第3章

信息检索的评价

3.1 信息检索的评价指标

信息检索的效果是指利用检索系统(或检索工具)开展信息检索服务时所产生的有效成果,它直接反映了信息检索系统的性能和信息检索服务的质量。对信息检索效果进行评价,找出影响检索效果的各种因素,可以为改善信息检索系统性能提供明确的参考依据,从而进一步满足用户的检索需求。

3.1.1 查全率

查全率(Recall Ratio)和查准率(Precision Ratio)是美国学者佩里(J. W. Perry)和肯特(A. Kent)在20世纪50年代最先提出的。查全率也称为检全率、召回率,查准率也称为检准率、精确率。作为信息检索效果评价的两个重要指标,不仅可以用来评价每次检索的全面性和准确性,也是在信息检索系统评价中衡量系统检索性能的重要方面。

在信息检索系统中,每进行一次检索,就把系统中所有的文献分为检出文献和未检出文献两个部分(如图3-1所示)。其中一部分是检出文献,指的是与检索策略相匹配并被检索出来的文献,用户根据自己的判断把它分成相关文献(a,合理的命中)和非相关文献(b,误检);另一部分是未检出文献,指的是未能与检索策略相匹配的文献,也可以把它分成相关文献(c,漏检)和非相关文献(d,合理的排除)。

可以看到,$a+b$ 表示检出的全部文献数量,相对整个系统(尤其在Internet环境下)规模来说是很小的;$c+d$ 表示未检出的文献数量,数量则非常大;$a+c$ 表示与检索相关的全部文献;$b+d$ 表示与检索不相关的全部文献;$a+b+c+d$ 则表示检索系统中的所有文献。

	相关文献	非相关文献
检出文献	a	b
未检出文献	c	d

图 3-1 检索中的系统文献

查全率 R 为从检索系统中检出的与检索策略相关文献数占系统中相关文献信息总数的百分比。即

R(查全率)=检出的相关文献数/系统中相关文献总数=$a/(a+c)\times 100\%$

查全率反映了信息检索的全面性。

例如,在一次检索中,共检出文献 100 篇,经过分析判定,其中与检索相关的文献为 80 篇,其余的 20 篇为误检文献,假如检索系统中还有 80 篇相关文献,由于各种原因而未被检出(漏检),那么按照上述公式,本次检索的查全率就等于 $80/(80+80)\times 100\%$,即 50%。

理论上来讲,利用上述公式,对每一次信息检索,都可计算出其查全率,对检索效率做出定量化的评价。但在实际量化的操作中却有着难以克服的困难,因为实际运行的检索系统中根本不可能浏览所有的文献信息,未被检出的相关文献信息数量和文献总量等都很难统计。

3.1.2 查准率

1. 查准率

查准率 P 为从检索系统中检出的相关文献数占检出文献信息总数的百分比。即

P(查准率)=检出的相关文献数/检出的文献总数=$a/(a+\mathrm{b})\times 100\%$

上式中,当进行检索时,与检索策略相匹配并被检索出来的文献,用户根据自己的判断把它分成相关文献(a,合理的命中)和非相关文献(b,误检)。查准率反映检索的准确性。

例如,在一次检索中,共检出文献 100 篇,经过分析判定,其中与检索相关的文献为 80 篇,其余的 20 篇为误检文献,那么按照上述公式,本次检索的查准率 P 就等于(80/100)×100%,即 80%。

2. 替代方法

除了信息检索的查全率和查准率以外,两位美国研究人员 H. Vernon Leighton 和 Jaideep Srivastava 提出了一种计算查准率的替代方法,即“相关性范畴”概念和“前 X 命中记录查准率”。下面对这两种方法进行简要的介绍。

1) 相关性范畴

相关性范畴是按照检索结果同用户需求的相关程度,把检索结果分别归入如下 4 个范畴。

(1) 范畴 0:重复链接,死链接和不相关链接。

(2) 范畴 1：技术上相关的链接。

(3) 范畴 2：潜在有用的链接。

(4) 范畴 3：十分有用的链接。

2) 前 X 命中记录查准率

一旦相关判断进行完毕，接下来的工作就是决定对检索工具的检索性能进行评价的具体计量指标。为了解决这个问题，Leighton 和 Srivastava 提出了前 X 命中记录查准率 $P(X)$，用来反映检索工具在前 X 个检索结果中向用户提供相关信息的能力。

这个解决办法的最大优点就是它的可操作性。评价实验者可以根据人力、物力上的实际情况来选择 X 的具体数值。理论上，X 越大，$P(X)$就越接近真实查准率，但这也意味着评价实验成本的增加。实验结果的精确程度和实验成本也是一种互相制约的关系。当然，在条件允许的情况下，X 应该尽可能大。

一种比较合理的情况是把 X 值定为 20，因为许多检索工具都会以 10 为单位输出检索结果，前 20 个检索结果就是检索结果的前两页。而检索用户对前两页的检索结果一般都会认真浏览。这样要计算的查准率就是 $P(20)$。在计算 $P(20)$时，要对处在不同位置的检索结果进行加权处理。因为检索工具都有某种排序算法，排在前面的检索结果在理论上应具有较大的相关系数，并且检索者一般都从头开始检验检索结果。因此，排在前面的检索结果应该被赋予高权值。

与真实查准率一样，$P(20)$也是一个比值，取值范围为 0～1。对 $P(20)$的计算，Leighton 和 Srivastava 的做法如下。

(1) 先根据对命中记录进行相关检验的结果，给每个检索结果赋予相关系数 0 或 1。判断为相关的检索结果赋值为 1，不相关的结果赋值为 0。在评价时，相关标准可以根据评价的需要来确定。例如，只要求满足基本的检索要求，范畴 1、2、3 都可以被认为是相关的结果。而要求最为满足检索要求时，就只有范畴 3 是相关的了。

(2) 把检索结果分为 3 组：1～3、4～10、11～20，然后在计算时分别赋予不同的权值。一般设第一组权值为 20，第二组权值为 17，第三组权值为 10。

(3) 计算 $P(20)$的分子，把每组的检索结果乘以各自的权值相加。

例如，某个检索工具对某个检索课题返回的检索结果中，第一组有 2 条相关记录，第二组有 5 条相关记录，第三组有 8 条相关记录，那么，它的 $P(20)$的分子就是：$2\times20+5\times17+8\times10=205$。

(4) 计算 $P(20)$的分母。如果返回的检索结果超过 20 条，那么分母就是所有的 20 条记录都相关时的权值之和，即 $3\times20+7\times17+10\times10=279$。如果返回的检索结果不超过 20 条，分母就需要进行一定的调整，以使计算结果更接近真实查准率。

在检索结果少于 20 时如果不对分母进行调整，会出现检索命中记录越少，$P(20)$值越高的现象。如果检索命中记录数为 0，分母就是 0，那么 $P(20)$就会是无穷大。因此对 $P(20)$分母的计算做如下调整：当检索输出结果少于 20 时，用 279 减去不够 20 的检索结果数量乘以 10。

例如，某次检索返回 15 条命中记录，其 $P(20)$的分母应该是 $279-5\times10=229$。如果返回命中记录数为 0，其 $P(20)$的分母为 $279-20\times10=79$。

综上所述，最后的计算公式为

$$P(20)=(R_{(1\sim3)}\times20+R_{(4\sim10)}\times17+R_{(11\sim20)}\times10)/(279-(20-N)\times10)$$

其中，R 代表各条命中记录的相关系数，N 为命中记录数(当命中记录数大于 20 时，$N=20$)。

这样，如果某一检索返回超过 20 条记录，其中前 15 条是相关记录，则 $P(20)=229/279$；如果命中记录数是 15，并且全部都是相关记录，则 $P(20)=229/229$；如果只返回一条记录且相关，$P(20)=20/89$；如果命中记录数是 0，$P(20)=0/79$。

Leighton 等人研究的替代方法很好地解决了网络环境下查准率难以确定所有相关信息数量的局限性。但上面的公式也存在一些问题，已有一些发表的成果对一些问题进行了改进。

3.1.3 查准率与查全率的关系

利用查准率和查全率指标，可以对每一次检索进行检索效率的评价，为检索的改进调整提供依据。利用这两个量化指标，也可以对信息检索系统的性能水平进行评价。

要评价信息检索系统的性能水平，就必须在一个检索系统中进行多次检索。每进行一次检索，计算其查准率和查全率，并以此作为坐标值，在平面坐标图上标示出来。通过大量的检索，就可以得到检索系统的性能曲线，如图 3-2 所示。

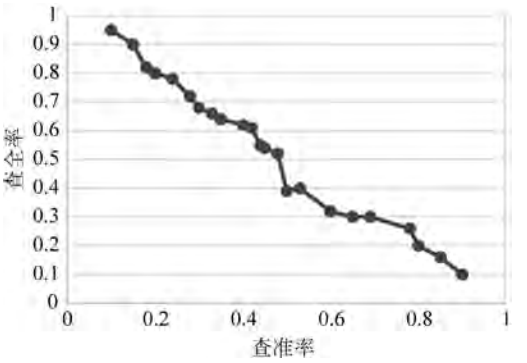

图 3-2 信息检索系统的性能曲线

大量的检索评价实验表明，查准率与查全率之间存在着特定的关系：在一个信息检索系统中，当查准率和查全率达到一定程度以后，两者就会呈现出非线性的反变关系。换句话说，在查准率不断提高的同时，查全率会持续下降；反之，在查全率不断提高的同时，查准率也会持续下降。一些专家认为，查全率大致在 60%～70%，查准率在 40%～50%时，查全率和查准率处于最佳比例关系，一旦查全率超过了 70%，要想提高查全率，就必须以牺牲查准率为前提条件。

查全率和查准率与文献的存储和信息检索是直接相关的，也就是说，与系统的收录范围、索引语言、标引工作和检索工作等有着非常密切的关系。要想做到查全，势必要对检索范围和限制逐步放宽，则结果是会把很多不相关的文献也带进来，影响了查准率。要使查全率和查准率都同时提高，并不是很容易。强调一方面，忽视另一方面，也是不妥的。因此，要根据具体信息检索需要，合理调节查全率和查准率，以保证检索效果。

值得注意的是，只有当查准率和查全率达到一定程度，两者之间才会呈现出这样的反变关系。如果查准率和查全率都很低，那么两者完全可以同时得到提高。查准率与查全率之间的这种反变关系，对于信息检索的实践具有极为重要的指导意义。

查准率和查全率是信息检索效率评价的量化指标，在检索系统的评价中也具有举足轻重的作用。其突出的好处在于，检索效率评价是一种结果评价，使检索评价变得简明、直观而易行。而其局限主要表现在以下两方面。

第一，它能够评价一次检索或一个系统的性能水平，却不能指出是什么原因产生了这样的检索效率。例如，两次检索或两个系统的查准率可能完全相同，但是其原因通常却不会完全相同。这样，就只能为检索的调整提供改进的方向，却不能指明需要改进的具体因素及措施。

第二，它以相关性为基础，具有相关性本身所固有的局限。例如，没有考虑文献的重要性程度等。

需要注意的是，信息检索的效率与信息检索系统的效率之间存在着密切的关联，但是也有着显著的区别。对于每一次检索而言，其检索效率的高低，不仅要依赖于检索系统的性能水平，而且还要取决于本次检索的具体措施和手段。

如果一个信息检索系统的查准、查全性能水平较低，那么在这样的系统中所进行的信息检索，一般而言查准率和查全率都会比较低；但是，倘若一次检索的措施和手段相当理想，也可能达到较高的检索效率。反之，如果一个信息检索系统具有较高的性能水平，那么在这样的系统中所进行的信息检索，通常就容易实现较高的查全率和查全率；但是，倘若一次检索的措施和手段都相当差，就会得到较低的检索效率。例如，对于传统的联机检索系统和现代的搜索引擎，在查准、查全的性能水平上前者要比后者高得多。但这并不意味每一次检索的结果必定如此。在利用联机系统进行检索时，如果选词不合理，措施和手段不当，就不可能达到系统的性能水平。同样，在利用搜索引擎进行检索时，如果检索的措施和手段相当理想，完全可以超越系统的平均性能水平。

3.1.4　漏检率和误检率

检索系统每进行一次检索，就把系统中所有的文献分为检出文献和未检出文献两个部分。前面已述，其中一部分是检出文献，指的是与检索策略相匹配并被检索出来的文献，用户根据自己的判断把它分成相关文献（a，合理的命中）和非相关文献（b，误检）；另一部分是未检出文献，指的是未能与检索策略相匹配的文献，也可以把它分成相关文献（c，漏检）和非相关文献（d，合理的排除）。

1. 信息检索的漏检率

漏检率（omission ratio）是查全率的补充指标，它们是一对互逆的指标，查全率高则漏检率就低，或反之。

$$\begin{aligned} O(\text{漏检率}) &= \text{未检出的相关文献数} / \text{系统中相关文献总数} \\ &= c/(a+c) \times 100\% \\ &= 1 - R(\text{查全率}) \end{aligned}$$

2. 信息检索的误检率

误检率(noise ratio 或 fallout ratio)是查准率的补充指标,它们是一对互逆的指标,查准率高则误检率就低,或反之。

$$\begin{aligned} N(\text{误检率}) &= \text{检出的不相关文献数} / \text{检出的文献总数} \\ &= b/(a+b) \times 100\% \\ &= 1 - P(\text{查准率}) \end{aligned}$$

3.1.5 响应时间

响应时间是从用户输入检索表达式开始查询到检出结果所需要的时间。显然,它也是检索中的一个重要指标。响应时间与多方面因素有关,不同的检索系统其响应时间的影响因素也各不相同。手工检索响应时间以人为因素较多,一般会比较长;单机检索系统的响应时间主要由系统的处理速度决定;网络环境下,响应时间则不仅取决于检索工具本身的响应速度,还在相当大的程度上取决于用户使用的通信设备和网络的拥挤程度等外部因素。

同一种检索系统在不同时间使用同一个检索表达式来检索同一课题,其响应时间可能有所不同。因此,在计算响应时间时,应该在相同的时间,在相同的软硬件环境下,对同一个检索课题的响应情况进行量化评价。另外,还要考虑系统是否具有记忆搜索结果加速调用的功能,方便用户使用常见词检索。

除了查全率、查准率和响应时间外,信息检索评价的指标一般还有收录范围、检索费用、信息的可用性、输出形式等。联机检索系统的评价指标体系如图 3-3 所示。

图 3-3 联机检索系统的评价指标体系

3.2 信息检索系统的评价

采用常规的方式来度量 ad hoc IR 系统的效果，需要一个测试集，它由以下三部分构成。

(1) 一个文档集。

(2) 一组用于测试的信息需求集合，信息需求可以表示成查询。

(3) 一组相关性判定结果，对每个查询-文档对而言，通常会赋予一个二值判断结果——要么相关，要么不相关。

常规的 IR 系统评价方法主要是围绕相关和不相关文档的概念来展开。对于每个用户信息需求，将测试集中的每篇文档的相关性判定看成一个二类分类问题进行处理，并给出判定结果：相关或不相关。这些判定结果称为相关性判定的黄金标准或绝对真理。测试集中的文档及信息需求的数目必须要合理：由于在不同的文档集和信息需求上的结果差异较大，所以需要在相对较大的测试集合上对不同信息需求的结果求平均。经验发现 50 条信息需求基本足够(同时 50 也是满足需要的最小值)。

需要指出的是，相关性判定是基于信息需求而不是基于查询来进行的，例如，可能有这样一个信息需求：在降低心脏病发作的风险方面，饮用红葡萄酒是否比饮用白葡萄酒更有效(原文是 whether drinking red wine is more effective at reducing your risk of heart attack than drinking white wine)。该需求可能会表达成查询 wine AND red AND white AND heart AND attack AND effective。一篇满足信息需求的文档是相关的，但这并不是因为它碰巧都包含查询中的这些词。由于信息需求往往并不显式表达，上述区别在实际上常常被误解。尽管如此，信息需求却始终存在。如果用户向 Web 搜索引擎输入“python”，那么他们可能想知道可以买宠物蛇的地方，或者想查找与编程语言 Python 相关的信息。对于单个词构成的查询，系统很难知道其背后的真实需求。当然，对于用户而言，他肯定有自己的信息需求，并且能够基于该需求判断返回结果的相关性。要评价一个系统，需要对信息需求进行显式的表达，以便利用它对返回文档进行相关性判定。迄今为止，我们对相关性都进行了简化处理，把相关性考虑为一个只具有如下尺度的概念：一些文档高度相关而其他却不太相关。也就是说，到现在为止，我们仅对相关性给出一个二值判定结果。应该说，这种简化具有一定的合理性。

许多系统都包含多个权重参数，改变这些参数能够调优系统的性能。通过调优参数而在测试集上获得最佳性能并报告该结果是不可取的。这是因为这种调节能在特定的查询集上获得最佳参数，而这些参数在随机给定的查询集上并不一定能够取得最佳性能，因此，上述做法实际上夸大了系统的期望性能。正确的做法是，给定一个或者多个开发测试集，在这个开发测试集上调节参数直至最佳性能，然后测试者再将这些参数应用到最后的测试集上，最后在该测试集上得到的结果性能才是真实性能的无偏估计结果。

习题

1. 查准率和查全率的定义分别是什么？它们之间有怎样的联系？

2. 什么是漏检率和误检率？

3. 响应时间的定义是什么？

4. 如何对一个信息检索系统进行评价？

5. 两个查询 q1、q2 的标准答案数目分别为 100 个和 50 个，某系统对 q1 检索出 80 个结果，其中正确数目为 40；系统对 q2 检索出 30 个结果，其中正确数目为 24。请计算每个系统的查准率和查全率。