frequency：跟 df_i 形成“反比关系”，$idf_i = \log(\frac{N}{df_i})$

idf_i 值越高，意味着 $term_i$ 对于文档的区别意义越大

N 为全部文档的数量。如果一个项目仅出现在一个文档中，$idf = logN$，
如果一个项目出现在所有文档中，$idf = log1 = 0$

tf.idf 加权（续）

索引项加权：给那些经常出现在一个文档中，而不常出现在其他文档中的项目以更高的权重，即让“特别的词”从“一般的词”中凸现出来。

在这个基本精神指导下，有许多不同的加权公式

公式一：
$$weight_{i,j} = tf_{i,j} \times idf_i = tf_{i,j} \times \log \frac{N}{df_i}$$

公式二：
$$weight_{i,j} = \begin{cases} (1 + \log(tf_{i,j})) \log \frac{N}{df_i}, & \text{当} tf_{i,j} \geq 1 \\ 0, & \text{当} tf_{i,j} = 0 \end{cases}$$

……

tf.idf 加权示例

Query = “夏夜湖畔的蛙鸣”

D1：湖畔的夏夜常常很凉爽，……

D2 :湖畔有家“湖畔”啤酒花园，花园中常常是鼓鼓的蛙鸣一片，……

D3：“蛙鸣”禅社举办“蛙鸣”诗会的消息……

Term	…	湖畔	夏夜	的	常常	蛙鸣	禅社	诗会	…
df	…	2	1	3	2	2	1	1	…
idf	…	0.176	0.477	0	0.176	0.176	0.477	0.477	…

$N = 3$

$$idf_i = \log(\frac{N}{df_i})$$

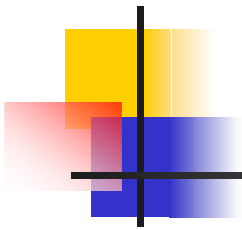
tf.idf 加权示例（续）

公式一： $weight_{i,j} = tf_{i,j} \times idf_i$

Term / $W_{i,j}$ / Doc	…	湖畔	夏夜	的	常常	蛙鸣	禅社	诗会	…
D_1	…	0.176	0.477	0	0.176	0	0	0	…
D_2		0.352	0	0	0.176	0.176	0	0	
D_3		0	0	0	0	0.352	0.477	0.477	
Q	…	0.176	0.477	0	0	0.176	0	0	…

$Cos(q,d_1) = 0.893$ $Cos(q,d_2) = 0.400$ $Cos(q,d_3) = 0.151$

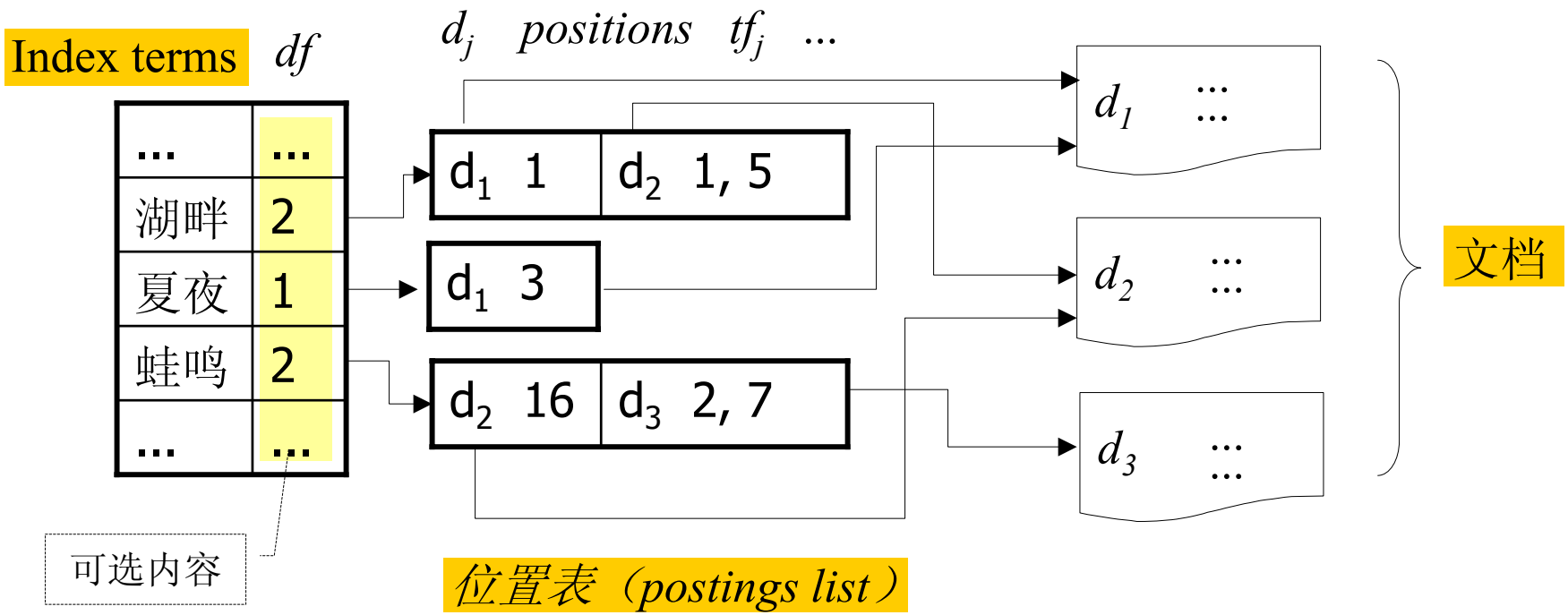
与查询 q 相似的文档顺序：$d_1 \succ d_2 \succ d_3$



停用词表（stop list）

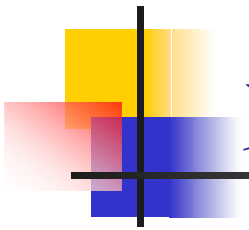
- 表达实际文档所需的 *term* 很多，空间开销很大
- 有些“词”在 *query* 时很少出现，即不大作为用户的查询目标，比如 “常常”，“of”，…
- 有些“词”在每个文档中都会出现，比如“的”，这些词的 *idf* 值通常为 0
- 一般把“的”，“of”这类词收集起来，构成一个停用词表
- 因此，在为文档建索引的时候，可以不停用词表中的词。这样可以节省资源，同时也不至于太影响检索效果

文档索引（inverted index）



Query: "湖畔" AND "蛙鸣" ⟶ 对两个term对应的位置表求交集

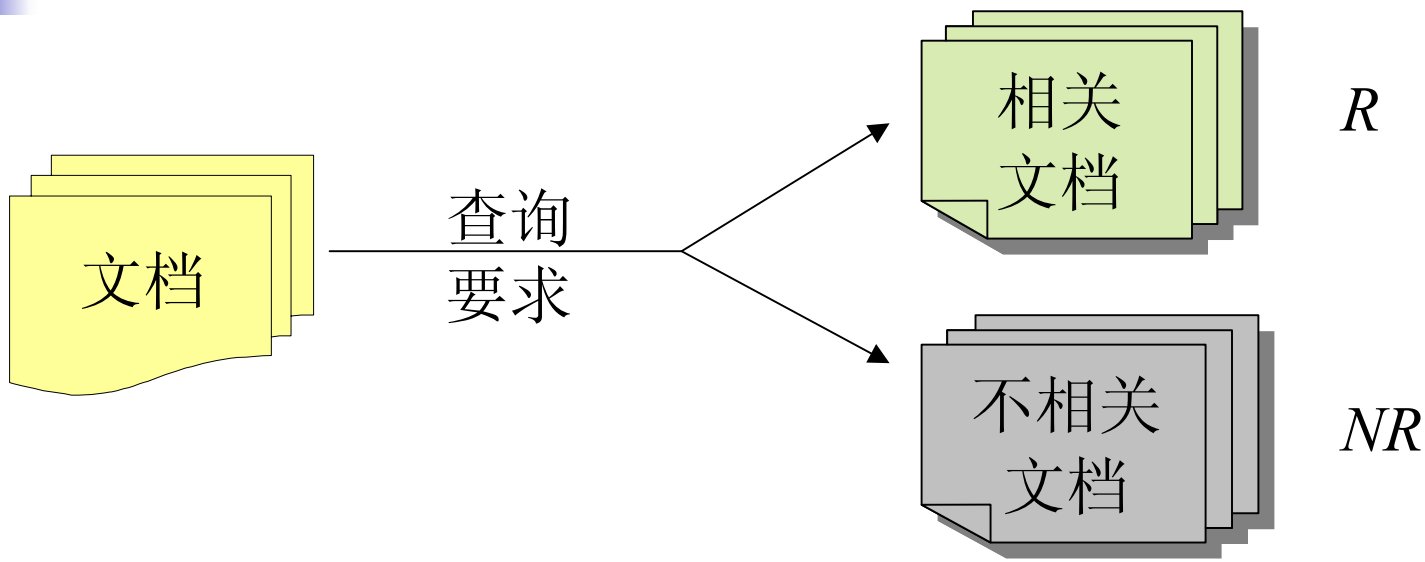
Query: "湖畔" OR "蛙鸣" ⟶ 对两个term对应的位置表求并集



文档分析的其他问题

- lemmatization
- stemming
- 索引项的选择（index terms selection）
- 文档的压缩、存贮

2.3 概率模型



检索问题即求条件概率问题

if $\text{Prob}(R \mid d_i, q) > \text{Prob}(NR \mid d_i, q)$ then d_i 是检索结果，否则不是检索结果

3 信息检索系统的评价

	文档	属于检索结果集合 Rt	不属于检索结果集合 NRt
NR	不相关	＋不相关 ＋ 在检索结果中 A	＋不相关＋不在检索结果中 C
R	相关	＋相关 ＋ 在检索结果中 B	＋相关 ＋ 不在检索结果中 D

$$\text{准确率} = \frac{\text{检索结果中和查询相关的文档数}}{\text{检索结果中的文档总数}} = \frac{B}{Rt} \qquad precision$$

$$\text{召回率} = \frac{\text{检索结果中和查询相关的文档数}}{\text{文档库中所有和查询相关的文档数}} = \frac{B}{R} \qquad recall$$

$$\text{误识率} = \frac{\text{检索结果中和查询不相关的文档数}}{\text{文档库中所有和查询不相关的文档数}} = \frac{A}{NR} \qquad fallout$$

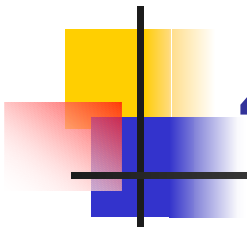
断点准确率（precision at cutoff）

评价	系统1	系统2	系统3
	D1 √	D10 ×	D6 ×
	D2 √	D9 ×	D1 √
	D3 √	D8 ×	D2 √
	D4 √	D7 ×	D10 ×
	D5 √	D6 ×	D9 ×
	D6 ×	D1 √	D3 √
	D7 ×	D2 √	D5 √
	D8 ×	D3 √	D4 √
	D9 ×	D4 √	D7 ×
	D10 ×	D5 √	D8 ×
断点5处的准确率	100%	0%	40%
断点10处的准确率	50%	50%	50%

系统1的查询结果排序优于系统3，系统3优于系统2

TREC评测

- Text REtrieval Conference
  http://trec.nist.gov/
- 组织者
  - NIST(National Institute of Standards and Technology)，美国政府部门
  - DARPA(Defense Advanced Research Projects Agency)，美国军方
- 1992 – 2001（每年一届）
  - 大测试集 - 测试语料主要来源： LDC语料
  - 自动评估与人工评估相结合，完全公开的评估体系和软件系统
  - 以评估促进研究成果实用化



4 小结

- 目前比较成熟的正在使用的IR系统并没有用到太多的语言学知识
- 理想的检索系统是所谓的 语义层（概念层）的检索系统，要求IR系统对文档库中的文档，以及用户的查询做到“真正的理解”
- 从IR系统向QA（question-answer）系统发展

汉语信息检索的特殊问题

- 汉字编码标准不统一 GB, GIG5, Unicode
- 按字索引 / 按词索引?
- 文本分词问题
  - 不分词：

    检索“中将” 误检 “地铁中将可使用移动电话”
  - 分词：

    检索“旱灾” 漏检 “抗旱、受旱地区、……”

进一步阅读文献

- 吴立德 等（1997）《大规模中文文本处理》，复旦大学出版社1997年版。第6.2节
- Christopher D. Manning & Hinrich Schutze, 1999, *Foundations of Statistical Natural Language Processing*, The MIT Press. Chapter 15.
- Ronald A. Cole, et al. eds., 1996, *Survey of the State of the Art in Human Language Technology*, Cambridge University Press. Chapter 7.2
- N. Fuhr, 1992, *Probabilistic Model in Information Retrieval*, In The Computer Journal, Vol. 35, No.3.
- C.van Rijsbergen, 1979, *Information Retrieval*, 2nd edtion, Butterworths, London, 1979

Web IR service （2002年）

http://www.google.com (google搜索引擎) -- 全球网页

http://e.pku.edu.cn/ (北大天网搜索引擎) -- 中文网页

http://www.baidu.com/home.html (百度搜索引擎) -- 中文网页

http://www.portal.com.hk/ （香港入门网） -- 香港、大陆

http://www.openfind.com.tw/ -- 台湾

http://www.profusion.com/ （profusion） -- 英文网页

复习思考题

1. 请说明布尔检索和向量空间检索模型各自的优缺点。
2. 对互联网上一些知名的搜索引擎进行一定规模的调查，从用户的角度撰写调查报告。