搜索评价指标——NDCG

概念

NDCG，Normalized Discounted cumulative gain 直接翻译为归一化折损累计增益，可能有些晦涩，没关系下面重点来解释一下这个评价指标。这个指标通常是用来衡量和评价搜索结果算法（注意这里维基百科中提到了还有推荐算法，但是我个人觉得不太适合推荐算法，后面我会给我出我的解释）。DCG的两个思想：

1、高关联度的结果比一般关联度的结果更影响最终的指标得分；

2、有高关联度的结果出现在更靠前的位置的时候，指标会越高；

累计增益（CG）

CG，cumulative gain，是DCG的前身，只考虑到了相关性的关联程度，没有考虑到位置的因素。它是一个搜素结果相关性分数的总和。指定位置p上的CG为：

$$\mathrm{CG_p} = \sum_{i=1}^{p} rel_i$$

rel_i 代表i这个位置上的相关度。

举例：假设搜索"篮球"结果，最理想的结果是：B1、B2、 B3。而出现的结果是 B3、B1、B2的话，CG的值是没有变化的，因此需要下面的DCG。

折损累计增益（DCG）

DCG， Discounted 的CG，就是在每一个CG的结果上处以一个折损值，为什么要这么做呢？目的就是为了让排名越靠前的结果越能影响最后的结果。假设排序越往后，价值越低。到第i个位置的时候，它的价值是 $1/\log_2(i+1)$，那么第i个结果产生的效益就是 $rel_i * 1/\log_2(i+1)$，所以：

$$\mathrm{DCG_p} = \sum_{i=1}^{p} \frac{rel_i}{\log_2(i+1)} = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2(i+1)}$$

当然还有一种比较常用的公式，用来增加相关度影响比重的DCG计算方式是：

$$\mathrm{DCG_p} = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{\log_2(i+1)}$$

百科中写到后一种更多用于工业。当然相关性值为二进制时，即 rel_i在{0,1}，二者结果是一样的。当然CG相关性不止是两个，可以是实数的形式。

归一化折损累计增益（NDCG）

NDCG， Normalized 的DCG，由于搜索结果随着检索词的不同，返回的数量是不一致的，而DCG是一个累加的值，没法针对两个不同的搜索结果进行比较，因此需要归一化处理，这里是处以IDCG。

$$\mathrm{nDCG_p} = \frac{DCG_p}{IDCG_p}$$

IDCG为理想情况下最大的DCG值。

$$\mathrm{IDCG_p} = \sum_{i=1}^{|REL|} \frac{2^{rel_i} - 1}{\log_2(i+1)}$$

其中 |REL| 表示，结果按照相关性从大到小的顺序排序，取前p个结果组成的集合。也就是按照最优的方式对结果进行排序。

实际的例子

假设搜索回来的5个结果，其相关性分数分别是 3、2、3、0、1、2

那么 CG = 3+2+3+0+1+2

可以看到只是对相关的分数进行了一个关联的打分，并没有召回的所在位置对排序结果评分对影响。而我们看DCG：

i	reli	$\log_2(i+1)$	$rel_i / \log_2(i+1)$
1	3	1	3
2	2	1.58	1.26
3	3	2	1.5
4	0	2.32	0
5	1	2.58	0.38
6	2	2.8	0.71

所以 DCG = 3+1.26+1.5+0+0.38+0.71 = 6.86

接下来我们归一化，归一化需要先结算 IDCG，假如我们实际召回了8个物品，除了上面的6个，还有两个结果，假设第7个相关性为3，第8个相关性为0。那么在理想情况下的相关性分数排序应该是：3、3、3、2、2、1、0、0。计算IDCG@6：

i	reli	$\log_2(i+1)$	$rel_i / \log_2(i+1)$
1	3	1	3
2	3	1.58	1.89
3	3	2	1.5
4	2	2.32	0.86
5	2	2.58	0.77
6	1	2.8	0.35

所以IDCG = 3+1.89+1.5+0.86+0.77+0.35 = 8.37

so 最终 NDCG@6 = 6.86/8.37 = 81.96%