BM25算法浅析

(/p/7)

标签： **BM25** (/t/BM25) **算法** (/t/%E7%AE%97%E6%B3%95)

BM25 is a bag-of-words retrieval function that ranks a set of documents based on the query terms appearing in each document, regardless of the inter-relationship between the query terms within a document (e.g., their relative proximity). It is not a single function, but actually a whole family of scoring functions, with slightly different components and parameters.

BM25算法，通常用来作搜索相关性评分。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：

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

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：

$$\mathrm{IDF}(q_i)=\log\frac{N-n(q_i)+0.5}{n(q_i)+0.5},$$

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：

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

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：

$$Score(Q,d)=\sum_{i}^{n}IDF(q_i)\cdot\frac{f_i\cdot(k_1+1)}{f_i+k_1\cdot(1-b+b\cdot\frac{dl}{avgdl})}$$

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。

来源：http://ipie.blogbus.com/logs/104136815.html

Alvan (/@Alvan) 发布于 2013-09-03 14:18

← 上一篇 (/p/6)

下一篇 → (/p/8)