**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Agreement（部分评价指标）**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **dumb** | **MSnDCG**  **@3** | **MSnDCG**  **@5** | **nDCG**  **@3** | **nDCG**  **@5** | **P**  **@3** | **P**  **@5** | **AP** | **AP**  **@3** | **AP**  **@5** | **RBP** | **P-measure** | **O-measure** | **Q-measure** |
| **MSnDCG@3** | 1. | 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@5** |  | 1. | 0.746 | 0.917 | 0.536 | 0.566 | 0.549 | 0.532 | 0.549 | 0.959 | 0.643 | 0.617 | 0.833 |
| **nDCG@3** |  |  | 1. | 0.830 | 0.735 | 0.556 | 0.617 | 0.741 | 0.617 | 0.750 | 0.539 | 0.535 | 0.803 |
| **nDCG@5** |  |  |  | 1. | 0.601 | 0.601 | 0.598 | 0.597 | 0.598 | 0.935 | 0.565 | 0.533 | 0.859 |
| **P@3** |  |  |  |  | 1. | 0.795 | 0.904 | 0.985 | 0.904 | 0.550 | 0.255 | 0.310 | 0.747 |
| **P@5** |  |  |  |  |  | 1. | 0.936 | 0.795 | 0.936 | 0.579 | 0.171 | 0.203 | 0.720 |
| **AP** |  |  |  |  |  |  | 1. | 0.918 | 1. | 0.564 | 0.204 | 0.241 | 0.753 |
| **AP@3** |  |  |  |  |  |  |  | 1. | 0.918 | 0.546 | 0.251 | 0.317 | 0.761 |
| **AP@5** |  |  |  |  |  |  |  |  | 1. | 0.564 | 0.204 | 0.241 | 0.753 |
| **RBP** |  |  |  |  |  |  |  |  |  | 1. | 0.572 | 0.541 | 0.809 |
| **P-measure** |  |  |  |  |  |  |  |  |  |  | 1. | 0.945 | 0.612 |
| **O-measure** |  |  |  |  |  |  |  |  |  |  |  | 1. | 0.605 |
| **Q-measure** |  |  |  |  |  |  |  |  |  |  |  |  | 1. |

* 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**

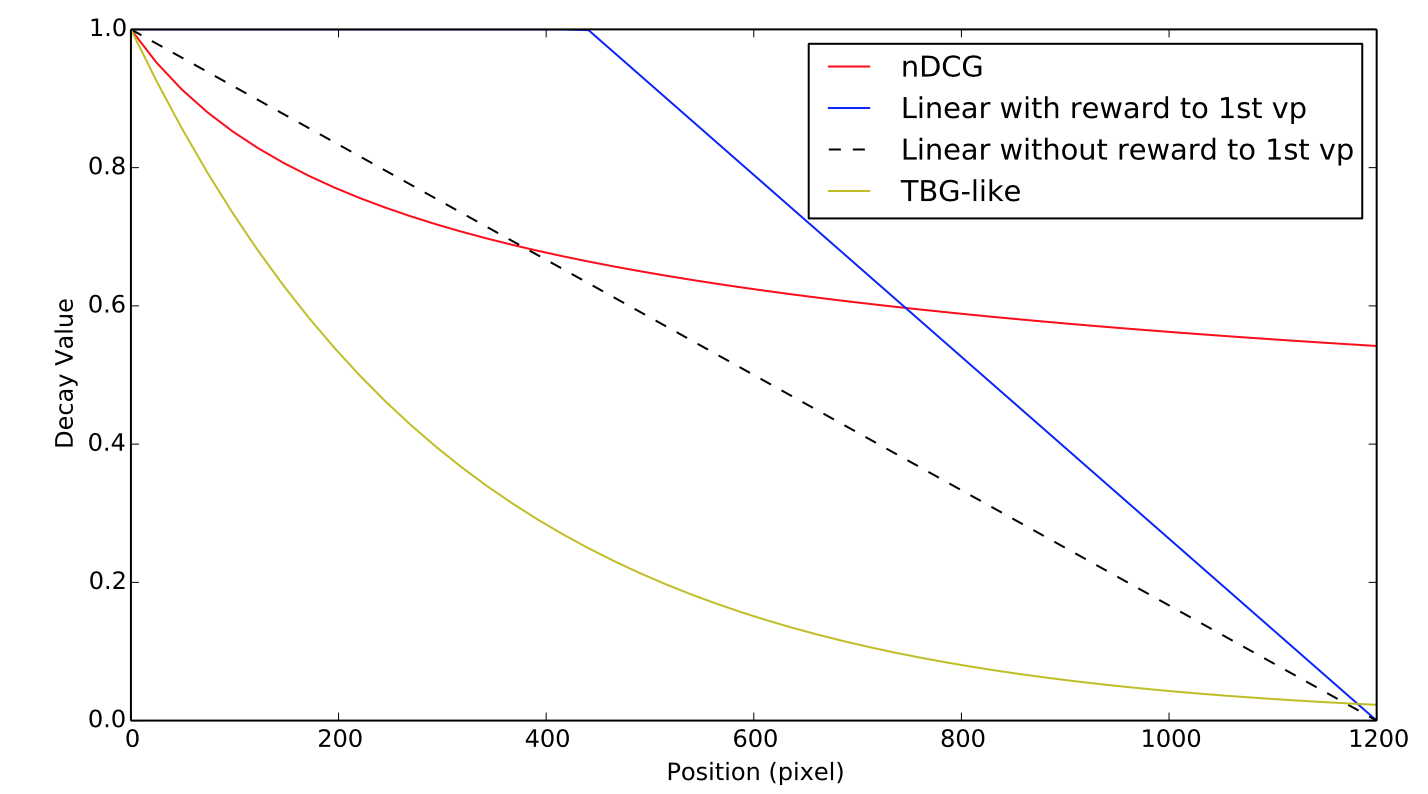
|  |  |  |  |
| --- | --- | --- | --- |
| **Metrics** | **Accuracy** | **Metrics** | **Accuracy** |
| nERR@3 | 0.720 | Q@5 | 0.660 |
| Q@3 | 0.707 | nDCG@5 | 0.647 |
| MSnDCG@3 | 0.693 | O-measure | 0.640 |
| ERR | 0.680 | NCUgu,P | 0.627 |
| P-measure | 0.680 | NCUrb,P | 0.480 |
| P-plus | 0.680 | AP@5 | 0.480 |
| nERR@5 | 0.680 | AP | 0.480 |
| NCUgu,BR | 0.673 | AP@3 | 0.473 |
| MSnDCG@5 | 0.673 | P@5 | 0.467 |
| NCUrb,BR | 0.673 | P@3 | 0.467 |
| nDCG@3 | 0.673 | RR | 0.380 |
| RBP | 0.667 | Hit@3 | 0.320 |
| Q-measure | 0.660 | Hit@5 | 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。

目前正在进行更细粒度的评价指标实验，已经抽取出200个SERP的源码，正在利用PhantomJS抽取每个结果的位置坐标，有一部分需要根据搜狗的标注数据进行手工对齐。

第二部分实验：

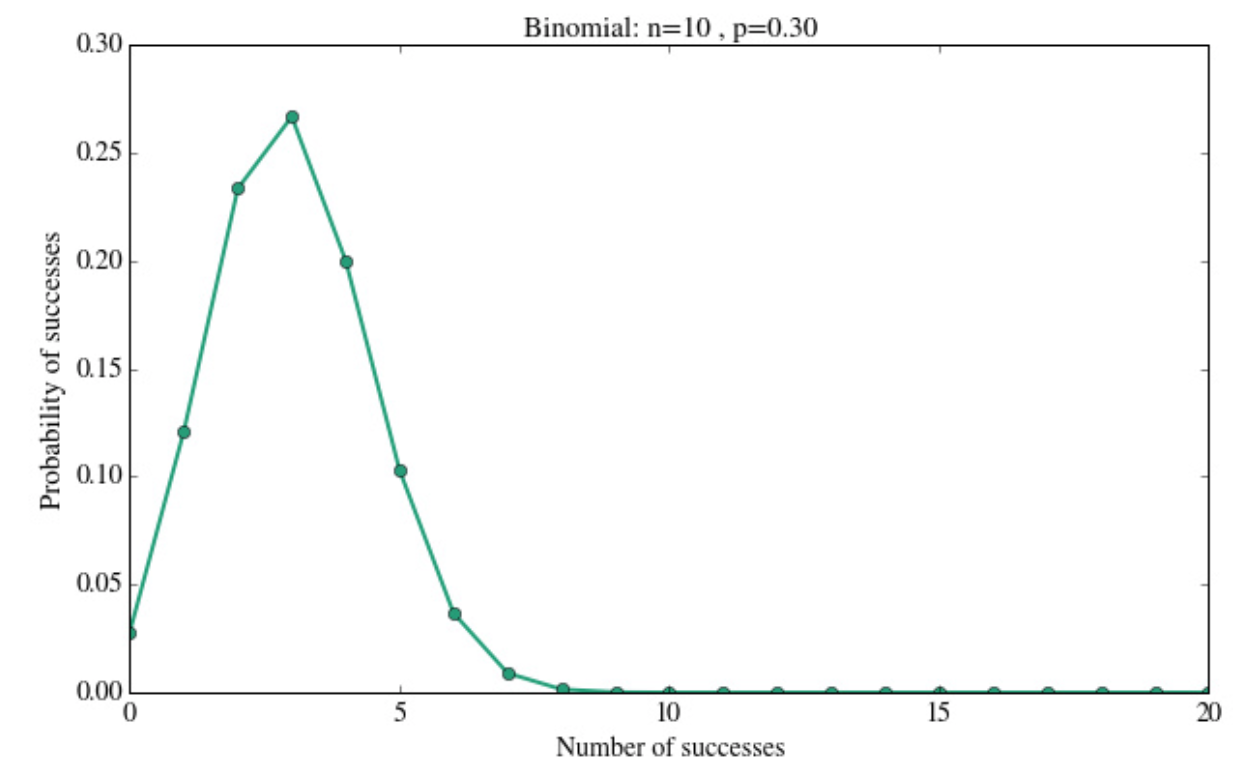
1. 对搜索引擎结果页上的结果进行了分析，抽取了每一个结果的位置（left, right, top, bottom），并且和搜狗同事提供的标注是一致的。
2. 忽略了所有位于第一个被标注结果之前的广告内容。
3. 设计了一些评价指标，重点考虑这么几个因素
4. 评价指标是否需要Normalized，考虑了一下应该是需要归一化的，因为这样使得不同长度的结果列表是可比较的。
5. 考虑decay function， 目前采用了如下几种Decay Function



* nDCG形式的1/log\_b\_r
* Linear with reward to first Viewport
* Linear without reward to first Viewport
* TBG式的exp(-t\*ln2/h) h是参数（TODO: 这个参数可调）

1. Gain的形式：采用了原始的分数（0~5）和指数（2^r）
2. Gain的累积方式：采用了阶跃函数（Gain均匀分布在前20%）

和Binomial的分布



具体计算得到Metrics 之间的一致性（kendall‘s tau），请见附件

20161022metric\_inner\_consistent.xlsx

表中紫色部分是新的metric和传统metric的一致性，还是比较低的，当然这个问题并没有一个ground truth。

第二部分结果是metric和SBS的一致性，请见附件20161022\_metric\_sbs\_agreerate.xlsx，可以看到最好的metric可以和传统指标表现接近，但是稍微差一点。

现在这些指标的问题是，没有太考虑结果本身的cost信息，如果一个结果排在第1位， 占据很大位置，但是得分很低，事实上惩罚了这个排在第1位以后的结果（对应的decay function的值变小了）。对这个结果本身没有太大的影响。

接下来会再调一下相关的参数，并且重点考虑Gain的问题。

/\*

2016年10月22日刘老师的comments

1. 需要进行一下细致的分析，看你的指标与排名靠前的指标在错误样例上是否一致，还是有互补效果

2. 需要考虑一下，移动环境下的评价，缺陷的本质在哪里，似乎是在于对于结果 cost 和 benefit 的计算方式都是基于同质性的假设，认为结果一样高，阅读每个结果的成本是一样的；同时结果只要看到就能得到他的全部 benefit；因此你的改进也要针对这两点来进行。

3. 移动环境下多级相关性评价似乎也面临着挑战，结果之间的 redundancy 变得非常重要，部分 card 类结果一个就能满足用户100%的需求，从而使得后面的评价变得不再必要。这种情况需要能够被你的计算方式所考虑到。

\*/

第三部分实验

1. 从查询的方面考虑，可以把分为这么几类：
   1. 单个结果有很大几率不能满足用户查询需求，比如knowledge型，比如“宝塔镇河妖的下一句是什么“，” 长江水文网“，” 笃字的拼音“
   2. 需要多个结果对比才能满足用户的查询需求；“关于秋天的作文500字”
   3. 从这个维度不能确定；
2. 从结果的考虑，可以分为这么几类：
   1. 结果是否不需要点击就可以获得信息，比如“笃字的拼音”，“济南天气预报”的结果，
   2. 结果必须通过点击获得相关信息，对于大部分结果是这样的。
   3. 不能确定

针对刘老师上周的点评，

1. 针对移动搜索环境下Cost和Gain的同质性假设，从移动搜索的结果特性来看，用户的点击的成本更高，因为用户可能要用额外的流量和等待时间。针对Cost，应该考虑用户为了得到相关的Gain付出的必要成本

其中p表示需要检验landing page的成本。

1. 针对Gain，一种做法是针对Query的不同，采用不同的评价指标（主要是decay function）， 近似于在P+Q中的，在Query是informational的时候和navigational的时候分别用Q和P+。

另外一种可能的做法是定义结果的可终结性（terminability）。如果一个结果是可终结的，意味着这个结果单独就可以满足用户需求。这个现象应该是由query和结果共同决定的。用户看到这个结果之后的其他结果的概率会显著减小。

这周准备面试花费的时间比较多，下一步进行查询结果的标注，计算不同的指标。

WWW方面，郑玉昆同学本周解决了链接关系抽取的一些技术问题，下周出结果过计划抽取子集合，可以发布Category B数据结合。