20161111实验结果

/Users/luocheng/Documents/coderepo/mobilesearch/metric_design/docs/20161111/metric随阈值变化的曲线.pdf

根据周中与刘老师讨论的，首先现在对metric与SBS是这样定义的。

定义1：

当采取相对的判断时，

采取绝对的判断时，

我们选取几个有代表性的指标，ERR， AP， nDCG@0003，以及我们自己开发的ERP和Nor-tbg-raw-step，在图中分别用不同的颜色代表指标，实线是绝对判断，虚线是相对判断。

可以得到以下这么几点结论：

1. 不同的指标在ratio变化的时候，虽然采用绝对和相对的判断有所不同，但是总体的趋势基本是一致的，这代表不同的指标在于SBS拟合的性能上天然存在差别。
2. 可以看到我们之前使用的0.05的这条线上，在相对判断的方法上，ERR是最好的。但是当ratio进一步增大的时候，Nor-tbg-raw-step的绝对和相对的数值超越了ERR，最终的性能也比ERR要更好。
3. 从绝对和相对两种方法来看，采用绝对数值的方法更好。因为这避免了一些SBS=0，同时两种评价指标的数值都比较低的情况下可能判断为不一致的情况。
4. 回顾我们之前仔细看过的采用ERR、相对判断、ratio=0.05的情况下，18个badcase中有10个case都是因为SBS=0，但是metric!=0。随着ratio增大，这样的case会被进一步解决。

这里面有一个问题是，我们在这里定义的一致性，并不能直接拿来预测SBS，因为在实际预测SBS的时候，是不知道SBS的值的。

进一步地，如果只用Metric数据预测SBS，可以有两种不同策略：

定义2：

如果

相关的结果请见下面两张图：

硬分类：

predict_hard.pdf

可以看到此时，随着threshold的增大，一致性先增大，后减小；先增大是因为相对来说宽松了判定为SBS=0的条件，这时候有一些metric差别很小的，在threshold很小的时候预测SBS!=0变成SBS=0，如果threshold过于大，会让更多的pair判定为SBS=0，与SBS标注不一致。

不同评价指标来看，采用Nor-tbg-raw-step-ABS的指标最好，比ERR还要显著地好。

软分类：

predict_soft.pdf

软分类这个做法不大妥当，也不太能放到真实环境下使用，因为相对来说在后面的case会越来越少，因为大部分的case都被预测为SBS=0不纳入统计范围了。这里想看的是，假设我们认为diff(metrics(A,B)) < threshold 的不能自动判定，需要人工判定对剩下自动预测的精度有多少。可以看Nor-tbg-step-ABS与0.9的交点，大概是threshold=0.06，这时候大概还剩下60%的case，可以对他们进行精度为90%的预测。