IR评估指标—MAP、NDCG和MRR

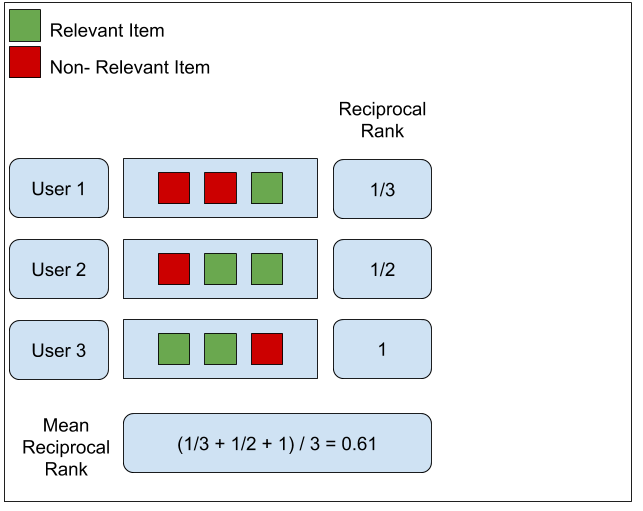
推荐系统需要把相关的物品放在推荐列表的最前面，用户不会通过滚动浏览200个条目来找到他们最喜欢物品。本文总结信息检索和推荐系统中常用的三种评估指标：MAP、NDCG和MRR。

# 平均排名的倒数（MRR）

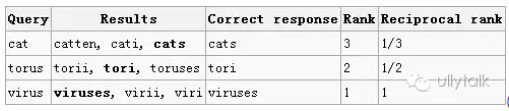
## 原理

MRR（Mean Reciprocal Rank）是最简单的评估指标，它的中文名称是平均排名的倒数。它试图度量“第一个相关的物品在哪里？”，把标准答案在被评价系统给出结果中的排序取倒数作为它的准确度，再对所有的问题取平均。

假设有以下三个针对三个用户的推荐列表。可以通过查找每个列表中第一个相关物品的排名来计算每个用户的倒数。然后我们对所有用户做一个简单的平均。



再比如：有3个query如下图所示：



(黑体为返回结果中最匹配的一项)

可计算这个系统的MRR值为：(1/3 + 1/2 + 1)/3 = 11/18=0.61。

## MRR的优点

* 该方法计算简单，解释简单。
* 这种方法高度关注列表的第一个相关元素。它最适合有针对性的搜索，比如用户询问“对我来说最好的东西”。
* 适用于已知项目搜索，如导航查询或寻找事实。

## MRR的缺点

* MRR指标不评估推荐项目列表的其余部分。它只关注列表中的第一个项目。
* 它给出一个只有一个相关物品的列表。如果这是评估的目标，那找个度量指标是可以的。
* 对于想要浏览相关物品列表的用户来说，这可能不是一个好的评估指标。用户的目标可能是比较多个相关物品。

# 平均精度均值（MAP）

## 原理

单个主题的平均准确率是每篇相关文档检索出后的准确率的平均值。主集合的平均准确率(MAP)是每个主题的平均准确率的平均值。MAP （Mean Average Precision）是反映系统在全部相关文档上性能的单值指标。系统检索出来的相关文档越靠前(rank 越高)，MAP就可能越高。如果系统没有返回相关文档，则准确率默认为0。

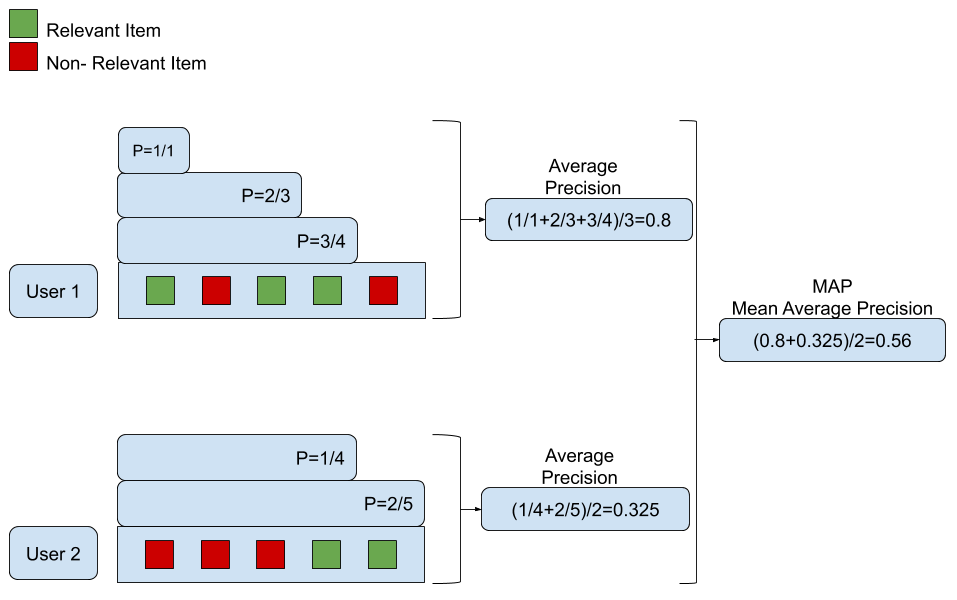
例如：假设有两个主题，主题1有4个相关网页，主题2有5个相关网页。某系统对于主题1检索出4个相关网页，其rank分别为1, 2, 4, 7；对于主题2检索出3个相关网页，其rank分别为1,3,5。对于主题1，平均准确率为(1/1+2/2+3/4+4/7)/4=0.83。对于主题2，平均准确率为(1/1+2/3+3/5+0+0)/5=0.45。则MAP= (0.83+0.45)/2=0.64。”

=========================================

假设我们有一个二进制相关性数据集。我们想要评估整个推荐项目列表，直到一个特定的截止值n。这个截止值之前使用Precision@N度量。决策支持度指标计算n个推荐中好的推荐的比例。此指标的缺点是，它不认为推荐列表是一个有序列表。P@N将整个列表视为一组条目，并平等对待推荐列表中的所有错误。

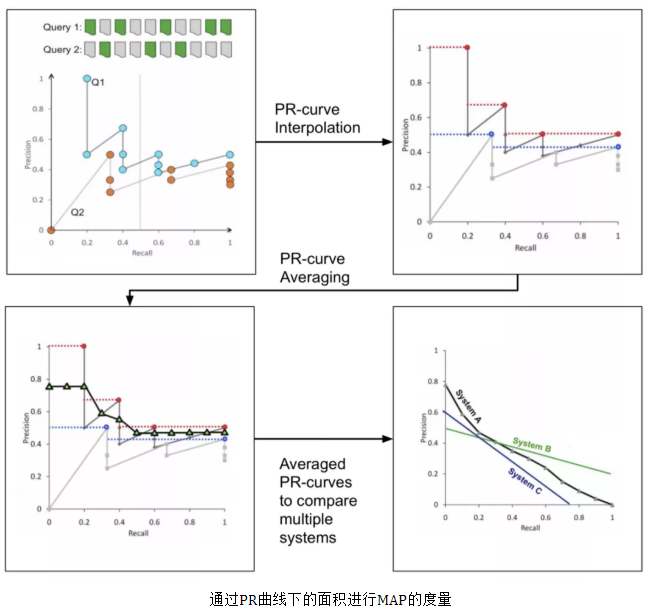
目标是在列表的前几个元素中减少错误，而不是在列表的后几个元素中。为此，我们需要一个度量来相应地对误差进行加权。这样做的目的是要在列表的顶部对错误的权重加大。然后，当我们沿着列表中较低的项目往下走时，逐渐减少错误的重要性。

平均预测(AP)度量试图近似这个加权滑动指标。它结合使用连续子列表上的精度，以及这些子列表中召回率的变化。计算如下：



从上图中，我们可以看到平均精度度量是在单个推荐列表，即用户级别。通过此项计算精度意味着细分推荐列表。当我们得到一个相关的项目时，我们就检查一个新的子列表。然后计算当前子列表的精度。我们对每个子列表都这样做，直到我们到达推荐的末尾。现在我们有了一组精度，我们对它们进行平均以获得单个用户的平均精度。然后我们得到所有用户的AP和平均精度。

这主要是AP度量的原始目标的近似值。AP度量表示精确度召回率率曲线下的面积。通过计算召回率作为召回值的函数，得到了精确度-召回率曲线。整个过程就是为每个用户推荐列表生成PR曲线。然后生成插值后的PR曲线，并对插值后的PR曲线求平均。这是视觉上的过程：



。。。。。

## MAP优点

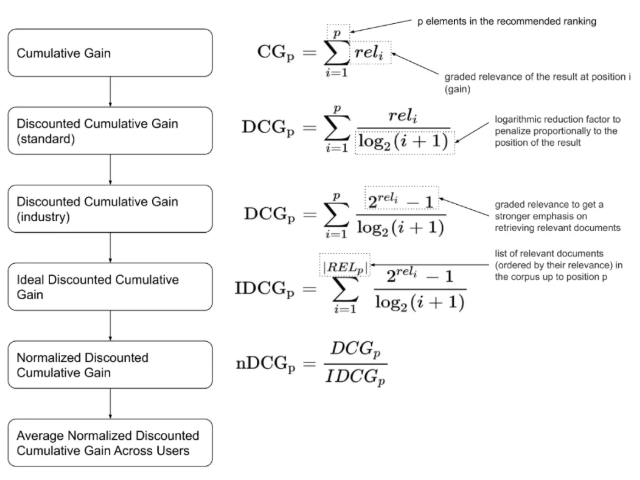
* 给出了一个代表精确度—召回率曲线下复杂区域的单一度量。这提供了每个列表的平均精度。
* 处理列表推荐物品的自然排序。这与将检索项视为集合的度量标准形成了对比。
* 这一指标能够给予发生在排序高的推荐名单中的错误更多的权重。相反，它对发生在推荐列表中较深位置的错误的权重较小。这符合在推荐列表的最前面显示尽可能多的相关条目的需要。

## MAP缺点

* 这个度量标准适用于二进制(相关/非相关)评级。然而，它不适合细粒度的数字评级。此度量无法从此信息中提取误差度量。
* 对于细粒度的评分，例如从1星到5星的评分，评估首先需要对评分进行阈值，以产生二元相关性。一种选择是只考虑大于4的评级。由于人工阈值的存在，这在评估度量中引入了偏差。此外，我们正在丢弃那些精细的信息。这个信息是在4星和5星之间的差异评级，以及在不相关的项目的信息。1星评级真的和3星评级一样吗？

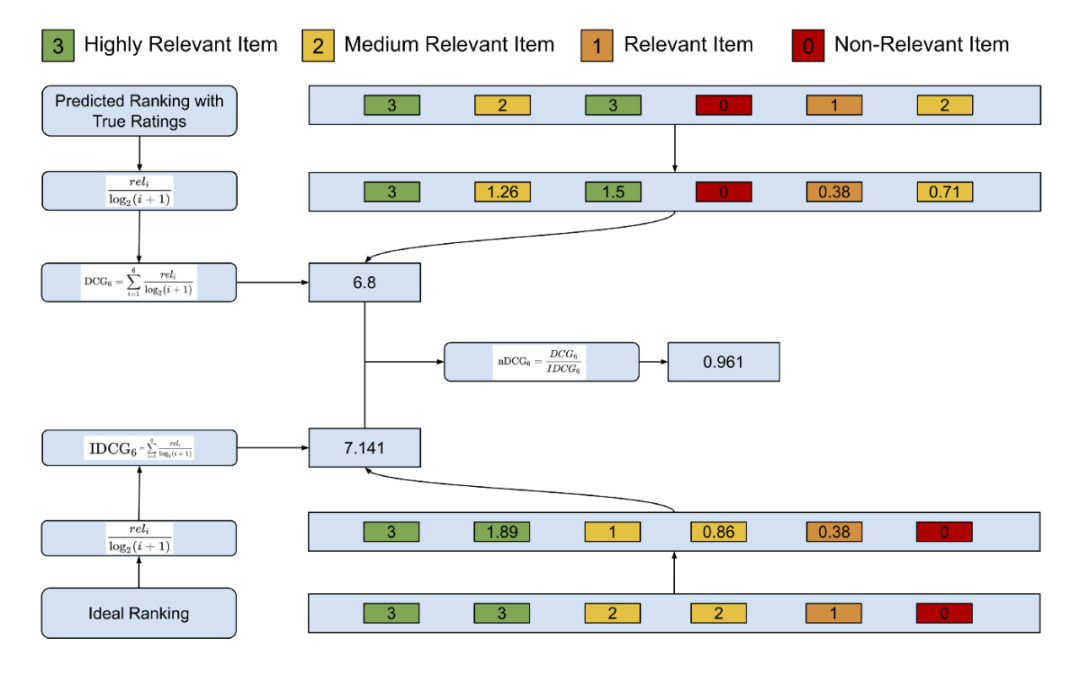
# 标准化折扣累积收益（NDCG）

MAP度量的目标与NDCG度量的目标相似。它们都重视将高度相关的文件排在推荐列表的前列。然而，NDCG进一步调整了推荐列表评估。它能够利用某些文档比其他文档“更”相关这一事实。高度相关的项目应该在中度相关项目之前，中度相关项目应该在非相关项目之前。我提供以下图表，它显示了阶段计算NDCG的步骤：

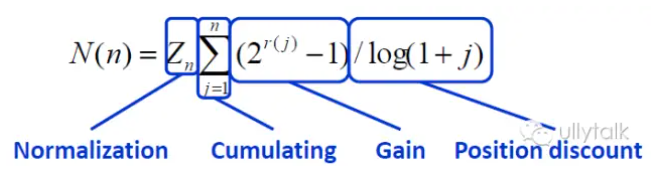


在NDCG之前我们有cumulative gain CG。这是一种基本的方法来积累等级相关度。这个度量不考虑元素在排序列表中的位置。对于排序任务，我们需要增加排序列表中元素位置的相对影响。standard Discounted Cumulative Gain(DCG)增加了一个对数衰减因子，以按比例惩罚项目的位置相关分数。此外，在工业应用中，为了强调检索相关文档，相关性分数得到提升是很常见的。这出现在industry DCG公式中。

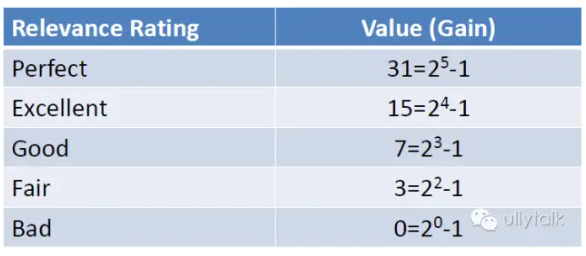
我们在处理动态系统。用户将得到数量可变的相关项目推荐。这使得DCG测量在用户之间没有可比性。我们需要标准化度量，使它在0和1之间。为此，我们确定用户的理想排名。然后用该排序作为Ideal Discounted Cumulative Gain IDCG。这提供了一个很好的归一化因子。它有助于计算 Normalized Discounted Cumulative Gain。因为这是一个针对每个用户的度量，所以我们需要为测试集中的所有用户计算这个度量。然后，这个平均值用于比较recsys系统之间的差异。为了可视化这个过程，我们在下面的图中计算单个用户的预测和理想排名。



NDCG（Normalized Discounted Cumulative Gain）计算相对复杂。对于排在结位置n处的NDCG的计算公式如下图所示：

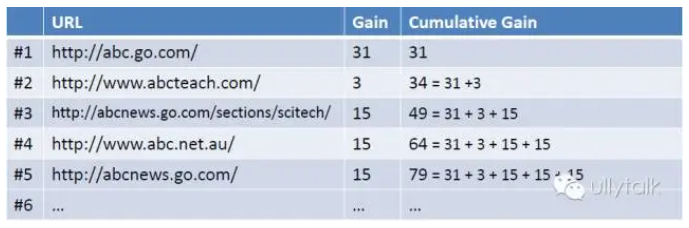


在MAP中，四个文档和query要么相关，要么不相关，也就是相关度非0即1。NDCG中改进了下，相关度分成从0到r的r+1的等级(r可设定)。当取r=5时，等级设定如下图所示：

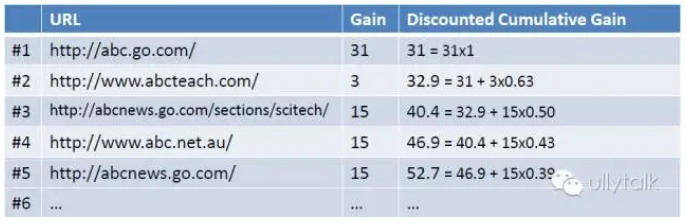


(应该还有r=1那一级，原文档有误，不过这里不影响理解)

例如现在有一个query={abc}，返回下图左列的Ranked List(URL)，当假设用户的选择与排序结果无关(即每一级都等概率被选中)，则生成的累计增益值如下图最右列所示：



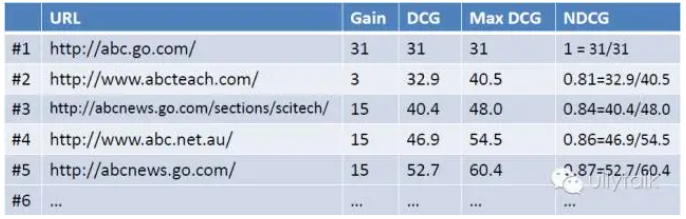
考虑到一般情况下用户会优先点选排在前面的搜索结果，所以应该引入一个折算因子(discounting factor): log(2)/log(1+rank)。这时将获得DCG值(Discounted Cumulative Gain)如下如所示：



最后，为了使不同等级上的搜索结果的得分值容易比较，需要将DCG值归一化的到NDCG值。操作如下图所示，首先计算理想返回结果List的DCG值：



然后用DCG/MaxDCG就得到NDCG值，如下图所示：



## NDCG优点

* NDCG的主要优势是它考虑到了分等级的相关性值。当它们在数据集中可用时，NDCG是一个很好的选择。
* 与MAP度量相比，它在评估排名项目的位置方面做得很好。它适用于二元的相关/非相关场景。
* 平滑的对数折现因子有一个很好的理论基础，该工作的作者表明，对于每一对显著不同的排名推荐系统，NDCG度量始终能够确定更好的一个。

## NDCG缺点

* NDCG在部分反馈方面有一些问题。当我们有不完整的评级时，就会发生这种情况。这是大多数推荐系统的情况。如果我们有完整的评级，就没有真正的任务去实现！在这种情况下，recsys系统所有者需要决定如何归罪于缺失的评级。将缺少的值设置为0将把它们标记为不相关的项。其他计算值(如用户的平均/中值)也可以帮助解决这个缺点。
* 接下来，用户需要手动处理IDCG等于0的情况。当用户没有相关文档时，就会发生这种情况。这里的一个策略是也将NDCG设置为0。
* 另一个问题是处理NDCG@K。recsys系统返回的排序列表的大小可以小于k。为了处理这个问题，我们可以考虑固定大小的结果集，并用最小分数填充较小的集合。
* 正如我所说的，NDCG的主要优势在于它考虑到了分级的相关性值。如果你的数据集有正确的形式，并且你正在处理分级相关性，那么NDCG度量就是你的首选指标。

参考

<https://mp.weixin.qq.com/s/0OTxBxriIGUXrJAe-rUyTQ>

<https://mp.weixin.qq.com/s/LyU4hrj67X3BiO5NOkojgQ>