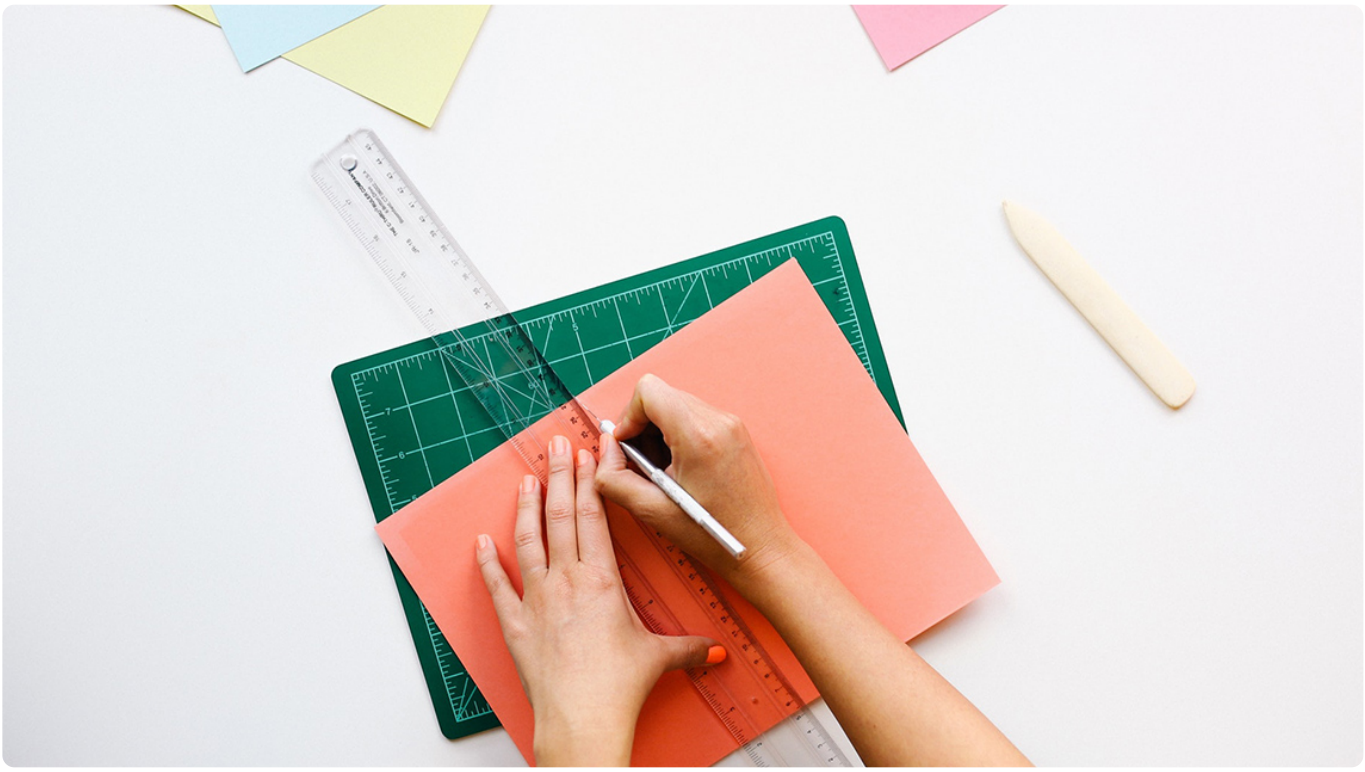
027 | 搜索系统评测，有哪些基础指标?

2017-12-04 洪亮劼

AI技术内参 进入课程 >



讲述：初明明
时长 08:45 大小 4.01M

我在之前几周的专栏文章里主要讲解了最经典的信息检索（Information Retrieval）技术和基于机器学习的排序学习算法（Learning to Rank），以及如何对查询关键字（Query）进行理解，包括查询关键字分类、查询关键字解析以及查询关键字扩展。这些经典的技术是 2000 年后开始流行的各类搜索引擎的核心技术。

在进一步介绍更多的搜索引擎技术前，我觉得有必要专门抽出一周时间，来好好地看一下搜索系统的评测（Evaluation）以及我们经常使用的各类指标（Metric）。俗话说得好，“如果你不能衡量它，你就不能改进它”（If You Can't Measure It, You Can't Improve It）。意思其实就是说，对待一个系统，如果我们无法去衡量这个系统的好坏，没有相应的评测指标，那就很难真正地去琢磨怎么改进这些指标，从而达到提升系统的目的。

虽然我们这里是在搜索系统这个重要场景中讨论评测和指标，但实际上我们这周要讨论的很多细节都可以应用到很多类似的场景。比如，我们后面要讨论的推荐系统、广告系统等，在这些场景中几乎就可以无缝地使用这周要讲的很多内容。

线下评测

假设你今天开发了一个新软件，比如说是一个最新的手机软件，你怎么知道你的用户是不是喜欢你的软件呢？你怎么知道你的用户是不是愿意为你的软件掏钱呢？

评测的核心其实就是了解用户的喜好。最直接的方法，当然是直接询问用户来获得反馈。例如你可以对每一个下载了你手机软件的用户强行进行问卷调查，询问他们对待新软件的态度。

然而，我们很快就会发现这样的方法是行不通的。姑且不说用户是否会因为这样强行的方式产生反感，我们是不是能通过这些调查问卷获得用户的真实反馈，这本身就是一个问题。这里面涉及到调查问卷设计的科学性问题。

即便这些调查问卷都能完整准确地反映出用户对手机软件的看法，真正实施起来也会面临种种困难。如果这款手机软件的用户数量有百万甚至千万，那我们就要进行如此大规模的问卷调查，还要处理调查后的数据，显然这样做的工作量非常大。而这些调查问卷是没法反复使用的，因为下一个版本的软件更新后，用户的态度就会发生改变，这样的方式就没法系统地来帮助软件迭代。

那么如何才能形成一组数据来帮助系统反复迭代，并且还能够减少人工成本，这就成了一个核心问题。

在信息检索系统开发的早年，研究人员和工程师们就意识到了这个核心问题的重要性。英国人赛利尔·克莱温顿（Cyril Cleverdon）可以算是最早开发线下测试集的计算机科学家。

赛利尔生于 1914 年，在英国的布里斯托（Bristol）图书馆工作了很长时间。从 1950 年开始，赛利尔就致力于开发信息检索系统，以提高图书馆查询的效率。1953 年他尝试建立了一个小型的测试数据集，用于检测图书管理员查找文档的快慢。这个工作最早发表于 1955 年的一篇论文（参考文献 [1]）。

这之后，英美的一些早期信息检索系统的研发都开始顺应这个思路，那就是为了比较多个系统，首先构造一个线下的测试数据集，然后利用这个测试集对现有的系统反复进行改进和提

升。如果你想对早期测试集的构造以及信息有所了解，建议阅读文末的参考文献 [2]。

那么，当时构造的这些测试数据集有些什么特点呢？

这些测试数据集都会包含一个查询关键字集合。这个集合包含几十到几百不等的查询关键字。一方面，这些关键字的选取大多来自于经验。另一方面，从赛利尔就开始认识到，需要保证有一些信息一定能够通过这些关键字来找到。其实，这里就是在测试我们后面要讲的“召回”。

在有了这些查询关键字以后，**这些测试数据集往往有几百到几千不等的文档。**这些文档中的某一部分，研究人员在构造数据集的时候就知道了会包含所对应查询关键字需要的信息，也就是我们后面要说的相关文档。

你可以看到，几十到几百的查询关键字以及几千个文档，很明显不能代表所有可能使用系统的用户的行为。你甚至可以说，这都无法代表绝大多数用户的行为。然而，这种测试集的好处是，查询关键字和文档本身是和要测试的系统无关的。也就是说，今天我们要测试系统 A，还是明天要测试系统 B，都可以反复利用同样一组测试数据集。这样做的好处相比于我们之前提到的问卷调查是显而易见的。

另外，我需要强调的是，“用户”这个概念在测试数据集中被“抽象”出去了。当我们在讨论文档相对于某个查询关键字的相关度时，我们假定这种相关度是恒定的，是对于所有用户都适用的。因此，究竟是哪位用户在使用这个系统并不重要。只要研发的系统能够在这些“标准化”的查询关键字和文档的集合表现优异，我们就相信这个系统能够满足所有用户的需要。

因为测试数据集并不是用户与产品交互产生的真实回馈结果，所以我们往往又把测试数据集叫作“线下评测数据”。

基于二元相关度的评测指标

从线下收集评测数据以后，我们最容易做到的就是利用“二元相关度”所定义的一系列评测指标来衡量手中系统的好坏。

什么叫“二元相关度”呢？简单来说，就是指**针对某一个查询关键字而言，整个测试集里的每一个文档都有一个要么“相关”要么“不相关”的标签。**在这样的情况下，不存在百分比的相关度。而每个文档针对不同的关键字，有不同的相关信息。

假定某个系统针对某个关键字，从测试数据集中提取一定量的文档而不是返回所有文档，我们就可以根据这个提取的文档子集来定义一系列的指标。

有两个定义在“二元相关度”上的指标就成了很多其他重要指标的基石。一个叫**“精度”**（Precision），也就是说，在提取了的文档中，究竟有多少是相关的。另一个叫“**召回**”（Recall），也就是说， 在所有相关的文档中，有多少是提取出来了的。

“精度”和“召回”的相同点在于，分子都是“即被提取出来了又相关的文档数目”。这两个指标所不同的则是他们的分母。“精度”的分母是所有提取了的文档数目，而“召回”的分母则是所有相关的文档数目。如果我们返回所有的文档，“精度”和“召回”都将成为1（也就是说，在这样的情况下是没有意义的）。因此，我们注意到，这两个指标其实都假定，提取的文档数目相比于全集而言是相对比较小的子集。

很快，大家从实践中就体会到，“精度”和“召回”就像是“鱼与熊掌不可兼得”。一个系统很难做到“精度”和“召回”都能够达到很高的数值。也就是说，我们往往需要在这两个指标之间做一些平衡。于是，研究人员开始寻找用一个数字来表达“精度”和“召回”的“平均水平”。来自英国的学者范·李杰斯博格（C. J. van Rijsbergen）最早在论文中采用了 “**调和平均数**” （Harmonic Mean）来计算 “精度”和“召回”的平均（参考文献 [3]）。这个方法被后人称为“**F 值**” ，并且一直沿用至今。

这里对“精度”和“召回”还需要注意一点，因为这两个指标都是基于“二元相关度”的。因此，这两个指标都不是“排序指标”（Ranking Metrics）。换句话说，这两个指标其实并不能真正评价排序系统。

比如，我们针对某个关键字提取 10 个文档，如果有 3 个相关文档被取出来，不管是“精度”还是“召回”都无法分辨这三个文档在最后序列中的位置，是头三位，还是后面的三位？很遗憾，“精度”和“召回”都无法解决这个问题。“二元相关度”的这个问题也就指引研究人员去开发真正能对排序进行评估的指标。

小结

今天我为你讲了现代搜索技术中一个非常重要的一个环节，那就是如何评价我们构建的系统。我们详细讲解了线下测试的由来以及这样的测试相比于调查问卷的优势。

一起来回顾下要点：第一，简要介绍了可重复使用的线下测试集的历史，以及这样的测试集都有什么特点与局限。第二，详细介绍了两个非常流行和重要的基于“二元相关度”的评测

指标，那就是“精度”和“召回”。

最后，给你留一个思考题，我们讲了排序的好坏不能简单地从“精度”和“召回”的数值看出，那能不能动一些手脚呢？如果我们就依赖“二元相关度”，有没有什么方法来看“排序”的好坏呢？

欢迎你给我留言，和我一起讨论。

参考文献

1. R.G. THORNE, B.Sc., A.F.R.Ae.S. The Efficiency Of Subject Catalogues and The Cost of Information Searches. Journal of Documentation, Vol. 11 Issue: 3, pp.130-148, 1995.
2. K. SPARCK JONES, C.J. VAN RIJSBERGEN. Information Retrieval Test Collections. Journal of Documentation, Vol. 32 Issue: 1, pp.59-75, 1976.
3. C.J. VAN RIJSBERGEN. Foundation of Evaluation. Journal of Documentation, Vol. 30 Issue: 4, pp.365-373, 1974.

上一篇　026 | “查询关键字理解”三部曲之扩展

下一篇　028 | 搜索系统评测，有哪些高级指标?

精选留言 (6)

写留言

黑翼天佑
2017-12-15

2

可以通过选择不同的阈值来画出roc曲线，然后计算auc来评价

展开

作者回复: 这是一个比较好的方法。

xinfeng1i
2019-05-08

依次查看top1,top3,top5等不同提取数量的精度和召回，可以估计排序的好坏。

牧云
2019-03-16

通过用户对搜索结果的选择，选择与不选择，构建二分，来判断排序的好坏

展开

幻大米
2018-12-24

「如果我们返回所有的文档，“精度”和“召回”都将成为 1」精度不是 1 吧?

无
2018-03-03

您好！请问如果正负样本比差别很大，比如1/1000的情况下，上述这些指标以及AUC是否就不准导致无法作为参考了？如果这样，应该如何应对正负样本比悬殊时模型评估的问题呢？ 谢谢！

展开

老敖

2017-12-04

在计算召回或者精确率的时候，不是只计算文档个数，而是把它变成一个浮点数，把排序的相关性也考虑进去。就是相关性大的，查出来一篇顶两篇。是这样吗？

展开

作者回复: 计算AUC是一种思路。