**Mobile Evaluation Metric 实验结果**

第一部分实验：

依据搜狗同事提供的标注结果（50个样本查询），包含Sogou, SM, Haosou, Baidu 四个列表，对于每个列表，针对top 5 结果标注了得分（近似认为是相关性），并提供了Sogou与其他三家公司的Side-by-Side的倾向性得分。

* 1. 采用不同的采用相关性的评价指标，不同的评价指标之间是否一致：

这一方法的不足在于：异质结果的相关性可能会不能比较。可行性：搜狗同事在标注的时候对于不同长度的结果进行了切分，比如A结果的Rank 1相当于B结果的Rank1+Rank2 的高度，将对A的Rank 1 进行切分成两个结果评价。这样可以认为，**在结果高度这个意义上，不同的结果之间是近似可比的。**这样的场景接近于传统的ten blue links评价的场景。截断选择了3和5，近似认为是第一页结果和前两页结果。

使用NTCIR-Eval工具，使用了Adhoc搜索评价常用的指标，选取了其中常用的部分， 在50个查询上，对于不同的来源的列表计算Kendall’s tau ，在所有的topic上取平均，列在表1 中。

从结果可以看到，

1. 表格中高亮了几个主要的区域，第一部分是nDCG-family的结果互相一致程度较高。AP&P-family的结果互相一致程度较高，Q-measure主要的区别是decay function 有不同，结果和nDCG-family的结果接近。
2. RBP和两个版本的nDCG都是在@5的时候效果更好，在@3的时候区分度差。
3. P-measure和O-measure互相比较一致，和其他的指标一致性很差。

**表1. 不同评价指标之间的一致性（部分评价指标） Agreement Between Metrics Pairs**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **dumb** | **MSnDCG**  **@0003** | **MSnDCG**  **@0005** | **nDCG**  **@0003** | **nDCG**  **@0005** | **P**  **@0003** | **P**  **@0005** | **AP** | **AP**  **@0003** | **AP**  **@0005** | **RBP** | **P-measure** | **O-measure** | **Q-measure** |
| **MSnDCG**  **@0003** | 1.000 | 0.750 | 0.952 | 0.799 | 0.684 | 0.507 | 0.581 | 0.699 | 0.581 | 0.742 | 0.635 | 0.640 | 0.816 |
| **MSnDCG**  **@0005** |  | 1.000 | 0.746 | 0.917 | 0.536 | 0.566 | 0.549 | 0.532 | 0.549 | 0.959 | 0.643 | 0.617 | 0.833 |
| **nDCG**  **@0003** |  |  | 1.000 | 0.830 | 0.735 | 0.556 | 0.617 | 0.741 | 0.617 | 0.750 | 0.539 | 0.535 | 0.803 |
| **nDCG**  **@0005** |  |  |  | 1.000 | 0.601 | 0.601 | 0.598 | 0.597 | 0.598 | 0.935 | 0.565 | 0.533 | 0.859 |
| **P@0003** |  |  |  |  | 1.000 | 0.795 | 0.904 | 0.985 | 0.904 | 0.550 | 0.255 | 0.310 | 0.747 |
| **P@0005** |  |  |  |  |  | 1.000 | 0.936 | 0.795 | 0.936 | 0.579 | 0.171 | 0.203 | 0.720 |
| **AP** |  |  |  |  |  |  | 1.000 | 0.918 | 1.000 | 0.564 | 0.204 | 0.241 | 0.753 |
| **AP@0003** |  |  |  |  |  |  |  | 1.000 | 0.918 | 0.546 | 0.251 | 0.317 | 0.761 |
| **AP@0005** |  |  |  |  |  |  |  |  | 1.000 | 0.564 | 0.204 | 0.241 | 0.753 |
| **RBP** |  |  |  |  |  |  |  |  |  | 1.000 | 0.572 | 0.541 | 0.809 |
| **P-measure** |  |  |  |  |  |  |  |  |  |  | 1.000 | 0.945 | 0.612 |
| **O-measure** |  |  |  |  |  |  |  |  |  |  |  | 1.000 | 0.605 |
| **Q-measure** |  |  |  |  |  |  |  |  |  |  |  |  | 1.000 |

* 1. 计算不同的评价指标和SBS倾向的一致性

针对Sogou v.s X，对于每一个指标，一共有150个观测点（50 queries \* 3 X system）。 如果评价指标和SBS的倾向性是完全一致的，则认为是一个正例，否则认为是一个负例。认为人工标注的结果是Ground Truth，计算评价指标的Accuracy，结果如表2：

**表2 评价指标和SBS的一致性（Consistency between Metrics and Preference）**

|  |  |  |  |
| --- | --- | --- | --- |
| **Metrics** | **Accuracy** | **Metrics** | **Accuracy** |
| *nERR@0003* | 0.720 | *Q@0005* | 0.660 |
| *Q@0003* | 0.707 | *nDCG@0005* | 0.647 |
| *MSnDCG@0003* | 0.693 | *O-measure* | 0.640 |
| *ERR* | 0.680 | *NCUgu,P* | 0.627 |
| *P-measure* | 0.680 | *NCUrb,P* | 0.480 |
| *P-plus* | 0.680 | *AP@0005* | 0.480 |
| *nERR@0005* | 0.680 | *AP* | 0.480 |
| *NCUgu,BR* | 0.673 | *AP@0003* | 0.473 |
| *MSnDCG@0005* | 0.673 | *P@0005* | 0.467 |
| *NCUrb,BR* | 0.673 | *P@0003* | 0.467 |
| *nDCG@0003* | 0.673 | *RR* | 0.380 |
| *RBP* | 0.667 | *Hit@0003* | 0.320 |
| *Q-measure* | 0.660 | *Hit@0005* | 0.300 |

从上面的结果可以看到，与SBS分数一致性最高的结果是nERR@3，nDCG类的指标也可以达到接近0.7的水平。相对来说，@3的指标都要比@5更好，这有可能和标注的方式有关系。