SEO技术：文本相似度-bm25算法原理及实现

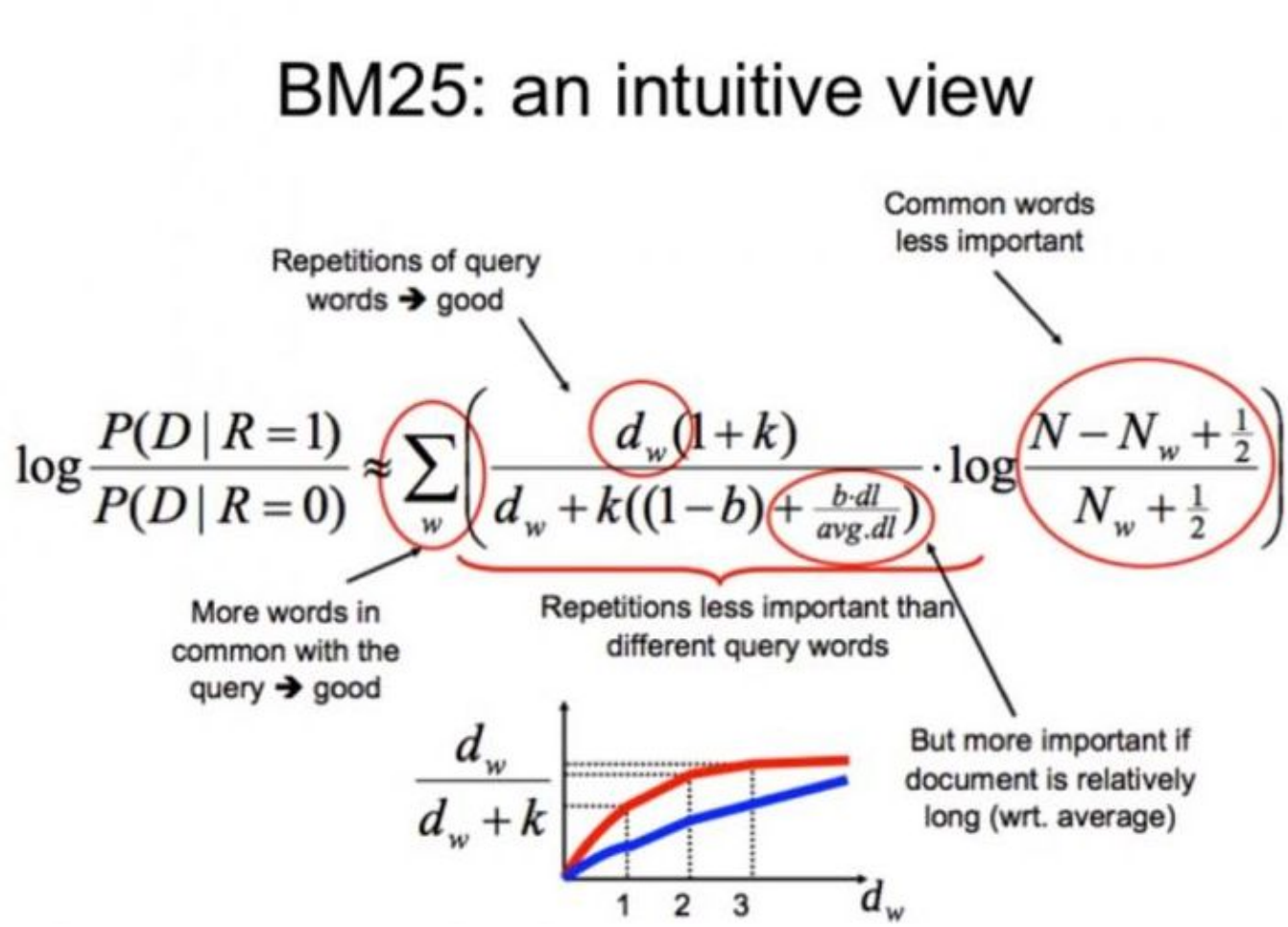
无忧 每天学点SEO 2019-10-27

前面提到过TF-IDF算法（TF-IDF算法原理及公式）与之更进一步算法BM25相关度也是处理关键词相关性中重要的算法其中。

那么 TF 和 IDF 谁更重要呢，怎么计算最终的相关性得分呢？那就是 BM25。

BM25算法，通常用来作搜索相关性平分。一句话概况其主要思想：对Query进行语素解析，生成语素qi；然后，对于每个搜索结果D，计算每个语素qi与D的相关性得分，最后，将qi相对于D的相关性得分进行加权求和，从而得到Query与D的相关性得分。

本文整理了多篇有关BM25相关度算法原理形成本文供各位SEOer阅读，内容比较深度也非常的装逼，反正无忧是看不懂。有兴趣的站长可以查看研究一下。



BM25算法原理及实现

原理

BM25算法，通常用来作搜索相关性平分。一句话概况其主要思想：对Query进行语素解析，生成语素qi；然后，对于每个搜索结果D，计算每个语素qi与D的相关性得分，最后，将qi相对于D的相关性得分进行加权求和，从而得到Query与D的相关性得分。
BM25算法的一般性公式如下：

其中，Q表示Query，qi表示Q解析之后的一个语素（对中文而言，我们可以把对Query的分词作为语素分析，每个词看成语素qi。）；d表示一个搜索结果文档；Wi表示语素qi的权重；R(qi，d)表示语素qi与文档d的相关性得分。
下面我们来看如何定义Wi。判断一个词与一个文档的相关性的权重，方法有多种，较常用的是IDF。这里以IDF为例，公式如下：

其中，N为索引中的全部文档数，n(qi)为包含了qi的文档数。
根据IDF的定义可以看出，对于给定的文档集合，包含了qi的文档数越多，qi的权重则越低。也就是说，当很多文档都包含了qi时，qi的区分度就不高，因此使用qi来判断相关性时的重要度就较低。
我们再来看语素qi与文档d的相关性得分R（qi，d）。首先来看BM25中相关性得分的一般形式：

$$R(q_i,d)=\frac{f_i\cdot(k_1+1)}{f_i+K}\cdot\frac{qf_i\cdot(k_2+1)}{qf_i+k_2}$$

$$K=k_1\cdot(1-b+b\cdot\frac{dl}{avgdl})$$

其中，k1，k2，b为调节因子，通常根据经验设置，一般k1=2，b=0.75；fi为qi在d中的出现频率，qfi为qi在Query中的出现频率。dl为文档d的长度，avgdl为所有文档的平均长度。由于绝大部分情况下，qi在Query中只会出现一次，即qfi=1，因此公式可以简化为：

从K的定义中可以看到，参数b的作用是调整文档长度对相关性影响的大小。b越大，文档长度的对相关性得分的影响越大，反之越小。而文档的相对长度越长，K值将越大，则相关性得分会越小。这可以理解为，当文档较长时，包含qi的机会越大，因此，同等fi的情况下，长文档与qi的相关性应该比短文档与qi的相关性弱。
综上，BM25算法的相关性得分公式可总结为：

从BM25的公式可以看到，通过使用不同的语素分析方法、语素权重判定方法，以及语素与文档的相关性判定方法，我们可以衍生出不同的搜索相关性得分计算方法，这就为我们设计算法提供了较大的灵活性。

BM25相关度打分公式

BM25算法是一种常见用来做相关度打分的公式，思路比较简单，主要就是计算一个query里面所有词和文档的相关度，然后在把分数做累加操作,而每个词的相关度分数主要还是受到tf/idf的影响。公式如下：

$$Score(Q,d)=\sum_{i}^{n} W_i \cdot R(q_i,d)$$

R(qi,d)是每个词和文档的相关度值，其中qi代表每个词，d代表相关的文档，Wi是这个词的权重，然后所有词的乘积再做累加。

自然语言处理-BM25相关度打分

BM25（Best Match25）是在信息检索系统中根据提出的query对document进行评分的算法。It is based on the probabilistic retrieval framework developed in the 1970s and 1980s by Stephen E. Robertson, Karen Spärck Jones, and others.BM25算法首先由OKapi系统实现，所以又称为OKapi BM25。

BM25属于bag-of-words模型，bag-of-words模型只考虑document中词频，不考虑句子结构或者语法关系之类，把document当做装words的袋子，具体袋子里面可以是杂乱无章的。It is not a single function, but actually a whole family of scoring functions, with slightly

different components and parameters. One of the most prominent instantiations of the function is as follows.

对于一个query Q, 包括关键字 $q_1, \cdots, q_n$, 一个文档的BM25得分:

$$\text{score}(D,Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{\text{avgdl}})},$$

其中IDF是上篇文章《TD-IDF》中的IDF，f是《TD-IDF》中的TF，|D|是文档D的长度，avgdl是语料库全部文档的平均长度。k_1和b是参数。usually chosen, in absence of an advanced optimization, as $k_1 \in [1.2, 2.0]$ and $b = 0.75$。

相关性

对每一个搜索查询，我们很容易给每个文档定义一个“相关分数”。当用户进行搜索时，我们可以使用相关分数进行排序而不是使用文档出现时间来进行排序。这样，最相关的文档将排在第一个，无论它是多久之前创建的（当然，有的时候和文档的创建时间也是有关的）。

有很多很多种计算文字之间相关性的方法，但是我们要从最简单的、基于统计的方法说起。这种方法不需要理解语言本身，而是通过统计词语的使用、匹配和基于文档中特有词的普及率的权重等情况来决定“相关分数”。

这个算法不关心词语是名词还是动词，也不关心词语的意义。它唯一关心的是哪些是常用词，那些是稀有词。如果一个搜索语句中包括常用词和稀有词，你最好让包含稀有词的文档的评分高一些，同时降低常用词的权重。

这个算法被称为 Okapi BM25。它包含两个基本概念 词语频率(term frequency) 简称词频 (“TF”) 和 文档频率倒数(inverse document frequency) 简写为(“IDF”). 把它们放到一起，被称为 “TF-IDF”，这是一种统计学测度，用来表示一个词语 (term) 在文档中有多重要。

TF-IDF

词语频率(Term Frequency), 简称 “TF”, 是一个很简单的度量标准：一个特定的词语在文档出现的次数。你可以把这个值除以该文档中词语的总数，得到一个分数。例如文档中有 100 个词， ‘the’ 这个词出现了 8 次，那么 ‘the’ 的 TF 为 8 或 8/100 或 8%（取决于你想怎么表示它）。

逆向文件频率（Inverse Document Frequency），简称 “IDF”，要复杂一些：一个词越稀有，这个值越高。它由总文件数目除以包含该词语之文件的数目，再将得到的商取对数得到。越

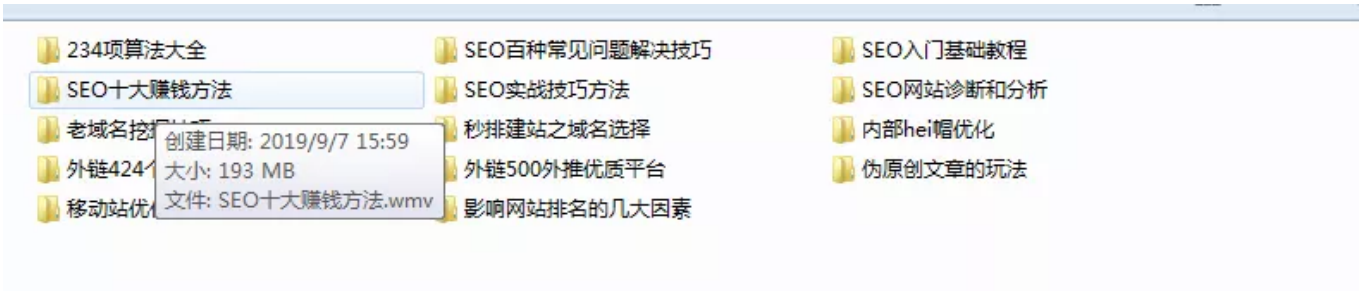
是稀有的词，越会产生高的 “IDF”。

如果你将这两个数字乘到一起 (TF*IDF), 你将会得到一个词语在文档中的权重。“权重”的定义是：这个词有多稀有并且在文档中出现的多么频繁？

你可以将这个概念用于文档的搜索查询。在查询中的对于查询中的每个关键字，计算他们的 TF-IDF 分数，并把它们相加。得分最高的就是与查询语句最符合的文档。

整理了相关的资料，乱七八糟写了这么多，能看懂的一定是大神级别了，供各位SEO从业者学习借鉴。有什么困惑欢迎留言各位探讨。

下面是老陈在公司整理到的一些行业优化教程和SEO工具包（部分截图） 扫描下方二维码即可免费领取，前3名还有神秘豪华礼包哦！



里面有100节SEO真人课程，是我们团队花费240多天制作的，课程涵盖数十种网站优化方法，课程文件加起来有30G，识别下方二维码即可免费领取学习！

END

原文及出处：https://www.ainiseo.com/suanfa/5339.html