

信息检索与数据挖掘实验报告

基本信息:

姓名: 逄沐一

班级: 2018级数据班

学号: 201800130020

实验内容: IR Evaluation

实验任务:

- 实现MAP: mean average precision
- 实现MRR: mean reciprocal rank
- 实现NDCG: normalized discounted cumulative gain
- 在实验提供的数据集上测试三种方法

实验数据:(这里提供在实验框架下处理后的数据格式)

qrels_dict:

```
'224': {'301224062649237505': 1, '299219806081650691': 1, '317331309884239874': 2,
'317298074215120897': 2, '309744560542715904': 2, '309733609185619968': 2, '313863522255790081': 2,
'313805653422505984': 1, '318463763571433472': 2, '306464103105441792': 2}, '225':
{'306443022537719808': 2, '299316199571980288': 2, '299316199580368896': 2, '311527001297137664': 2,
'308666163024498689': 2, '299254723641495552': 2, '306463905960579072': 2, '306807696261271552': 2,
'306776742306123776': 2, '306579647796240384': 2, '306461020266844161': 2, '29921809899934977': 2,
'308707242054660097': 2, '308651755602866177': 1, '306459132851023872': 2, '299236121920016385': 2,
'306477050900930560': 2, '308805892055388160': 2, '308641248867135488': 2, '308654490263879682': 2,
'308654528021032960': 2, '299238298776379393': 2, '308671338808242177': 2, '299306858865373184': 1,
'308659586360020992': 2, '308694793335209984': 1, '297674112782962688': 1, '309178790842626048': 1,
'308862632629260288': 2, '299320972706734080': 2, '308961345536151552': 2, '299223337677250560': 1,
'301458306143240192': 2}}
```

test_dict:

```
'307029906305478657', '309312278774312961'], '224': ['311641149284700160', '311679019651325952',
'314214568694013952', '314330591527047170', '301224062649237505', '299219806081650691',
'317331309884239874', '317298074215120897', '309744560542715904', '309733609185619968',
'313863522255790081', '306043221463683073', '313805653422505984', '318463763571433472',
'306464103105441792', '304749857497829376', '315113953300738050', '303630708113547264'], '225':
['306443022537719808', '299316199571980288', '299316199580368896', '311527001297137664',
'308666163024498689', '299254723641495552', '306463905960579072', '306807696261271552',
'306776742306123776', '306579647796240384', '306461020266844161', '299218098999934977',
'308707242054660097', '308651755602866177', '306459132851023872', '299236121920016385',
'306477050900930560', '308805892055388160', '308641248867135488', '308654490263879682',
'308654528021032960', '299238298776379393', '308671338808242177', '299306858865373184',
'308659586360020992', '304829897384288256', '308694793335209984', '297674112782962688',
'309178790842626048', '308862632629260288', '299320972706734080', '308961345536151552',
'299223337677250560', '301458306143240192']}]
```

实验分析:

首先对所给数据进行分析：对于qrels_data来说它是一个字典字典的每一项为一次索引以及与他们对应的文档的相关度；对于test_dict同样是一个字典每一项为一次索引对应的查询出的文档id（已经按照打分排序了）。我们之后就需要根据这两个数据 对这组推荐情况进行一个打分，判断推荐算法的性能。

MAP：平均精度分析，这种评分方式通过对每一次的推荐的情况打一个average precision (AP)分数，然后对所有的AP进行一次平均，并以这个值作为对推荐算法性能的评测标准。接下来我们主要探讨AP的计算方式，平均精度只关注我们推荐出来的前k个文档，并且对于不相关的文档我们并不需要知道它的精度，我们认为一组推荐应该会将与查询相关的文档推荐到前面，并且我们认为推荐的前i个 ($i \leq k$) 个文档里相关文档越多那么这组推荐就越

好，因此设计出了AP这种评分标准他的公式为： $AP(j) = \frac{\sum_{i=1}^m \frac{R(i)}{i}}{m}$ R(i)为当前 推荐的前i ($i \leq k$) 个推荐中相关文档的数量，m为当前推荐前k个推荐中相关文档的总数量。i的范围为1到 $\min(k, \text{推荐的文档数量})$ ，之后我们在对全部的查询的AP做一个平均就可以得到最后的算法的MAP分数。

MRR：平均倒数排名，这种评分方式仅仅关注推荐中出现的第一个相关文档的位置（前k个不出现则为0），同样对所有的文档做一个平均，他的公式为 $RR(j) = \frac{1}{i}$ i表示第j次查询中第一次出现相关文档的位置，之后对所有查询的RR做一次平均即可得到MRR评分。

NDCG：标准折现累计收益，这是一种累计的评分，首先我们对CG进行定义，我们认为文档的相关度越高推荐越可靠，因此我们认为前k个文档的累计相关度越高则越可靠 故 $CG(j) = \sum rel(i)$ ，rel(i)表示第j次查询的第i个推荐文档的相关度，但仅考虑CG我们会忽略相关文档顺序对推荐的影响因此我们定义DCG如下：

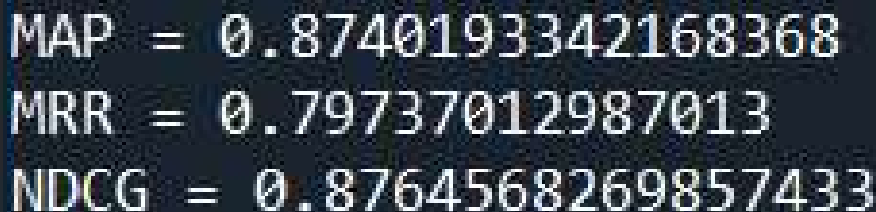
$CG(j) = rel(1) + \sum \frac{rel(i)}{\log 2(i)}$ 通过一个log函数 使得推荐在前面的文档的权重更高。

除此之外NDCG更具整体的查询做了归一化处理，他通过对当前推荐出的所有文档可以产生的最佳的DCG(IDCg)的计算来对原DCG做归一化处理，我们可以这样理解，同样的一组文档如果算法推荐出的顺序的DCG与最佳的顺序的DCG越接近那么我们认为这个推荐算法更优，IDCG的计算方式与DCG相同，只不过在计算前我们需要根据文档的相关度进行排序（仍为前k个的）。当计算出IDCG后我们可以根据以下公式计算NDCG, $NDCG(j) = \frac{DCG(j)}{IDCG(j)}$ 。

之后在对所有的NDCG做平均就可以得到这个算法的NDCG了。

实验结果与分析:

实验结果如图



```
MAP = 0.8740193342168368
MRR = 0.79737012987013
NDCG = 0.8764568269857433
```

我们发现实验提供数据的MRR分数(0.79)最低, MAP较高(0.874),NDCG(0.876)最高, 据此我们可以分析出这组推荐在整体相关度的考量上是比较好的, 但可能对于前几个推荐的关联度可能不高因此MRR会是最低的, 中间部分相关度可能也比较低因此MAP不是最高的, 相关度高的集中到了前1/4, 因此得出的NDCG会比较高, 总的来说这组推荐数据还是不错的, 但对于主题推荐来说这组推荐就相对来说差一点的, 因为主题推荐会更关注前几个推荐的准确度。而且实际上我们对于NDCG的评分是不应该有太大的参考度的因为实际查询中我们并不能知道文档的具体相关度。

收获与反思:

这次的实验总体来说比较简单, 虽然过程中也出现了对于某些算法流程的细节的不清晰导致结果非常反常的现象, 但通过与同学的交流以及查阅资料也顺利的完成了, 但以后学习过程中应该注意对于细节出的关注, 不然可能就会掌握不到算法的精髓。

除此之外通过实验可以发现这些评价指标的一些问题比如: NDCG对于一些坏文档不能做出很好的评价(NDCG更关注相关度而不关注文档的好坏), 而且NDCG更容易出现相同的度量值(如果对相关度的划分不是很细致); MRR对于推荐的整体水平没有什么体现仅仅关注最开始的, MAP更多的关注整体的情况, 因此可能会导致更关注前几条推荐的用户的体验较差。