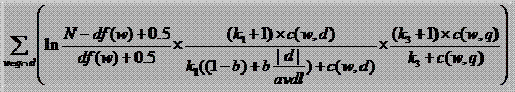
**Okapi BM25公式详解：**

[](http://by1.storage.msn.com/y1py6mK33o__4f5ENrfPKs6PZHx96EO-cNuPZD2Hp2ExTu8I3YxufWoKtPr3hl05jjM5SOserQMH6nKcAv9bfGdikeJrHgVrEsp?PARTNER=WRITER)

其中*q*、*d*、*C*分别表示查询、文档和文档集合，|*q*|和|*d*|分别是查询*q*和文档*d*的长度，*avdl*是文档集合中文档的平均长度。*w*表示特征词项(Term)，*c*(*w*,*d*)和*c*(*w*,*q*)分别表示*w*出现在*d*和*q*中的个数，*N*是文档集合中的文档总数，*df*(*w*)表示出现*w*的文档个数。*p*(*w*|*C*)表示*C*中出现*w*的概率。公式中的*k*1,*k*3,*b*，都是人工经验调节的参数。

BM25模型是一个经验启发式模型，存在着巨大的鲁棒性问题，具体表现在：

(1) 这些模型有需要人工经验调节的参数。同一模型在面对同一查询、同一文档集的情况下，不同的参数其取值结果性能相差较大；

(2) 如果参数不变，同一模型对同一查询，在不同的文档集上结果也会性能相差很大；

(3) 同一模型对不同查询，同一组参数，同一文档集合上的结果性能也会相差很大。

BM25F 对比BM25 增加了域加权

