

现代信息检索

Modern Information Retrieval

第8讲 概率检索模型

Probabilistic Information Retrieval

提纲

- ① 向量空间模型回顾
- ② 基本概率统计知识
- ③ Logistic回归模型
- ④ BIM模型
- ⑤ BM25模型

提纲

- ① 向量空间模型回顾
- ② 基本概率统计知识
- ③ Logistic回归模型
- ④ BIM模型
- ⑤ BM25模型

向量空间模型回顾

- 文档表示成向量
- 查询也表示成向量
- 计算两个向量之间的相似度：余弦相似度、内积等等
- 在向量表示中的词项权重计算方法主要是tf-idf公式，实际考虑tf、idf及文档长度3个因素

tf-idf权重计算的三要素

词项频率tf		文档频率df		归一化方法	
n(natural)	$tf_{t,d}$	n(no)	1	n(none)	1
l(logarithm)	$1 + \log(tf_{t,d})$	t(idf)	$\log \frac{N}{df_t}$	c(cosine)	$\frac{1}{\sqrt{w_1^2 + w_2^2 + \dots + w_M^2}}$
a(augmented)	$0.5 + \frac{0.5 \times tf_{t,d}}{\max_t(tf_{t,d})}$	p(prob idf)	$\max \left\{ 0, \log \frac{N - df_t}{df_t} \right\}$	u(pivoted unique)	$1/u(17.4.4节)$
b(boolean)	$\begin{cases} 1 & \text{if } tf_{t,d} > 0 \\ 0 & \text{otherwise} \end{cases}$			b(byte size)	$1/CharLength^a, a < 1$
L(log ave)	$\frac{1 + \log(tf_{t,d})}{1 + \log(\text{ave}_{t \in d}(tf_{t,d}))}$				

向量空间模型的优缺点

- 优点：
 - 简洁直观，可以应用到很多其他领域(文本分类、生物信息学)。
 - 支持部分匹配和近似匹配，结果可以排序
 - 检索效果不错
- 缺点：
 - 理论上不够严谨，往往基于直觉的经验性公式
 - 词项之间的独立性假设与实际不符：实际上，词项的出现之间是有关系的，并不是完全独立的。如：“张继科”、“乒乓球”的出现不是独立的。

本讲内容

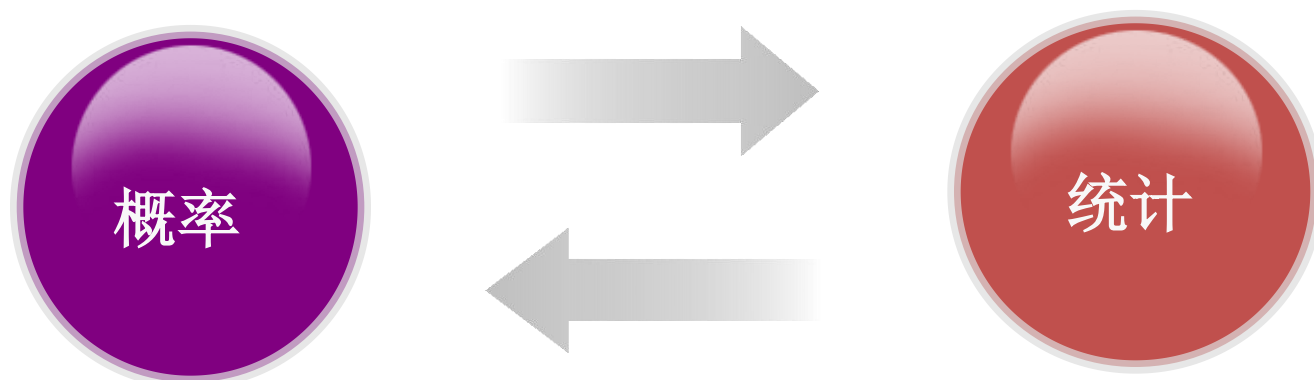
- 概率基础知识
- 基于概率理论的检索模型
- Logistic回归模型
- 二值独立概率模型 BIM: 不考虑TF和文档长度
- 考虑TF和文档长度的BM25模型

提纲

- ① 向量空间模型回顾
- ② 基本概率统计知识
- ③ Logistic回归模型
- ④ BIM模型
- ⑤ BM25模型

概率 vs. 统计

概率是统计的理论基础



统计是概率的实际应用

典型问题：已知某数据总体满足某分布，抽样得到某数据的概率是多少？

典型问题：已知某抽样数据(或总体分布)，判断总体的分布(或分布参数) 是多少？

概率统计初步

- 随机试验与随机事件
- 概率和条件概率
- 乘法公式、全概率公式、贝叶斯公式
- 随机变量
- 随机变量的分布

随机试验和随机事件

- **随机试验**：可在相同条件下重复进行；试验可能结果不止一个，但能确定所有的可能结果；一次试验之前无法确定具体是哪种结果出现。
 - 掷一颗骰子，考虑可能出现的点数
- **随机事件**：随机试验中可能出现或可能不出现的情况叫“随机事件”
 - 掷一颗骰子，4点朝上

概率和条件概率

- **概率**：直观上来看，事件A的概率是指事件A发生的可能性，记为 $P(A)$
 - 掷一颗骰子，出现6点的概率为多少？
- **条件概率**：已知事件A发生的条件下，事件B发生的概率称为A条件下B的条件概率，记作 $P(B|A)$
 - 30颗红球和40颗黑球放在一块，请问第一次抽取为红球的情况下第二次抽取黑球的概率？
 - 是否放回结果不同

乘法公式、全概率公式和贝叶斯公式

- 乘法公式：
 - $P(AB)=P(A)P(B|A)$
 - $P(A_1A_2...A_n)=P(A_1)P(A_2|A_1)...P(A_n|A_1...A_{n-1})$
- 全概率公式： $A_1A_2...A_n$ 是整个样本空间的一个划分

$$P(B)=\sum_{i=1}^n P(A_i)P(B|A_i)$$

- 贝叶斯公式： $A_1A_2...A_n$ 是整个样本空间的一个划分

$$P(A_j|B)=\frac{P(A_j)P(B|A_j)}{\sum_{i=1}^n P(A_i)P(B|A_i)}, (j=1,...,n)$$

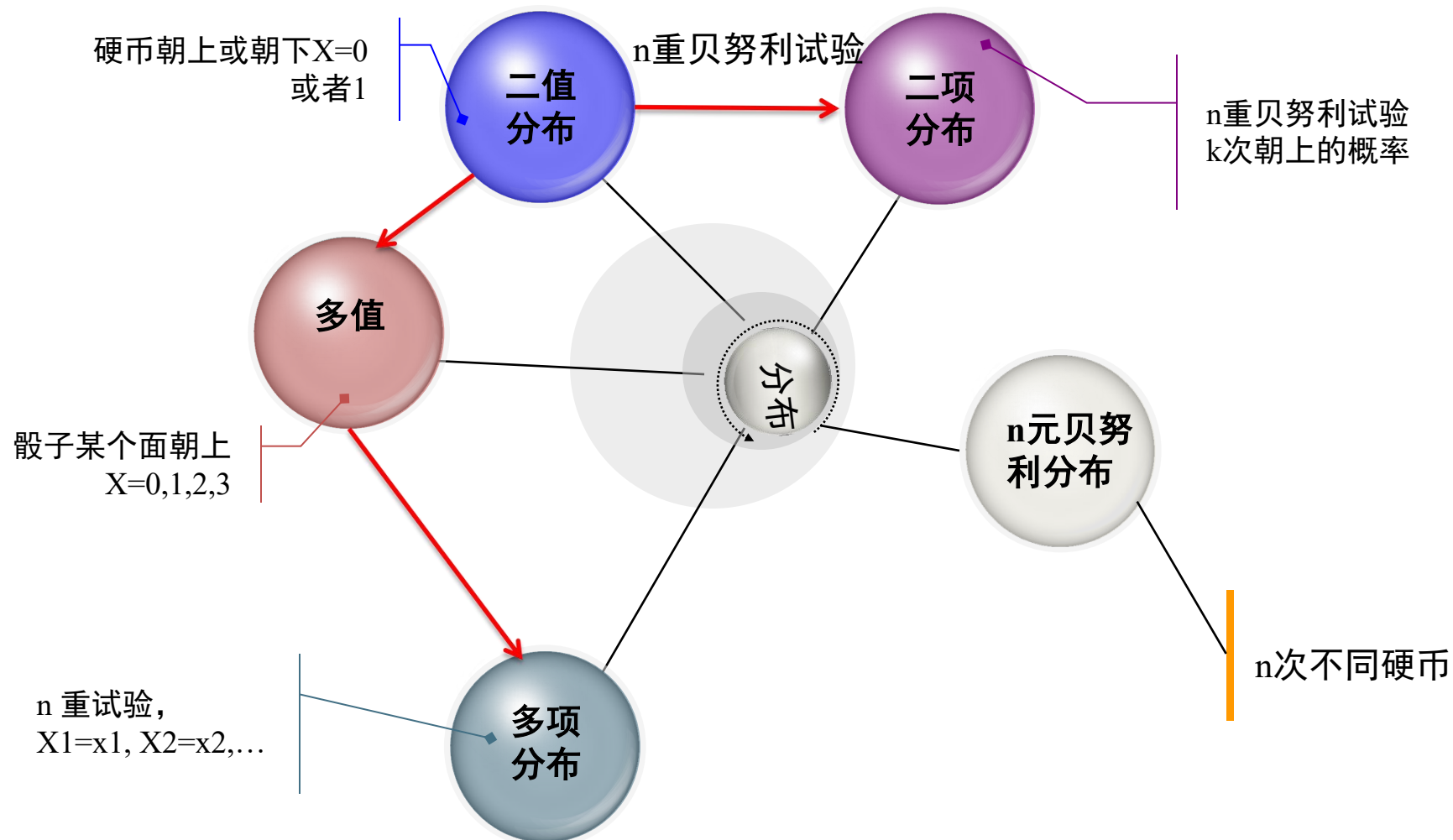
事件的独立性

- 两事件独立：事件A、B，若 $P(AB)=P(A)P(B)$ ，则称A、B独立
- 三事件独立：事件A B C，若满足 $P(AB)=P(A)P(B)$ ， $P(AC)=P(A)P(C)$ ， $P(BC)=P(B)P(C)$ ， $P(ABC)=P(A)P(B)P(C)$ ，则称A、B、C独立
- 多事件独立：两两独立、三三独立、四四独立....

随机变量

- 随机变量：若随机试验的各种可能的结果都能用一个变量的取值（或范围）来表示，则称这个变量为**随机变量**，常用 X 、 Y 、 Z 来表示
 - (离散型随机变量)：掷一颗骰子，可能出现的点数 X (可能取值1、2、3、4、5、6)
 - (连续型随机变量)：北京地区的温度(-15~45)

各种分布关系图



概率检索模型

- 检索系统中，给定查询，计算每个文档的相关度
- 检索系统对用户查询的理解是非确定的(uncertain)，对返回结果的猜测也是非确定的
- 而概率理论为非确定推理提供了坚实的理论基础
- 概率检索模型可以计算文档和查询相关的可能性

概率检索模型

- 概率检索模型是通过概率的方法将查询和文档联系起来
 - 定义3个随机变量 R 、 Q 、 D ：相关度 $R=\{0,1\}$ ，查询 Q 可以是 q_1, q_2, \dots 中的一个查询，文档 D 可以是 d_1, d_2, \dots 中的一篇文档，则可以通过计算条件概率 $P(R=1|Q=q, D=d)$ 来度量文档和查询的相关度。
- 概率检索模型包括一系列模型，如Logistic Regression(回归)模型及最经典的二值独立概率模型BIM、BM25模型等等(还有贝叶斯网络模型)。
- 1998出现的基于统计语言建模的信息检索模型本质上也是概率模型的一种。

概率排序原理(PRP)

- 简单地说：如果文档按照与查询的相关概率大小返回，那么该返回结果是所有可能获得结果中效果最好的。
- 严格地说：如果文档按照与查询的相关概率大小返回，而这些相关概率又能够基于已知数据进行尽可能精确的估计，那么该返回结果是所有基于已知数据获得的可能的结果中效果最好的。

几种概率检索模型

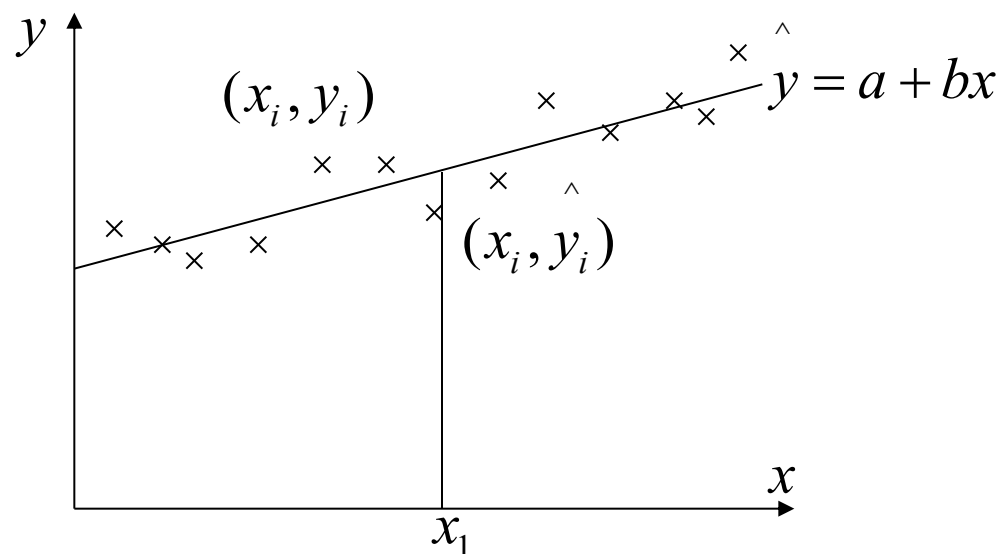
- 基于Logistic回归的检索模型
- 经典的二值独立概率模型BIM
- 经典的BM25模型 (BestMatch25)
- 贝叶斯网络模型：本讲义不介绍，请参考有关文献。
- 基于语言建模的检索模型：1998年兴起，研究界的热点。下一讲介绍。

提纲

- ① 向量空间模型回顾
- ② 基本概率统计知识
- ③ Logistic回归模型
- ④ BIM模型
- ⑤ BM25模型

回归(Regression)

- 回归分析：回归分析是处理变量之间相关关系的一种工具，回归的结果可以用于预测或者分类
- 一元线性回归：根据观测点，拟合出一条直线，使得某种损失 (如离差平方和)最小



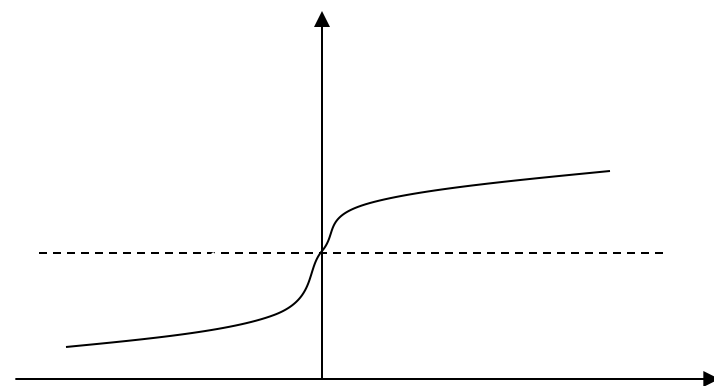
- 多元线性回归：

$$y = \beta_0 + \sum_i \beta_i x_i$$

Logistic 回归

- Logistic回归是一种非线性回归
- Logistic (也叫Sigmoid)函数(S型曲线):

$$y = f(x) = \frac{1}{1 + e^{-(\alpha + \beta x)}} = \frac{e^{\alpha + \beta x}}{1 + e^{\alpha + \beta x}}$$



- Logistic回归可以转化成线性回归来实现

$$\frac{y}{1-y} = e^{\alpha + \beta x}, \quad \ln \frac{y}{1-y} = \alpha + \beta x$$

Logistic 回归IR模型

- 基本思想：为了求 Q 和 D 相关的概率 $P(R=1|Q,D)$ ，通过定义多个特征函数 $f_i(Q,D)$ ，认为 $P(R=1|Q,D)$ 是这些函数的组合。
- Cooper等人提出：定义 $\log(P/(1-P))$ 为多个特征函数的线性组合。则 P 是一个Logistic函数，即：

$$\log \frac{P}{1-P} = \beta_0 + \sum_i \beta_i f_i(Q, D)$$

$$P = \frac{1}{1 + e^{-\beta_0 - \sum_i \beta_i f_i(Q, D)}}$$

*William S. Cooper , Fredric C. Gey , Daniel P. Dabney, Probabilistic retrieval based on staged logistic regression, Proceedings of ACM SIGIR'92, p.198-210, June 21-24, 1992, Copenhagen, Denmark

特征函数 f_i 的选择

$$X_1 = \frac{1}{M} \sum_1^M \log QAF_{t_j}$$

$$X_2 = \sqrt{QL}$$

$$X_3 = \frac{1}{M} \sum_1^M \log DAF_{t_j}$$

$$X_4 = \sqrt{DL}$$

$$X_5 = \frac{1}{M} \sum_1^M \log IDF_{t_j}$$

$$IDF = \frac{N - n_{t_j}}{n_{t_j}}$$

$$X_6 = \log M$$

- QAF: 查询平均词频
- DAF: 文档平均词频
- QL: 查询长度
- DL: 文档长度
- M: 词典大小

Logistic 回归IR模型(续)

- 求解和使用过程：
 - 通过训练集合拟和得到相应系数 $\beta_0 \sim \beta_6$ ，对于新的文档，代入公式计算得到概率 P
 - *Learning to Rank*中*Pointwise*方法中的一种
 - 判别式(discriminate)模型
- 优缺点：
 - 优点：直接引入数学工具，形式简洁。
 - 缺点：特征选择非常困难，实验中效果一般。
 - 以文档为样本(Pointwise)训练模型，无法解决不同查询之间的差异

提纲

- ① 向量空间模型回顾
- ② 基本概率统计知识
- ③ Logistic回归模型
- ④ BIM模型
- ⑤ BM25模型

二值独立概率模型BIM

- 二值独立概率模型(Binary Independence Model, 简称BIM): 伦敦城市大学Robertson及剑桥大学Sparck Jones 1970年代提出, 代表系统OKAPI
- 贝叶斯公式

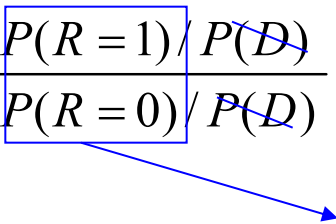
$$P(A|B) = \frac{P(A,B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

- BIM模型通过贝叶斯公式对所求条件概率 $P(R=1|Q,D)$ 展开进行计算。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 $P(R=1|D)$

BIM模型(续)

- 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

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。

有的资料将该评分函数称为Log-odd

文档是怎么生成的？

- 类比：
 - 班级是怎样构成的？
 - 博士是怎么读成的？
 -
- 概率的观点：
 - 词项满足某个总体分布，然后从该总体分布中抽样，将抽样出的词项连在一起，组成文档
 - 对于 $P(D|R=1)$ 或者 $P(D|R=0)$ ，可以认为 $R=1$ 或 0 的文档的词项满足某个总体分布，然后抽样生成 D

两种常用的文档生成的总体分布

- 多元贝努利分布(Multi-variate Bernoulli distribution)
 - 词项词典大小为 M ， M 个不规则硬币分别对应 M 个词项，第 i 个硬币朝上的概率为 p_i
 - 假设 $M=4$ (四个词项分别为 I you can fly), $p_1=0.7$, $p_2=0.4$, $p_3=0.1$, $p_4=0.05$
 - 则: $P(\text{I can fly fly})=0.7*(1-0.4)*0.1*0.05$
 - 多元贝努利分布不考虑出现位置
 - 多元贝努利分布考虑出现和不出现



两种常用的文档生成的总体分布(续)

- 多项式分布(Multinomial distribution)
 - 词项大小为M，某个不规则骰子共有M个面，每个面对应一个词项(假设每次抛掷必有某个面稳定朝上或下)，第*i*个面朝上的概率为 p_i
 - 假定M=4 (四个词项分别为 I you can fly), $p_1=0.4$, $p_2=0.3$, $p_3=0.2$, $p_4=0.1$
 - 则: $P(\text{I can fly fly})=P(X_1=1, X_2=0, X_3=1, X_4=2)$
 $=C \cdot 0.4 \cdot 0.2 \cdot 0.1 \cdot 0.1$
 - 其中 $C = C_4^1 \times C_3^1 \times C_2^2 = 12$
 - 多项式分布考虑词项的多次出现
 - 多项式分布不考虑词项的不出现
 - 多项式分布同样不考虑词项的出现位置和次序



BIM中 $P(D|R=1)$ 或 $P(D|R=0)$ 的计算

- 类比： M 次独立试验 (多元贝努利模型)
 - 假想词项空间中有 M 个词项，相当于有 M 个不规则硬币，第 i 个硬币对应词项 i ，正面写着“出现 t_i ”，反面写着“不出现 t_i ”，独立地抛这 M 个硬币，然后记录下每个硬币朝上的面对应的词项便组成文档 D 。
 - 因此，求 $P(D|R)$ 就是抛这个 M 个硬币得到 D 的概率。假设抛不同硬币之间是独立的(独立性假设)，并且不考虑 t_i 出现的次数，只考虑 t_i 要么出现要么不出现(二值)。同时，也不考虑抛硬币的次序(词袋模型)
 - $P(D|R=1)$ 和 $P(D|R=0)$ 相当于有两组硬币，因此需要求解 $2M$ 个概率参数

一个例子

- 查询为：信息 检索 教程

所有词项的在相关、不相关情况下的概率 p_i 、 q_i 分别为：

词项	信息	检索	教材	教程	课件
R=1时的概率 p_i	0.8	0.9	0.3	0.32	0.15
R=0时的概率 q_i	0.3	0.1	0.35	0.33	0.10

文档D1： 检索 课件

$$\text{则： } P(D|R=1)=(1-0.8)*0.9*(1-0.3)*(1-0.32)*0.15$$

$$P(D|R=0)= (1-0.3)*0.1*(1-0.35)*(1-0.33)*0.10$$

$$P(D|R=1)/P(D|R=0)=4.216$$

BIM模型公式的推导

将 D 看成 $\bigwedge_{t_i \in D} t_i \bigwedge_{t_j \notin D} \bar{t}_j$

$$\begin{aligned} P(D | R = 1) &= \prod_{t_i \in D} P(t_i | R = 1) \prod_{t_i \notin D} P(\bar{t}_i | R = 1) \\ &= \prod_{t_i} p_i^{e_i} (1 - p_i)^{1-e_i}, \text{if } t_i \in D \text{ then } e_i = 1, \text{else } e_i = 0 \end{aligned}$$

$$p_i = P(t_i | R = 1)$$

$$q_i = P(t_i | R = 0)$$

$$\begin{aligned} P(D | R = 0) &= \prod_{t_i \in D} P(t_i | R = 0) \prod_{t_i \notin D} P(\bar{t}_i | R = 0) \\ &= \prod_{t_i} q_i^{e_i} (1 - q_i)^{1-e_i}, \text{if } t_i \in D \text{ then } e_i = 1, \text{else } e_i = 0 \end{aligned}$$

注： $P(t_i|R=1)$ 表示在相关情况下， t_i 出现在文档中的概率(也就是说某个、或者某几个 $P(t_i|R=1)$ 可以为1)，注意：不是在相关文档集中出现的概率，因此所有 $P(t_i|R=1)$ 的总和不为1。这个可以和前面抛硬币的过程对照一下就明白了。

BIM模型公式的推导

继续推导，去掉公式中的只依赖查询 Q 的常数项，得所有出现在文档 $D(e_i=1)$ 中的词项的某个属性值之和。再假定对于不出现在 Q 中的词项，有 $p_i=q_i$ ，则得到所有出现在 $Q \cap D$ 中的词项的属性值之和

$$\begin{aligned}
 \log \frac{P(D | R=1)}{P(D | R=0)} &= \log \frac{\prod_{t_i \in D \cup \bar{D}} p_i^{e_i} (1-p_i)^{1-e_i}}{\prod_{t_i \in D \cup \bar{D}} q_i^{e_i} (1-q_i)^{1-e_i}} = \sum_{t_i \in D \cup \bar{D}} \log \left(\frac{p_i}{q_i} \right)^{e_i} \left(\frac{1-p_i}{1-q_i} \right)^{1-e_i} \quad \text{常数} \\
 &= \sum_{t_i \in D \cup \bar{D}} \left(e_i \log \frac{p_i}{q_i} + (1-e_i) \log \frac{1-p_i}{1-q_i} \right) = \sum_{t_i \in D \cup \bar{D}} \left(e_i \log \frac{p_i}{q_i} - e_i \log \frac{1-p_i}{1-q_i} + \log \frac{1-p_i}{1-q_i} \right) \quad \text{假设对不属于Q的 term, } p_i=q_i, \text{ 则此项为零} \\
 &\propto \sum_{t_i \in D \cup \bar{D}} \boxed{e_i \log \frac{p_i/(1-p_i)}{q_i/(1-q_i)}} = \sum_{t_i \in D} \log \frac{p_i/(1-p_i)}{q_i/(1-q_i)} = \sum_{t_i \in Q \cap D} \log \frac{p_i/(1-p_i)}{q_i/(1-q_i)} + \boxed{\sum_{t_i \notin Q \wedge t_i \in D} \log \frac{p_i/(1-p_i)}{q_i/(1-q_i)}} \\
 &\approx \sum_{t_i \in Q \cap D} \log \frac{p_i/(1-p_i)}{q_i/(1-q_i)} \quad \text{类似于向量内积计算} \\
 &= \sum_{t_i \in D \cap Q} W_i^{BIM} \quad \text{ } t_i \text{ 在 } Q \text{ 中权重, 只与 } Q \text{ 相关}
 \end{aligned}$$

最原始的BIM模型的计算公式，其中最关键是 p_i 、 q_i 的计算！

$p_i q_i$ 参数的计算

理想情况下，可以将整个文档集合根据是否和查询相关、是否包含 t_i 分成如下四个子集合，每个集合的大小已知。

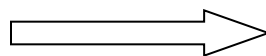
	相关 R_i (100)	不相关 $N-R_i$ (400)
包含 t_i n_i (200)	r_i (35)	$n_i - r_i$ (165)
不包含 t_i $N-n_i$ (300)	$R_i - r_i$ (65)	$N - R_i - n_i + r_i$ (235)

其中， N 、 n_i 分别是总文档以及包含 t_i 的文档数目。 R_i 、 r_i 分别是相关文档及相关文档中包含 t_i 的文档数目。括号中列举的数值是给出的一个总文档数目为500的计算例子。则：

$$p_i = \frac{r_i}{R_i} = \frac{35}{100} = 0.35$$

$$q_i = \frac{n_i - r_i}{N - R_i} = \frac{165}{400} = 0.413$$

平滑



$$p_i = \frac{r_i + 0.5}{R_i + 1}$$

$$q_i = \frac{n_i - r_i + 0.5}{N - R_i + 1}$$

p_i q_i 参数的计算(续)

- 由于在真实情况下，对于每个查询，无法事先得到相关文档集和不相关文档集，所以无法使用理想情况下的公式计算，因此必须进行估计
- 有多种估计方法
 - 初始检索：第一次检索之前的估计
 - 基于检索结果：根据上次检索的结果进行估计

p_i q_i 参数的计算(续)

- 初始情况：检索初始并没有相关和不相关文档集合，此时可以进行假设： p_i 是常数， q_i 近似等于term i 在所有文档集合中的分布(假定相关文档很少， $R_i=r_i=0$)

$$p_i = 0.5$$

$$q_i = \frac{n_i}{N}$$

$$\begin{aligned} \sum_{t_i \in D \cap Q} \log \frac{p_i / (1 - p_i)}{q_i / (1 - q_i)} &= \sum_{t_i \in D \cap Q} \log \frac{N - n_i}{n_i} \\ &\approx \sum_{t_i \in D \cap Q} \log \frac{N - n_i + 0.5}{n_i + 0.5} = \sum_{t_i \in D \cap Q} W_i^{IDF} \end{aligned}$$

因此，BIM在初始假设情况下，其检索公式实际上相当于对所有同时出现在 q 和 d 中的词项的IDF的求和

RSJ权重

- Robertson & Spärck Jones权重 (RSJ权重)

$$W_i^{RSJ} = \log \frac{(r_i + 0.5)(N - R - n_i + r_i + 0.5)}{(n_i - r_i + 0.5)(R - r_i + 0.5)}$$



$p_i q_i$ 参数的计算(续)

- 基于前面的检索结果：假定检索出的结果集合 V (可以把 V 看成全部的相关文档集合)，其中集合 V_i 包含 term i ，则可以进一步进行计算
- 为避免较小的 V 和 V_i 集合，加入常数或非常数平滑因子(以下用 V 和 V_i 表示同名集合的大小)

$$\begin{array}{ccc}
 p_i = \frac{V_i}{V} & \xrightarrow{\text{平滑方式1}} & p_i = \frac{V_i + 0.5}{V + 1} \\
 q_i = \frac{n_i - V_i}{N - V} & \searrow \text{平滑方式2} & q_i = \frac{n_i - V_i + 0.5}{N - V + 1}
 \end{array}$$

$$\begin{array}{ccc}
 & & p_i = \frac{V_i + \frac{n_i}{N}}{V + 1} \\
 & & q_i = \frac{n_i - V_i + \frac{n_i}{N}}{N - V + 1}
 \end{array}$$

BIM模型小结

- BIM计算过程：目标是求排序函数 $P(D|R=1)/P(D|R=0)$
 - 首先估计或计算每个term分别在相关文档和不相关文档中的出现概率 $p_i=P(t|R=1)$ 及 $q_i=P(t|R=0)$
 - 然后根据独立性假设，将 $P(D|R=1)/P(D|R=0)$ 转化为 p_i 和 q_i 的某种组合，将 p_i 和 q_i 代入即可求解。

BIM模型的优缺点

- 优点：
 - BIM模型建立在数学基础上，理论性较强
- 缺点：
 - 需要估计参数
 - 原始的BIM没有考虑TF、文档长度因素
 - BIM中同样存在词项独立性假设
 - BIM实质上是一个idf权重公式，仅考虑了全局信息，缺少局部信息。因此需要和TF权重配合使用

提纲

- ① 向量空间模型回顾
- ② 基本概率统计知识
- ③ Logistic回归模型
- ④ BIM模型
- ⑤ BM25模型

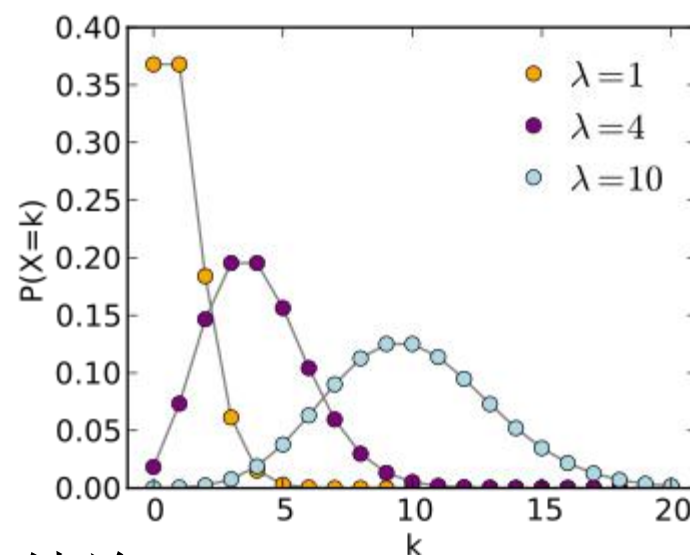
二重泊松分布 (2-Poisson Distribution)

- 泊松分布是一个经典的随机分布
- 分布公式参数：均值 λ
- 分布形式随参数取值变化

$$f(k; \lambda) = \Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

- 二重泊松分布

- 关于文本中词频分布的一个经典结论
- 在高质量精英文档集(Elite Set)中：均值较高，接近正态分布
- 在整个语料中：均值低，接近指数分布



基于二重泊松模型的 tf 评分建模

- 考虑 tf ，则RSJ公式变为：

$$w = \log \frac{p(1-q)}{q(1-p)} = \log \frac{p_{tf} q_0}{q_{tf} p_0}$$

- 将二重泊松模型带入到上述公式，我们可以得到：

$$w = \log \frac{(p' \lambda^{tf} e^{-\lambda} + (1-p') \mu^{tf} e^{-\mu}) (q' e^{-\lambda} + (1-q') e^{-\mu})}{(q' \lambda^{tf} e^{-\lambda} + (1-q') \mu^{tf} e^{-\mu}) (p' e^{-\lambda} + (1-p') e^{-\mu})},$$

- λ 和 μ 分别是在elite set和整个语料中的泊松分布均值参数， $p' = (\text{doc elite for } t \mid R)$, q' 是 \bar{R} 中相应的概率
- 估计问题：四个参数，均无直接计算依据
 - Elite是一个隐含变量

一个粗略的 tf 模型

- 上面的二重泊松模型公式可变换为:

$$w = \log \frac{(p' + (1 - p')(\mu/\lambda)^{tf} e^{\lambda - \mu}) (q' e^{\mu - \lambda} + (1 - q'))}{(q' + (1 - q')(\mu/\lambda)^{tf} e^{\lambda - \mu}) (p' e^{\mu - \lambda} + (1 - p'))}$$

- μ 远小于 λ . 当 $tf \rightarrow \infty$, $(\mu/\lambda)^{tf}$ 趋近于0, 因此这些部分都可以省略掉. 同时 $e^{(\mu - \lambda)}$ 很小, 因此以上公式可化简为:

$$w = \log \frac{p'(1 - q')}{q'(1 - p')} \sim \log \frac{p(1 - q)}{q(1 - p)}$$

- 频繁词项? (μ 远小于 λ 的假设不成立)
 - 使用Okapi BM25时, 需要去除停用词保证模型可靠性

对化简后二重泊松模型的估计

- 基本原则
 - 当 $tf=0$ ，权重为0
 - 权重对 tf 单调递增
 - 但增幅逐渐降低直至某最大值
 - 是对RSJ权重的估计，并且是 $eliteness$ 的直接指标

一个简单的 tf 权重公式

- 基于以上四项基本原则，有以下 tf 权重公式：

$$w = \frac{tf}{k_1 + tf} w^{(1)}$$

- 基本原则
 - 当 $tf=0$ ，权重为0
 - 权重对 tf 单调递增
 - 但增幅逐渐降低直至某最大值
 - 是对RSJ权重的估计，并且是 $eliteness$ 的直接指标

查询词项权重 (Query Term Weighting)

- 采用类似的权重函数：

$$w = \frac{qtf}{k_3 + qtf} \cdot \frac{tf}{k_1 + tf} \cdot w^{(1)}$$

- qtf : 词项在查询中的词频(通常都是1)
- k_3 : 与 k_1 类似的可调参数，通常取值从8到1000不等
 - 在此范围内变动对检索效果影响不大
- 考虑了在查询中不同词项的相对重要性
 - 但仅仅基于 qtf ，很难真正区分
 - 可以通过相关反馈修改查询词项权重，从而有效区分词项的重要性

文档长度问题

- 与向量空间模型中长度归一化类似的问题，概率的解释
- 泊松分布假设了罐子/抽屉模型 (Urn Model)
 - 文档：罐子/抽屉
 - 词项：相同颜色的球
- tf 分布描述的是文档 d 中观察到 tf 个颜色为 t 的球的概率
 - 先验概率：若有 N 个罐子，将一个颜色为 t 的球随机扔出，落入罐子 d 的概率
- 泊松分布假设先验概率服从均匀分布，即 $P(d)=1/N$
- 但在实际应用中并不严格成立，因此需要长度归一化/长度规整化(length normalization)

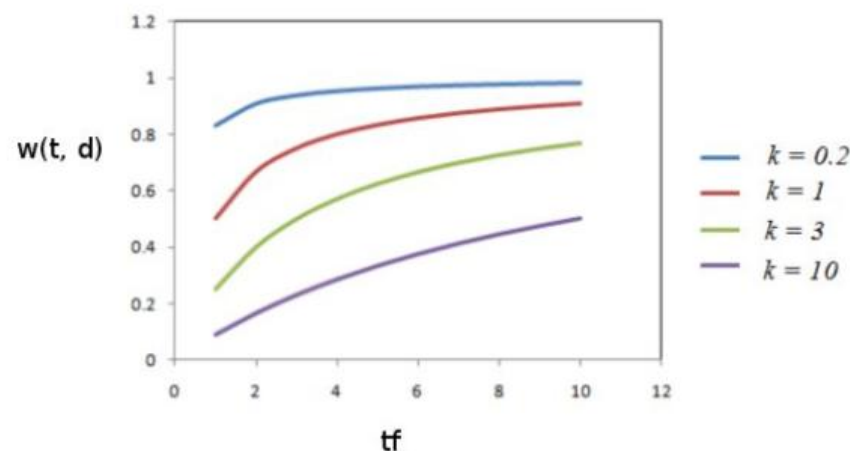
Okapi's TF (S. E. Robertson 1990s)

- 加入文档长度因素，进一步得到：

$$TF = \frac{k_1 \cdot tf}{tf + k_1 \left(1 - b + b \frac{l_d}{avg_l} \right)}$$

也叫饱和函数 (saturation function)

- l_d : 文档长度
- avg_l : 平均文档长度



BM25最终公式（引入了查询权重和长度）

$$score(Q, d) = \sum_{t \in Q} w(t, d)$$

TF权重（基于二重泊松分布）

查询权重

$$w(t, d) = \frac{qtf}{k_3 + qtf} \cdot \frac{k_1 \cdot tf}{tf + k_1(1 - b + b \cdot l_d / avg_l)} \cdot w^{(1)}$$

$$w^{(1)} = \log_2 \frac{N - df + 0.5}{df + 0.5}$$

IDF权重，基于BIM模型

- $w^{(1)}$: 当 $R=r=0$ 时的RSJ权重； qtf : 查询中词频；
- tf : 文档中词频； l_d : 文档长度； avg_l : 平均文档长度
- N : 文档数量； df : 文档频率； b, k_1, k_3 : 可调参数

BM25的实验结果

- BM25在一系列TREC竞赛项目中得到了最好结果
 - TREC-4, 5 ad-hoc tasks
 - Web tracks
 - TREC-2009 relevance feedback track
 - 等等.....
- 在大规模文本检索评测语料上被证明非常有效

BM25: Pros & Cons

- 优点
 - 一定程度上的理论化模型
 - 基于二重泊松假设——适用于绝大多数文本语料上的IR检索应用
 - 实验证明有效
- 缺点
 - 待调参数多且参数敏感性高
 - 必须去停用词

BM25被视为现实应用中最好的IR模型之一。即便现在基于BERT预训练语言模型的方法可以获得更好的效果，仍然需要使用BM25进行无监督过滤来保证检索精度。

BM25与概率模型：参考文献

- Harter S.P. A probabilistic approach to automatic keyword indexing. Journal of the American Society for Information Science 1975; 26:197–206 and 280–289
- Robertson S.E., Van Rijsbergen C.J. & Porter M.F. Probabilistic models of indexing and searching. Information Retrieval Research (pp.35–56). Butterworths London, 1981
- Robertson S.E. and Sparck Jones K. Relevance weighting of search terms. Journal of the American Society for Information Science 1976; 27:129–146
- S. E. Robertson and S. Walker. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. ACM SIGIR 1994.
- S. E. Robertson, S. Walker, M. M. Beaulieu, M. Gatford and A. Payne. Okapi at TREC-4. 1995.

近期研究进展

- 传统文档长度归一方法对不同文档采用同样的归一化函数（例如向量空间模型的Cosine归一法）
- 但是不同的作者的写作风格不一，冗长(Verbosity)程度不一
- 近期一些研究试图在TF-IDF和语言模型中引入冗长信息
- Jiaul H. Paik. A Novel TF-IDF Weighting Scheme for Effective Ranking. SIGIR 2013.
- Ronan Cummins, Jiaul H. Paik, Yuanhua Lv. A Pólya Urn Document Language Model for Improved Information Retrieval, TOIS 2015.

参考资料

- 《现代信息检索》第11章
- BM25模型的推导
 - S.E Robertson and S. Walker, Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval, SIGIR'94
 - S.E Robertson, S. Walker, S. Jones, Okapi at TREC-3, in Proceedings of TREC-3
- DFR模型
 - Type II模型: Gianni Amati, C. J. van Rijsbergen: Probabilistic models of information retrieval based on measuring the divergence from randomness. ACM Trans. Inf. Syst. 20(4): 357-389 (2002)
 - Type I模型: Gianni Amati: Frequentist and Bayesian Approach to Information Retrieval. ECIR 2006: 13-24
- 近年研究进展
 - Jiaul H. Paik. A Novel TF-IDF Weighting Scheme for Effective Ranking. SIGIR 2013.
 - Ronan Cummins, Jiaul H. Paik, Yuanhua Lv. A Pólya Urn Document Language Model for Improved Information Retrieval, TOIS 2015.