20161105

移动评价指标设计方面，首先本周修改判断和SBS指标一致性的标准，认为如果SBS=0且|ma-mb| < 0.05\*max(ma, mb)的情况下，也是一致的。也就是说，如果sbs=0时，如果metric得到的结果在5%以内，可以认为metric预测的结果也认为AB两个list是没有差别的。这样的话新的Metric v.s. Preference的结果如下：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Metric** | **Agree** | **Disagree** | **Metric** | **Agree** | **Disagree** |
| **ERR** | **132** | **18** | Nor-lwfv-exp-binom | 111 | 39 |
| **nERR@0005** | **132** | **18** | Nor-lwfv-raw-binom | 111 | 39 |
| Nor-tbg-raw-step | 127 | 23 | MSnDCG@0003 | 110 | 40 |
| Nor-tbg-raw-binom | 126 | 24 | NCUgu,BR | 107 | 43 |
| **nERR@0003** | **124** | **26** | RBP | 106 | 44 |
| **NCUrb,BR** | **122** | **28** | nDCG@0005 | 105 | 45 |
| Q@0005 | 120 | 30 | P-measure | 102 | 48 |
| Q-measure | 120 | 30 | P-plus | 102 | 48 |
| Nor-nDCG-raw-binom | 119 | 31 | nDCG@0003 | 101 | 49 |
| Nor-lwofv-raw-step | 119 | 31 | O-measure | 96 | 54 |
| Nor-lwofv-raw-binom | 119 | 31 | NCUgu,P | 96 | 54 |
| Q@0003 | 118 | 32 | **ERP** | **81** | **69** |
| Nor-nDCG-raw-step | 118 | 32 | NCUrb,P | 72 | 78 |
| Nor-tbg-exp-binom | 114 | 36 | AP@0005 | 72 | 78 |
| Nor-lwofv-exp-binom | 112 | 38 | AP | 72 | 78 |
| Nor-lwfv-exp-step | 111 | 39 | P@0005 | 70 | 80 |
| Nor-lwofv-exp-step | 111 | 39 | AP@0003 | 69 | 81 |
| Nor-lwfv-raw-step | 111 | 39 | P@0003 | 68 | 82 |
| Nor-tbg-exp-step | 111 | 39 | RR | 57 | 93 |
| MSnDCG@0005 | 111 | 39 | Hit@0003 | 46 | 104 |
| Nor-nDCG-exp-step | 111 | 39 | Hit@0005 | 45 | 105 |
| Nor-nDCG-exp-binom | 111 | 39 |  |  |  |

可以看出排在靠前位置的NCU-family（ERR，NCU）的指标表现都很好，这个意思是说如果用metric预测SBS，准确率可以将近90%，如果可以进一步在大范围的数据上验证，应该已经是一个可以使用的指标了。

分析其中NCU family的指标特性，有这么几个特点：

1. 这些指标不包含recall的部分，包括NDCG等其他指标，其实都在ideal 结果列表中包含了Recall的部分。NCU完全不考虑。这与用户在移动搜索中的情景是比较一致的。用户更在乎查准，并不太在乎查全。

2. NCU指标有一个Diminishing Return的特性，即如果两个结果R1,R2，如果R1是相关的，访问到R2的概率会极大地衰减；反之如果R1是不相关的，R2访问到的概率比较大，此外通过1/R来衰减结果位置（固有的）的影响。这样的一种设计事实上强调了上下文的因素，在相关结果后的结果的重要性相对不重要了。这也和用户在移动场景下使用搜索引擎的一些特点比较一致。

稍微改动了一些ERR，生成了一个新的指标ERP，主要是：1) 手动标注了一些  
“一次点击即满足需求”的结果，让用户满意的概率为1. 2) 原有的根据rank衰减的函数改成了根据position，采用类似TBG的指数函数进行衰减。这个结果表现比较差。可能的原因是标注人员在做SBS标注的时候，不会完全忽略“一次点击即满足需求的结果”之后的结果，这和评价指标的假设不一样。

总体上看ERR还是一个比较有价值的改进方向。对ERR里面的不一致的情况作了分析：

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Target System | Query | Consist | Sogou Relevance | Target System Relevance | SBS | ERR Sogou | ERR Target System |
| baidu | 0023:笃字的拼音 | P=M> | 4.5|4|3.5|2|2.5 | 4.5|2|3.5|0|0 | 0 | 0.7492 | 0.6827 |
| baidu | 0043:我的世界籽岷 | P<M= | 4|4|3.5|2|2 | 4.5|3|3|2|4 | -1 | 0.7492 | 0.733 |
| sm | 0002:49天之同人 | P=M< | 0|0|0|2|0 | 0|0|2|0|0 | 0 | 0.05 | 0.0667 |
| sm | 0008:球球大作战兑换码礼品 | P=M> | 0|0|0|2|3 | 0|0|0|0|2 | 0 | 0.114 | 0.04 |
| sm | 0010:wipe cache partition | P<M= | 3|3|4|3|0 | 3|4|0|3|0 | -1 | 0.6064 | 0.604 |
| sm | 0019:长江水文网 | P=M> | 5|2|3|2|0 | 3|2|2|3|2 | 0 | 0.8461 | 0.5396 |
| sm | 0024:天气预报济南 | P=M< | 3|3|4|4|3 | 4.5|3|4|3|4 | 0 | 0.6182 | 0.7445 |
| sm | 0031:果糖的主要功能 | P=M< | 0|4|4|4|3 | 3|2|3|3|0 | 0 | 0.4091 | 0.5528 |
| sm | 0032:上海商学院志愿者服务网 | P>M= | 4.5|0|2.5|3|3 | 4.5|0|4.5|3|0 | 1 | 0.674 | 0.696 |
| sm | 0035:申通快递 | P=M< | 4.5|5|3|2|3.5 | 5|5|2|2|4 | 0 | 0.7761 | 0.8873 |
| sm | 0043:我的世界籽岷 | P<M= | 4|4|3.5|2|2 | 4.5|2|4|3|4 | -1 | 0.7492 | 0.726 |
| sm | 0048:罪恶成魔 | P=M> | 4|3|3|3|2 | 3.5|3|3|3|3.5 | 0 | 0.7299 | 0.6 |
| haosou | 0002:49天之同人 | P=M> | 0|0|0|2|0 | 0|0|0|0|0 | 0 | 0.05 | 0 |
| haosou | 0004:广西流浪汉月半小夜曲快手 | P>M< | 0|4|4|2|4 | 3|0|0|2|0 | 2 | 0.4034 | 0.43 |
| haosou | 0020:梦见失去工作 | P>M< | 2.5|4|3|4|3 | 3|3|4|0|4 | 1 | 0.5176 | 0.6093 |
| haosou | 0024:天气预报济南 | P=M< | 3|3|4|4|3 | 4|4|4|3|4 | 0 | 0.6182 | 0.763 |
| haosou | 0042:韩剧步步惊心丽 | P>M= | 4.5|3|3|2|3 | 4.5|3|3|2|3 | 1 | 0.7284 | 0.7284 |
| haosou | 0043:我的世界籽岷 | P<M= | 4|4|3.5|2|2 | 4.5|3|3|2|3 | -1 | 0.7492 | 0.7284 |

可以看到，比较多的不一致(10/18)还是因为SBS=0，但是ERR预测有差别。其他一些原因:

1. 在结果特别差的时候（0002sm, 0008sm, 0002haosou）ERR测的比SBS深，总会体现差别；
2. SBS更侧重Top1或者Top2的结果，位于标注系统第一个Viewport；
3. SBS存在一些误差。