# Liam's Blog

陈亮亮




# BM25 概率检索模型介绍

Dec 8, 2016

BM25 和向量空间模型（Vector Space Model, VSM）都是经典的信息检索模型，在信息检索领域得到了广泛采用。相比较而言，BM25 比向量空间模型有更坚实的数学理论基础，并且在文档长度、词项频率等对相关性影响判断方面有更优异的表现。这里对 BM25 概率检索模型的原理做一个简单介绍。

## 1、基础知识

了解 BM25 算法需要有一定的概率论基础知识。为了便于后续理解，我们这里对涉及到的概念做一个简单回顾，进一步的了解请参考专业的书籍。

假定变量 A 代表一个事件，它是所有可能的结果构成的空间上的一个子集。同样，我们可以通过随机变量（random variable）来代表该子集，它是从结果到实数的一个映射函数。对应该子集，随机变量 A 会取一个具体的值。通常我们并不能确信某个事件是否为真，此时，我们想知道事件 A 发生的概率 P(A)，它满足 $0 \leq P(A) \leq 1$。对于两个事件 A、B，它们的联合事件发生的可能性通过联合概率 P(A,B) 来描述。条件概率 $P(A|B)$ 表示在事件 B 发生的条件下 A 发生的概率。联合概率和条件概率的关系可以通过如下的链式法则（chain rule）来体现：

$$P(AB) = P(A \bigcap B) = P(A|B)P(B) = P(B|A)P(A)$$

上式表明，在不做任何假设的条件下，两个事件的联合概率等于其中一个事件的概率乘上在该事件发生的条件下另一个事件发生的条件概率。

事件 A 的补集的概率记为 $P(\overline{A})$，同样有

$$P(\overline{A}B) = P(B|\overline{A})P(\overline{A})$$

概率论中的全概率定理（partition rule）为：如果事件 B 可以划分成互不兼容（即互斥）的多个子事件，那么 B 的概率将是所有子事件概率的和。下式给出的是该定理的一个特例：

$$P(B) = P(AB) + P(\overline{A}B)$$

基于上述结果可以推导出贝叶斯定理（Bayes's rule）：

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)}=\left[\sum_{X\in\{A,\overline{A}\}}\frac{P(B|A)}{P(B|X)P(X)}\right]P(A)$$

最后值得一提的是另外一个常用的概念是事件的优势率（odds），它提供了一种反映概率如何变化的“放大器”（multiplier）：

$$\text{优势率：}O(A)=\frac{P(A)}{P(\overline{A})}=\frac{P(A)}{1-P(A)}$$

## 2、二值独立模型

假设有一个用户的查询请求 q，和文档集中的一篇文档 d，利用概率模型来估计每篇文档和查询请求的相关概率 $P(R=1|d,q)$（R 代表 d 和 q 的相关性关系，并假设为二值，要么相关，要么不相关，当 d 和 q 相关时记 R=1，不相关时记 R=0），然后根据相关性大小对结果进行排序。这就是概率排序的原理（Probability Ranking Principle, PRP）

概率排序是一种直接对用户需求相关性进行建模的方法，不同于向量空间模型以查询和文档的相似性来作为相关性的代替品。但概率排序模型只是一种指导思想，实际中，该如何对用户需求和文档的相关性进行建模呢？这里介绍的二值独立模型（Binary Independence Model, BIM）就是传统上随同 PRP 一起使用的一种模型。

在二值独立模型中，“二值”等价于布尔值：文档和查询都表示为词项出现与否的布尔向量。也就是说，文档 d 表示为向量$\vec{x}=(x_1,\cdots,x_M)$，其中当词项 t 出现在文档 d 中时，$x_t=1$，否则 $x_t=0$。类似地，我们将查询 q 表示成词项出现向量 $\vec{q}$。“独立性” 指的是词项在文档中的出现是互相独立的，BIM 并不识别词项之间的关联。另外，我们再假设每篇文档的相关性与其它文档的相关性无关。至此，在 BIM 模型下，我们可以基于词项出现向量的概率 $P(R|\vec{x},\vec{q})$ 对概率 $P(R|d,q)$ 建模。然后，利用贝叶斯定理，有

$$P(R=1|\vec{x},\vec{q})=\frac{P(\vec{x}|R=1,\vec{q})|P(R=1|\vec{q})}{P(\vec{x},\vec{q})} \tag{1}$$

$$P(R=0|\vec{x},\vec{q})=\frac{P(\vec{x}|R=0,\vec{q})|P(R=0|\vec{q})}{P(\vec{x},\vec{q})} \tag{2}$$

其中，$P(\vec{x}|R=1,\vec{q})$ 和 $P(\vec{x}|R=0,\vec{q})$ 分别表示当返回一篇相关或不相关文档时文档表示为 $\vec{x}$ 的概率。

### 2.1、排序函数的推导

> 排序函数的推导较多的引用了<信息检索导论>（见参考文献[1]）中的相关内容，更详细的信息可以参考其“概率检索模型”章节。

给定查询 q，我们将按照 $P(R=1|d,q)$ 从高到低将所有文档排序。在 BIM 模型下，也就是要按照 $P(R=1|\vec{x},\vec{q})$ 排序。由于检索系统关心的只是文档的相对次序，所以这里并不需要直接估计出这个概率值，而是采用其它的更容易计算的排序函数，这中间只需要保证采用排序函数和直接计算概率所得到的文档次序一致即可。具体地，我们可以根据文档相关性的优势率来对文档排序，它是相关性概率的单调递增函数，这样的话就可以忽略公式 (1) 和 (2) 中的公共分母，使得计算起来更容易。文档相关性的优势率定义如下：

$$O(R|\vec{x},\vec{q})=\frac{P(R=1|\vec{x},\vec{q})}{P(R=0|\vec{x},\vec{q})}=\frac{\frac{P(\vec{x}|R=1,\vec{q})|P(R=1|\vec{q})}{P(\vec{x},\vec{q})}}{\frac{P(\vec{x}|R=0,\vec{q})|P(R=0|\vec{q})}{P(\vec{x},\vec{q})}}=\frac{P(R=1|\vec{q})}{P(R=0|\vec{q})}\cdot\frac{P(\vec{x}|R=1,\vec{q})}{P(\vec{x}|R=0,\vec{q})} \quad (3)$$

对于给定查询来说，上述公式中的 $\frac{P(R=1|\vec{q})}{P(R=0|\vec{q})}$ 是个常数。由于我们只关注文档排序，所以没有必要估计这个常数。因此，我们只需要估计 $\frac{P(\vec{x}|R=1,\vec{q})}{P(\vec{x}|R=0,\vec{q})}$，但是这个问题看上去很难，因为我们要精确估计整个词项出现向量的概率。这里我们引入了朴素贝叶斯条件独立性假设（Naive Bayes conditional independence assumption），即在给定查询的情况下，认为一个词的出现与否与任意一个其他词的出现与否是互相独立的，即

$$\frac{P(\vec{x}|R=1,\vec{q})}{P(\vec{x}|R=0,\vec{q})}=\prod_{t=1}^{M}\frac{P(x_t|R=1,\vec{q})}{P(x_t|R=0,\vec{q})}$$

因此根据公式 (3)，有

$$O(R|\vec{x},\vec{q})=O(R|\vec{q})\cdot\prod_{t=1}^{M}\frac{P(x_t|R=1,\vec{q})}{P(x_t|R=0,\vec{q})}$$

由于每个 $x_t$ 的取值要么为 0，要么为 1，所以有

$$O(R|\vec{x},\vec{q})=O(R|\vec{q})\cdot\prod_{t:x_t=1}\frac{P(x_t=1|R=1,\vec{q})}{P(x_t=1|R=0,\vec{q})}\cdot\prod_{t:x_t=0}\frac{P(x_t=0|R=1,\vec{q})}{P(x_t=0|R=0,\vec{q})}$$

简便起见，记 $p_t=P(x_t=1|R=1,\vec{q})$，即词项出现在一篇相关文档中的概率，同样令 $u_t=P(x_t=1|R=0,\vec{q})$，即词项出现在一篇不相关文档中的概率。这些值之间的关系展示在如下的列联表中，其中每列值的和为 1。

| 文档 | 相关 (R=1) | 不相关 (R=0) |
|---|---|---|
| 词项出现 $x_t=1$ | $p_t$ | $u_t$ |
| 词项不出现 $x_t=0$ | $1-p_t$ | $1-u_t$ |

可以对上述式子进行进一步的简化，假定没有在查询中出现的词项在相关和不相关文档中出现的概率相等，即当 $q_t=0$ 时，$p_t=u_t$。因此，我们只需要考虑在查询中出现的词项的概率的乘积，于是有

$$O(R|\vec{x},\vec{q}) = O(R|\vec{q}) \cdot \prod_{t:x_t=q_t=1} \frac{p_t}{u_t} \cdot \prod_{t:x_t=0,q_t=1} \frac{1-p_t}{1-u_t}$$

其中， 第二个因子计算的是出现在文档中的查询词项的概率乘积，而第 3 个因子计算的是不出现在文档中的查询词项的概率乘积。对上述公式可以进一步转换，有

$$O(R|\vec{x},\vec{q}) = O(R|\vec{q}) \cdot \prod_{t:x_t=q_t=1} \frac{p_t(1-u_t)}{u_t(1-p_t)} \cdot \prod_{t:q_t=1} \frac{1-p_t}{1-u_t}$$

此时，可以发现公式中的第 3 个因子只和查询词项 $q_t$ 有关，考虑的只是所有查询词项 。因此，对于给定的查询而言，第 3 个因子就像 $O(R|\vec{q})$ 一样是个常数，可以约去。所以文档排序中唯一需要估计的量就剩下了第 2 个因子。这个最后用于排序的量也被称为“检索状态值”（Retrieval Status Value, RSV） 。对其取 log 值后

$$RSV_d = \log \prod_{t:x_t=q_t=1} \frac{p_t(1-u_t)}{u_t(1-p_t)} = \sum_{t:x_t=q_t=1} \log \frac{p_t(1-u_t)}{u_t(1-p_t)} = \sum_{t:x_t=q_t=1} \log \frac{p_t}{1-p_t} + \log \frac{1-u_t}{u_t}$$

这就是 BIM 概率模型中用于计算排序的主要公式。

### 2.2、二值独立模型的不足

在二值独立模型中，为了能够在实际中对概率函数 $P(R|d,q)$ 进行估计，二值独立模型中引入了很多简单的假设。这些假设包括：

- 文档、查询及相关性的布尔表示；
- 词项的独立性；
- 查询中不出现的词项不会影响最后的结果；
- 不同文档的相关性之间是互相独立的。

这些假设显然和实际很不相符。比如文档、查询及相关性的布尔表示，在这种表示下，由于不考虑词项出现的次数及顺序，许多不同的文档可能都有相同的向量表示。另外词项的独立性假设和实际也很不相符，相同一篇文档中的词项通常都有前后的关联性（但词项的独立性假设在实际中常常却能给出令人满意的结果） 。不同文档的相关性之间是互相独立的这个假设显然也是错误的，当允许系统返回冗余或者近似冗余文档时，这个假设在实际当中尤其有害。或许正是因为这些假设过于严厉，因此在现实中 BIM 模型的检索效果往往难以达到较好的水平。

## BM25概率检索模型

虽然 BIM 模型计算相关性的实际效果并不好，但 BIM 却是 BM25 模型的基础，BM25 模型是在 BIM 公式的变体上经过一系列实验诞生的。BM25 在 BIM 的基础上加入了词项频率（tf）和文档长度等统计因子，拟合出了一个综合排序公式，并通过实验引入了一些经验参数。根据 TREC评测结果表明，BM25 是目前最好的内容排序模型。

关于 BM25 概率模型具体的实验及推导过程我们这里就不去介绍了，感兴趣的同学可以参考 S. Robertson 的相关论文（见参考文献[3]），最终的 BM25 概率排序模型的公式是这样的：

$$RSV_d = \sum_{t \in q} \left[ \log \frac{N}{df_t} \right] \cdot \frac{\log(k_1+1)tf_{td}}{k_1\left[(1-b)+b\times(L_d/L_{ave})\right]+tf_{td}} \cdot \frac{(k_3+1)tf_{tq}}{k_3+tf_{tq}}$$

其中，第一个因子即为上面介绍的 BIM 模型公式。但由于这里已对其进行了计算，所以显示的上述有些不一致。我们假设所有查询中的词项的 $p_t$ 是个常数 $p_t = 0.5$ （也即相关文档中出现的每个词项都有一个等值的优势率），所以 $p_t$ 和 $(1-p_t)$ 因子可以在 RSV 的计算公式中约掉（这种估计比较弱）。另外我们再假设相关文档只占所有文档的极小一部分，那么就可以通过整个文档集的统计数字来计算与不相关文档有关的量。在该假设下，在某查询下不相关文档中出现词项 t 的概率 $u_t = df_t/N$，于是有 $\log[(1-u_t)/u_t] = \log[(N-df_t)/df_t] \approx \log N/df_t$，也就是上面显示的第一个因子了。

第二个因子和第三个因子分别是引入了词项频率和文档长度的文档权重因子和查询权重因子（如果查询很短，也可以忽略这个权重）。第二个因子中，$tf_{td}$ 是词项 t 在文档 d 中的权重，$L_d$ 和 $L_{ave}$ 分别是文档 d 的长度及整个文档集中文档的平均长度。$k_1$ 是一个取正值的调优参数，用于对文档中的词项频率进行缩放控制。如果 $k_1$ 取 0，则对应 BIM 模型；如果 $k_1$ 取较大的值，那么对应于使用原始词项频率。b 是另外一个调节参数（$0 \le b \le 1$），决定文档长度的缩放程度：b = 1 表示基于文档长度对词项权重进行完全的缩放，b = 0 表示归一化时不考虑文档长度因素。在第三个因子中，$tf_{tq}$ 是词项 t 在查询 q 中的权重。这里 $k_3$ 是另一个取正值的调优参数，用于对查询中的词项频率进行缩放控制。

理想的情况下，$k_1, b, k_3$ 这几个公式中的经验参数可以在各个产品各自的数据集中进行结果优化而得到，也就是通过真实的数据迭代调优来实现最大化的检索效果。如果没有条件这么做，那么也可以根据现有的一些实验结果来取值，这些参数的一个合理的取值范围是：$k_1$ 和 $k_3$ 的取值区间为 1.2~2，b 取 0.75。

BM25 概率排序模型推出后，获得了极大的成功，被多个研究队伍作为权重计算公式并广泛的用于不同的搜索任务系统中。陆续也有也有基于 BM25 的改进算法出现，比如 BM25F，一个针对有多个文档域（field）场景的算法改进等。

## 3. 参考文献

[1]. 信息检索导论
[2]. 现代信息检索
[3]. S. Robertson and K. S. Jones. Relevance weighting of search terms. Journal of the American Society for Information Sciences. 1976