第11章　如何确定网页和查询的相关性

前面已经谈过了如何自动下载网页、如何建立索引、如何衡量网页的质量（PageRank）。接下来谈谈如何确定一个网页和某个查询的相关性。了解了这四个方面，有一定编程基础的读者就可以写出一个简单的搜索引擎了，比如为自己所在的学校或院系搭建一个小型搜索引擎。

我们还是看看前面介绍的例子，查找关于“原子能的应用”的网页。第一步是在索引中找到包含这三个词的网页（详见第8章关于布尔运算的内容）。现在任何一个搜索引擎能提供几十万甚至是上百万个与这个查询词组多少有点关系的网页，比如Google返回了大约一千万个结果。那么哪个应该排在前面呢？显然应该把网页本身质量好，且网页和查询关键词“原子能的应用”相关性高的网页排在前面。第10章已经介绍了如何度量网页的质量。这里介绍另外一个关键技术：如何度量网页和查询的相关性。

1　搜索关键词权重的科学度量TF-IDF

短语“原子能的应用”可以分成三个关键词：原子能、的、应用。根据直觉，我们知道，包含这三个词较多的网页应该比包含它们较少的网页相关。当然，这个办法有一个明显的漏洞，那就是内容长的网页比内容短的网页占便宜，因为长的网页总的来讲包含的关键词要多些。因此，需要根

据网页的长度，对关键词的次数进行归一化，也就是用关键词的次数除以网页的总字数。我们把这个商称为“关键词的频率”，或者“单文本词频”（Term Frequency），比如，某个网页上一共有1000词，其中“原子能”、“的”和“应用”分别出现了2次、35次和5次，那么它们的词频就分别是0.002、0.035和0.005。将这三个数相加，其和0.042就是相应网页和查询“原子能的应用”的“单文本词频”。

因此，度量网页和查询的相关性，有一个简单的方法，就是直接使用各个关键词在网页中出现的总词频。具体地讲，如果一个查询包含N个关键词$w_1, w_2, ..., w_N$，它们在一个特定网页中的词频分别是：$TF_1, TF_2, ..., TF_N$。（TF：Term Frequency，是词频一词的英文缩写）。那么，这个查询和该网页的相关性（即相似度）就是：

$$TF_1 + TF_2 + ... + TF_N \qquad (11.1)$$

读者可能已经发现了又一个漏洞。在上面的例子中，“的”这个词占了总词频的80%上，而它对确定网页的主题几乎没什么用处。我们称这种词叫“停止词”（Stop Word），也就是说，在度量相关性时不应考虑它们的频率。在汉语中，停止词还有“是”、“和”、“中”、“地”、“得”等几十个。忽略这些停止词后，上述网页和查询的相关性就变成了0.007，其中“原子能”贡献了0.002，“应用”贡献了0.005。

细心的读者可能还会发现另一个小漏洞。在汉语中，“应用”是个很通用的词，而“原子能”是个很专业的词，后者在相关性排名中比前者重要。因此，需要对汉语中的每一个词给一个权重，这个权重的设定必须满足下面两个条件：

1. 一个词预测主题的能力越强，权重越大，反之，权重越小。在网页中看到“原子能”这个词，或多或少能了解网页的主题。而看到“应用”一词，则对主题基本上还是一无所知。因此，“原子能”的权重就应该比应用大。

2. 停止词的权重为零。

很容易发现，如果一个关键词只在很少的网页中出现，通过它就容易锁定搜索目标，它的权重也就应该大。反之，如果一个词在大量网页中出现，看到它仍然不很清楚要找什么内容，因此它的权重就应该小。

概括地讲，假定一个关键词 w 在 D_w 个网页中出现过，那么 D_w 越大，w 的权重越小，反之亦然。在信息检索中，使用最多的权重是“逆文本频率指数”（Inverse Document Frequency，缩写为 IDF），它的公式为 $\log\left(\frac{D}{D_w}\right)$，其中 D 是全部网页数。比如，假定中文网页数是 D = 10 亿，停止词“的”在所有的网页中都出现，即 D_w = 10 亿，那么它的 $IDF = \log(10\text{亿} / 10\text{亿}) = \log(1) = 0$。假如专用词“原子能”在 200 万个网页中出现，即 D_w = 200 万，则它的权重 $IDF = \log(500) = 8.96$。又假定通用词“应用”出现在五亿个网页中，它的权重 $IDF = \log(2)$，则只有 1。

也就是说，在网页中找到一个“原子能”的命中率（Hits）相当于找到九个“应用”的命中率。利用 IDF，上述相关性计算的公式就由词频的简单求和变成了加权求和，即

$$TF_1 \cdot IDF_1 + TF_2 \cdot IDF_2 + \cdots + TF_N \cdot IDF_N \qquad (11.2)$$

在上面的例子中，该网页和“原子能的应用”的相关性为 0.0161，其中“原子能”贡献了 0.0126，而“应用”只贡献了 0.0035。这个比例和我们的直觉比较一致了。

TF-IDF（Term Frequency / Inverse Document Frequency）的概念被公认为信息检索中最重要的发明。在搜索、文献分类和其他相关领域有着广泛的应用。讲起 TF-IDF 的历史蛮有意思。IDF 的概念最早是剑桥大学的斯巴克·琼斯[1]（Karen Spärck Jones）提出来的。斯巴克·琼斯 1972 年在一篇题为“关键词特殊性的统计解释和它在文献检索中的应用”的论文中提出 IDF 的概念。遗憾的是，她既没有从理论上解释为什么权

1 斯巴克·琼斯，剑桥大学计算机女科学家，最著名的言论：“计算机是如此重要，因此不能只留给男人去做！”在程序界广为流传。

重 IDF 应该是对数函数 $\log\left(\frac{D}{D_w}\right)$（而不是其他函数，比如平方根 $\sqrt{\frac{D}{D_w}}$），也没有在这个题目上作进一步的深入研究，以至于在以后的很多文献中人们提到 TF-IDF 时没有引用她的论文，绝大多数人甚至不知道斯巴克·琼斯的贡献。同年剑桥大学的罗宾逊写了一个两页纸的解释，解释得很不好。倒是后来康奈尔大学的萨尔顿（Salton）多次撰文、写书讨论 TF-IDF 在信息检索中的用途，加上萨尔顿本人的大名（信息检索领域的世界级大奖就是以萨尔顿的名字命名的），很多人都引用萨尔顿的书，甚至以为这个信息检索中最重要的概念是他提出的。当然，世界并没有忘记斯巴克·琼斯的贡献。2004 年，在纪念《文献学学报》创刊 60 周年之际，该学报重印了斯巴克·琼斯的大作。罗宾逊在同期期刊上写了篇文章，用香农的信息论解释 IDF，这回的解释是对的，但文章写得并不好，非常冗长（足足 18 页），把简单问题搞复杂了。其实，信息论的学者们已经发现并指出，所谓 IDF 的概念就是一个特定条件下关键词的概率分布的交叉熵（Kullback-Leibler Divergence）（详见本书第 6 章“信息的度量和作用”）。这样，关于信息检索相关性的度量，又回到了信息论。

现在的搜索引擎对 TF-IDF 进行了不少细微的优化，使得相关性的度量更加准确了。当然，对有兴趣写一个搜索引擎的爱好者来讲，使用 TF-IDF 就足够了。如果结合网页排名（PageRank）算法，那么给定一个查询，有关网页的综合排名大致由相关性和网页排名的乘积决定。

2 延伸阅读：TF-IDF 的信息论依据

读者背景知识：信息论和概率论。

一个查询（Query）中每一个关键词（Key Word）w 的权重应该反映这个词对查询来讲提供了多少信息。一个简单的办法就是用每个词的信息量作为它的权重，即

$$I(w) = -P(w)\log P(w)$$

$$= -\frac{TF(w)}{N}\log\frac{TF(w)}{N} = \frac{TF(w)}{N}\log\frac{N}{TF(w)} \qquad (11.3)$$

其中，N 是整个语料库的大小，是个可以省略的常数。上面的公式可以简化成

$$I(w) = TF(w)\log\frac{N}{TF(w)} \qquad (11.4)$$

但是，公式（11.4）有一个缺陷：两个词出现的频率 TF 相同，一个是某篇特定文章中的常见词，而另外一个词是分散在多篇文章中，那么显然第一个词有更高的分辨率，它的权重应该更大。显然，更好的权重公式应该反映出关键词的分辨率。

如果做一些理想的假设，

1）每个文献大小基本相同，均为 M 个词，即 $M = \frac{N}{D} = \frac{\sum_w TF(w)}{D}$。

2）一个关键词在文献一旦出现，不论次数多少，贡献都等同，这样一个词要么在一个文献中出现 $c(w) = \frac{TF(w)}{D(w)}$ 次，要么是零。注意，$c(w) < M$，那么从公式（11.4）出发可以得到下面的公式：

$$TF(w)\log\frac{N}{TF(w)} = TF(w)\log\frac{MD}{c(w)D(w)}$$

$$= TF(w)\log(\frac{D}{D(w)}\frac{M}{c(w)}) \qquad (11.5)$$

这样，我们看到 TF-IDF 和信息量之间的差异就是公式（11.6）中的第二项。因为 $c(w) < M$，所以第二项大于零，它是 $c(w)$ 的递减函数。把上面的公式重写成

$$TF\text{-}IDF(w) = I(w) - TF(w)\log\frac{M}{c(w)} \qquad (11.6)$$

可以看到，一个词的信息量 $I(w)$ 越多，TF-IDF 值越大；同时 w 命中的文献中 w 平均出现的次数越多，第二项越小，TF-IDF 也越大。这些结论和信息论完全相符。

3 小结

TF-IDF 是对搜索关键词的重要性的度量，从理论上讲，它有很强的理论根据。因此，如果对搜索不是很精通的人，直接采用 TF-IDF 效果也不会太差。现在各家搜索引擎对关键词重要性的度量，都在 TF-IDF 的基础上有些改进和微调。但是，在原理上与 TF-IDF 相差不远。

参考文献：

1. Spärck Jones, Karen "A statistical interpretation of term specificity and its application in retrieval". Journal of Documentation28 (1): 11-21, 1972

2. Salton, G. and M. J. McGill, *Introduction to modern information retrieval*. McGraw-Hill, 1986

3. H.C. Wu, R.W.P. Luk, K.F. Wong, K.L. Kwok "Interpreting tf idf term weights as making relevance decisions". *ACM Transactions on Information Systems* 26 (3): 1-37, 2008