028 | 搜索系统评测，有哪些高级指标?

2017-12-06 洪亮劼

AI技术内参

进入课程 >

讲述：初明明
时长 07:16　大小 3.33M

周一我们介绍了基于“二元相关”原理的线下评测指标。可以说，从 1950 年开始，这种方法就主导了文档检索系统的研发工作。然而，“二元相关”原理从根本上不支持排序的评测，这就成了开发更加准确排序算法的一道障碍。于是，研究人员就开发出了基于“多程度相关”原理的评测标准。今天我就重点来介绍一下这方面的内容。

基于多程度相关原理的评测

从“二元相关”出发，自然就是给相关度更加灵活的定义。在一篇发表于 NIPS 2007 的文章中（参考文献 [1]），雅虎的科学家介绍了雅虎基于五分标准的相关评价体系，从最相关到最不相关。而在同一年的 SIGIR 上，谷歌的科学家也发表了一篇文章（参考文献 [2]），介绍了他们的“多程度”相关打分机制。至此之后，基于“多程度相关”原理的评价标准慢慢被各种搜索系统的研发者们所接受。

在这样的趋势下，基于“二元相关”的“精度”（Precision）和“召回”（Recall）都变得不适用了。我们需要新的、基于“多程度相关”的评价指标。

芬兰的科学家在 2000 年的 SIGIR 上（参考文献 [3]）发表了一种计算相关度评测的方法。这种方法被广泛应用到了“多程度相关”的场景中。那么，芬兰科学家发明的方法是怎样的呢?

这种方法被称作是“**折扣化的累积获得**”（Discounted Cumulative Gain），简称“DCG”。 在介绍 DCG 之前，我们首先假定，位置 1 是排位最高的位置，也就是顶端的文档，而位置数随着排位降低而增高，位置 10 就是第一页最后的文档。

DCG 的思想是这样的。

首先，一个排序的整体相关度，是这个排序的各个位置上的相关度的某种加权。这样用一个数字就描述了整个排序。只要排序的结果不同，这个数字就会有所不同。因此，这就避免了“精度”或“召回”对排序不敏感的问题。

其次，每个位置上面的“获得”（Gain）是和这个文档原本定义的相关度相关的，但是，根据不同的位置，要打不同的“折扣”。位置越低（也就是位置数越大），折扣越大。这就是 DCG 名字的由来。

在原始的 DCG 定义中，“折扣”是文档的相关度除以位置的对数转换。这样，既保证了位置越低（位置数大），折扣越大，还表达了，高位置（位置数小）的差别要大于低位置之间的差别。这是什么意思呢？意思就是，如果某一个文档从位置 1 挪到了位置 2，所受的折扣（或者说是损失）要大于从位置 9 挪到了位置 10。在这样的定义下，DCG 鼓励把相关的文档尽可能地排到序列的顶部。

事实上，假设我们有 5 个文档，假定他们的相关度分别是 1、2、3、4、5，分别代表“最不相关”、“不相关”、“中性”、“相关”和“最相关”。那么，在 DCG 的定义下，最佳的排序就应该是把这 5 个文档按照相关度的顺序，也就是 5、4、3、2、1 来排定。任何其他的顺序因为根据位置所定义的“折扣获得”的缘故，都会取得相对较小的 DCG，因此不是最优。DCG 比“精度”和“召回”能够更好地表达对排序的评估。

但直接使用 DCG 也存在一个问题。如果我们有两个查询关键字，返回的文档数不一样，那么直接比较这两个查询关键字的 DCG 值是不“公平”的。原因在于 DCG 的“加和”特

性，结果肯定是越加越大，因此不能直接比较两个不同查询关键字的 DCG 值。

有没有什么办法呢？把 DCG 加以“归一化”的指标叫做 **nDCG** （Normalized Discounted Cumulative Gain）。nDCG 的思路是下面这样的。

首先，对某一个查询关键字的排序，根据相关信息，来计算一组“理想排序”所对应的 DCG 值。理想排序往往就是按照相关信息从大到小排序。然后，再按照当前算法排序所产生的 DCG 值，除以理想的 DCG 值，就产生了“归一化”后的 DCG，也就是我们所说的 nDCG 值。简单来说，nDCG 就是把 DCG 相对于理想状态进行归一化。经过 nDCG 归一化以后，我们就可以比较不同查询关键字之间的数值了。

这里需要说明的是，我们上面介绍的是 DCG 的原始定义。后来微软的学者们在 2005 年左右发明了另外一个变种的 DCG，基本原理没有发生变化，只是分子分母有一些代数变形。这个新的版本后来在工业界得到了更加广泛的应用。如果你感兴趣，可以查看文末的参考文献 [4]。

直到今天，nDCG 以及 DCG 依然是评价排序算法以及各种排序结果的标准指标。

比较两个不同的排序

不管是我们之前谈到的“精度”和“召回”，还是今天介绍的 nDCG，我们都是使用一个“数”来描述了相对于某个查询关键字，一组结果的好坏。当我们有多个查询关键字的时候，我们该如何比较两个不同排序的结果呢？

这里面的一个问题是，相对于两个不同的排序 A 和 B 来说，可能结果各有千秋，也许对于某一个关键字 A 比 B 的表现要好，但是另外一个关键字 B 就比 A 的结果更棒。这怎么办呢？

也许你会想到用平均值来描述 A 和 B 的表现。这的确是很好的第一步。于是，我们就计算 A 和 B，两个排序的平均表现。这样对于这两个排序而言，我们就有了两个数值来表达这两个排序的好坏。

然而，很快我们就会遇到问题。假设 A 的 nDCG 平均值是 0.781，B 的 nDCG 平均值是 0.789，我们可以下结论认为 B 是比 A 更好的排序算法吗？

答案当然是不一定。这种情况，我们就需要依赖统计工具“**假设检验**”来评价两个排序的好坏。

我这里就不去复习假设检验的细节了，简单说一个经常使用的工具。

如果我们比较 A 和 B 是在同一组查询关键字上的话，那我们常常可以使用“**两个样本的配对 T 检验**”（Two Sample Paired T-Test）。这里所谓的“配对”是指 A 和 B 的结果是可以一一比较的。这里的“T 检验”其实就是说借助“T 分布”或者我们通常所说的“学生分布”来进行假设检验。如果我们是在不同查询关键字集合中进行比较的话，还有其他的假设检验工具，这里就不展开了。

值得注意的是，假设检验本身也不是“万灵药”。第一，怎么最有效地在排序结果上进行假设检验还是一个研究话题，包括我们刚说的“两个样本的配对 T 检验”在内的所有方法都不是“金科玉律”。第二，依靠假设检验得出来的结论，仅仅是统计意义上的“好坏”，和这些系统在用户面前的表现可能依然会有很大差距。因此，**对于假设检验的结果也要带有“批判”的眼光**。

小结

今天我为你讲了现代搜索技术中如何评价我们构建的系统，特别是如何评价排序系统。

一起来回顾下要点：第一，简要讲解了基于“多程度相关”的评价体系，包括其由来和 DCG 以及 nDCG 的概念。第二，详细介绍了如何来比较两个排序的好坏。

最后，给你留一个思考题，如果我们只有“二元”相关信息，能不能用 nDCG 来评价好坏呢？

欢迎你给我留言，和我一起讨论。

参考文献

1. Ben Carterette and Rosie Jones. Evaluating search engines by modeling the relationship between relevance and clicks. Proceedings of the 20th International Conference on Neural Information Processing Systems (NIPS'07), J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis (Eds.). Curran Associates Inc., USA, 217-224,2007.

2. Scott B. Huffman and Michael Hochster. How well does result relevance predict session satisfaction? Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '07). ACM, New York, NY, USA, 567-574, 2007.
3. Kalervo Järvelin and Jaana Kekäläinen. IR evaluation methods for retrieving highly relevant documents. Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '00). ACM, New York, NY, USA, 41-48, 2000.
4. Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. Proceedings of the 22nd international conference on Machine learning (ICML '05). ACM, New York, NY, USA, 89-96, 2005.

© 版权归极客邦科技所有，未经许可不得传播售卖。 页面已增加防盗追踪，如有侵权极客邦将依法追究其法律责任。

上一篇　027 | 搜索系统评测，有哪些基础指标?

下一篇　029 | 如何评测搜索系统的在线表现?

精选留言 (3)

写留言

侯永胜
2018-01-16

1

请大师也讲讲推荐系统相关的内容哈 谢谢啦

展开

白杨
2018-05-16

另外一直有个问题，烦请老师解答一下：
为什么这些高级评价指标都是不可微的呢？直观上体现在哪里？数学形式上又体现在哪里？

展开

白杨
2018-05-16

我的想法是，用文档出现的频率来代替位置的角色，然后来打折扣，这样应该在某些场景下可行。