arXiv:2304.00615v1 [cs.IR] 2 Apr 2023

信息检索评估指标的内在框架

Fernando Giner

E.T.S.I. Inform´atica UNED，C/ Juan del Rosal，16，28040-马德里，西班牙，
fginer3@gmail.com ORCiD：
0000-0002-9161-0458

摘要。 信息检索（IR）评估指标是衡量检索系统适用性和任务执行效率的基石。其度量与尺度属性使我们能够比较不同系统，从而确定它们之间的差异或相似之处。本文基于测量表征理论，利用检索度量本身所包含的信息来确定这些属性。本文确立了检索度量的 *intrinsic framework*，这是未明确指定定义域集合时的常见场景。本文提出了一种确定任意检索度量的度量与尺度属性的方法，只需了解其取得的部分数值。该方法依据检索度量的内在属性，将其划分为三个主要类别。本文还根据所提出的分类体系，对一些常见的面向用户和面向系统的评价指标进行了分类。

关键词： 信息检索、评价、度量、尺度

1 引言

信息检索（IR）是一个处理用户信息需求的分析、存储和检索的领域 [84]。为了有效比较并逐步开发更优秀的信息检索系统，评价发挥着重要作用。尽管评价是一种推动信息检索领域发展的丰富场景，但仍需要更深入地理解评估指标。具体而言，需要根据 Stevens [91]，的尺度类型对其进行分类，这有助于明确评估指标背后的假设，并证明根据所得结果得出的结论的有效性。

由于信息检索具有实证性质，这项任务也无法避免长期以来围绕尺度类型以及可采用的统计方法所存在的混淆与争议 [44,63,96]。近期，有两种观点探讨了信息检索评估度量[1] 作为区间尺度的作用。一方面，Moffat [66] 指出

[1] 在此，通常使用的术语“信息检索评估 *metric*”与数学术语“*metric*”相冲突，后者将在本文后续部分使用。为解决这一问题，本文其余部分将把术语“信息检索评估 *metrics*”称为“信息检索评价 *measures*”，而保留术语“*metric*”的数学含义。

只要检索度量具有现实基础（即外部效度），并且可以被阐释为与系统输出的效用相对应，就可以将其视为区间尺度。这一观点以修辞性论证为依据，与操作范式 [63,64]有相似之处。相比之下，Ferrante 等人 [31,32]以测量表征理论（RTM）为基础 [33]，探讨检索度量作为区间尺度的影响 [72,59,60,61]。在这一观点下（可称为表征范式 [63,64],），区间尺度是真实映射，其取值等间距。这一立场似乎有坚实的理论基础；然而，近期提出的论据 [33] 适用范围有限 [41]。

本文的目标并非质疑或赞同这两种观点中的任何一种，也不是寻找二者的共同点。事实上，由于争议在于是否假设 RTM，而 Moffat 的著作并未提及这一点，因此这两种观点所依据的测量理论并不相同。本文旨在通过将检索度量视为等间距映射 [31]，对信息检索评估度量进行分类，从而说明这一观点，并通过回应 [41] 的关切，进一步探讨支持该观点的论据。

检索度量可根据多个因素进行分类，例如定义域、值域、关系或表达式等。解析表达式可能是其中最重要的方面之一，因为它决定了检索系统的性能 [86]，或区分不同的用户浏览模型 [17,67]。除了解析表达式之外，定义域集合与值域集合中的元素之间的关系也决定了检索度量的类型。例如，定义域集合的度量性质可以判断待测的每个元素是否都能彼此区分 [37], ；而值域集合上的关系则决定了可执行的运算以及可应用的统计方法 [91]。本文基于 RTM 说明，检索度量所取得的值决定了其度量与尺度属性。本文由此建立了内在框架，利用检索度量本身所包含的信息，对信息检索评估度量进行分类。

内在框架并不局限于信息检索领域。在数据库或数据挖掘等领域，许多实证研究也采用了本文介绍的相同或相似的测量指标。在依赖 RTM 的研究中，尺度刻画对于针对待测属性进行统计推断至关重要。度量刻画也有助于确定与该测量指标相关的距离是否类似于欧几里得几何中的距离，抑或属于更具弹性的几何，其中“接近”的概念允许不同点之间的距离为零。在理论层面上，当检索度量未明确指定其经验域时，内在框架为其提供了默认的形式化基础。本文探讨了在 RTM 假设下检索度量的一些推论，这些内容对于挑战或认同表征范式的信息检索理论研究可能有所助益。

本文其余部分安排如下：第2节介绍一些相关工作。第3节和第4节阐述内在框架，并刻画其度量与尺度属性。第5节给出一个分类体系，用于确定检索度量的内在类别。第6节运用内在框架对一些常见的信息检索评估度量进行分类。最后，第7节总结全文。

2 相关工作

在信息检索领域，关于评价的研究成果丰硕 [45,74], ，因此评价被视为一个关键领域 [1]。TREC [98], 、NTCIR [83], 、CLEF [34], 、FIRE 和 INEX 等大规模评测活动和倡议，推动了学术界和工业界的发展。克兰菲尔德2号实验 [20], 被视为信息检索领域首次尝试进行评价 [24], ，也是许多现代实验所依据的框架。信息检索系统通常通过一组主题或搜索请求进行比较 [16],，并据此计算信息检索评估度量 [97,99]。信息检索评估度量的所得数值通常会结合显著性检验结果或置信区间来呈现 [18,49,80,87,93]。对信息检索系统的性能进行经验量化和统计评估，有助于确定其差异与相似之处 [79]。

对信息检索评估度量进行形式化分析，有助于更好地理解其含义。一些研究表明，检索度量对应着不同的用户浏览模型 [6,17,19,67,100,101]。另一些研究则通过形式属性刻画检索度量的效能 [2,4,48,65,88,92], ，这些属性有助于了解检索度量在特定场景中的适用性。Fang等人 [27,28,29,30],推动了将形式属性作为一种方法，用于探索检索模型及其改进方式，以实现更高的检索效能；这一方法已成功应用于基础模型 [26,27], 、伪相关反馈方法 [21,22,69], 、翻译检索模型 [54,71] 以及神经网络检索模型 [76]的研究。

已有若干研究从测量的视角探讨检索度量。van Rijsbergen [9,94] 通过基于精确率和召回率的联合（加性）结构，探讨了信息检索中测量的基础，随后考察了查准率—查全率结构上度量的性质。Bollman等人 [11]定义了类似的联合结构，但该结构基于二元检索的列联表；随后，他们研究了所提出的MZ度量的性质。后来，Bollman [10] 证明，只要满足提出的两条公理，检索度量就可以表示为相关/不相关已检索文档数量的线性组合。Flach [35] 通过混淆矩阵对经验域进行建模，进而勾勒出测量理论与机器学习评价之间的关系。然而，这些研究并未探讨检索度量的尺度性质。

一些研究将金标准视为测量对象，分析序数分类中的尺度性质 [7,39,95], 或

真实值、人工标注或预测变量的尺度性质 [43]。另一些研究则将系统输出和金标准视为相互独立的测量结果。这使得人们能够针对评估者量表与系统量表的相似度引入公理 [14,62]；并为大多数分类、排序和聚类测量指标提供单一且统一的解释 [5], ，或提供一种方法，用于确定给定数据挖掘问题最适合的任务／指标形式化 [3]。

最近，Fuhr [38]提出了一些用于衡量信息检索系统有用性的实验方案，例如避免使用平均倒数排名和期望倒数排名，因为它们违反了度量的基本要求。相比之下，Sakai [81] 对 Fuhr 的一些论述提出异议，因为这些论述并未解释检索度量与用户对有用性的感知之间的实验对齐。因此，作为这种分歧的结果或与之并行，目前存在一场活跃的对话 [31,66,32],，将检索度量视为用于量化检索系统有用性的真实映射。基于 RTM，Ferrante 等人 [31,32] 研究了信息检索测量指标作为区间尺度的影响。这一观点认为，区间尺度是真实映射，其取值等间距。他们提出，区间化是一种可行的技术，可用于获得有意义性。相比之下，Moffat [66] 认为，文档排序属于类别数据，检索度量是由数据集上下文定义的数值映射，并受某种外部现实的约束，只能取一组目标值。Moffat 声称，只要检索度量具有现实基础，即具备外部效度，就可以将其视为区间尺度；例如，将奖项分配给不同的排序类别。

这两种观点基于不同的假设来确定检索度量的量表类型。前者可称为表征范式，其论据基于 Ferrante 等人 [33]的研究。他们针对基于集合和基于排名的信息检索评估度量，以及二元和多级相关性，构建了一个框架，用以确定检索度量是否属于区间尺度。然而，所提出的论据适用范围有限 [41], ；通过允许检索度量自身指定定义域集合，可以解决这一问题。本文提出的内在框架遵循这一观点。

3 内在框架的形式化

在批量评估中，系统会向信息检索系统提交一个主题或查询，后者返回一个搜索引擎结果页面 （SERP）[2],$\hat{}$**r**作为输出。随后，评估指标以数值形式量化检索系统的效能，也就是它在检索相关文档的同时排除不相关文档的能力。因此，效能度量可以看作一种映射，将一组可能的排名**R**与实数关联起来；换言之，信息检索评价度量会为这些排名赋予数值

[2] 通常，搜索结果页面（SERP）以非同质的方式呈现内容，例如图像、查询建议、知识面板等。然而，这里我们考虑经典的有序（或无序）文档列表，因为这是研究排序模型评价时通常采用的结构。

（数值范围），它对应一组排名（经验域）。一旦定义了检索度量，就会同时对经验域和数值范围产生影响。

在经验域中，每个检索度量 f 都建立了一种隐含的排序关系 $\preceq_f$，该关系由其取得的值定义如下：

$$\hat{\mathbf{r}}_1 \preceq_f \hat{\mathbf{r}}_2 \iff f(\hat{\mathbf{r}}_1) \leq f(\hat{\mathbf{r}}_2)\ , \tag{1}$$

对于所有 $\hat{\mathbf{r}}_1, \hat{\mathbf{r}}_2 \in \mathbf{R}$；因此，每个信息检索评价度量都自然关联着一种排序关系，该关系本质上源自度量本身。根据这种二元关系 $\preceq_f$，任意一对元素都具有可比性，且其传递性显然成立；因此，它是一个弱序[3] [72]。排序结构 $(\mathbf{R}\ \preceq_f)$ 可用哈斯图表示，记作 G_f，其中节点标记为 $\mathbf{R}$ 中的元素。节点根据 f 的取值置于不同层级。若一个元素的取值大于另一个元素，且二者之间不存在其他排名，则用一条边表示这种关系。例如，某个检索度量 f 的关联图满足$f(\hat{\mathbf{r}}_1) < f(\hat{\mathbf{r}}_2) = f(\hat{\mathbf{r}}_3) = f(\mathbf{r}_4) < f(\hat{\mathbf{r}}_5) < f(\hat{\mathbf{r}}_6)$，如图 1 所示。

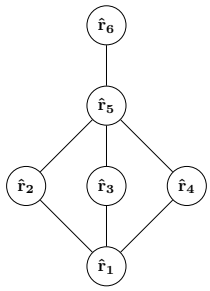


图 1. 与某个检索度量相关联的哈斯图示例，G_f。

在数值范围内，信息检索评价度量所取得的值使我们能够比较不同系统，从而确定它们之间的关系、差异或相似之处。通常通过考察所取得值的绝对差异或相对差异，或者基于这种差异进行验证性检验，来比较一对检索系统的效能 [86]。因此，为了进行比较，任何检索度量都自然关联着一个距离，

[3] 相关的弱序 $\preceq_f$ 可以通过考虑以下等价关系转换为全序：$\hat{\mathbf{r}}_1 \sim_f \hat{\mathbf{r}}_2 \Leftrightarrow f(\hat{\mathbf{r}}_1) = f(\hat{\mathbf{r}}_2)$。令 R^* 为等价类的集合，并令 $\hat{\mathbf{r}}_1^*$ 和 $\hat{\mathbf{r}}_2^*$ 为该集合中的两个元素，分别包含各个系统输出排序 $\hat{\mathbf{r}}_1, \hat{\mathbf{r}}_2 \in \mathrm{R}$。可以在 R^* 上定义如下排序关系：$\hat{\mathbf{r}}_1^* \preceq_f^* \hat{\mathbf{r}}_2^* \Leftrightarrow \hat{\mathbf{r}}_1 \preceq_f \hat{\mathbf{r}}_2$。于是，$(\mathrm{R}^*, \preceq_f^*)$ 称为 $(\mathrm{R}\ \preceq_f)$ 的*reduction*或 *quotient*，其中 $\preceq_f^*$定义良好，且 $(\mathrm{R}^*, \preceq_f^*)$ 是一个全序集 [72]。

$d_f: \mathbf{R} \times \mathbf{R} \longrightarrow \mathbb{R}$，即两个取得值之差的绝对值，通常也称为它们之间的欧几里得距离：

$$d_f(\hat{\mathbf{r}}_1, \hat{\mathbf{r}}_2) = |f(\hat{\mathbf{r}}_1) - f(\hat{\mathbf{r}}_2)| \tag{2}$$

注 *1.* 通过为相关哈斯图 G_f 的每条边赋予权重，即将值 $|f(\hat{\mathbf{r}}_1) - f(\hat{\mathbf{r}}_2)|$赋给 $\hat{\mathbf{r}}_1\hat{\mathbf{r}}_2$, ，显然可知，一对元素之间的距离 d_f 就是该边加权图上的最短路径距离；也就是说， d_f 是哈斯图 G_f 上的自然距离。

因此，每个信息检索评价度量 f 都有一个与之自然关联的上下文，该上下文由度量本身内在导出。此上下文由其关联的排序关系 $\preceq_f$ 和距离 d_f 构成。这两种数学工具与检索度量内在相连，因为排序关系直接反映了其设计目的，而距离则是对所得数值进行比较的常用方式。它们是只取决于检索度量本身定义的内在实体，可将二者结合起来，建立以下概念。

定义 1. 信息检索评价度量 f 的内在框架 是配有相应排序关系 $\preceq_f$ 和相应距离 d_f 的所有可能排名的集合；它表示为$(\mathbf{R}, \preceq_f, d_f)$。

内在框架与曲线和曲面的内在几何 [25,40] [4]这一数学概念密切相关。考虑与检索度量相关的排序关系的约化或商集（$\mathbf{R}^*, \preceq_f^*$）。由于 $\preceq_f^*$ 是全序，其相应的哈斯图 G_f^* 是一条链，用直线连接相邻元素。按照注释 1 所示为每条边赋予权重后，哈斯图 G_f^* 可以被视为一条独立的曲线，其中任意一对相邻元素之间的距离就是相应的权重。内在框架表示 G_f^* 的内在几何，其距离通过最短路径长度计算。严格来说，内在框架并非 G_f^* 的内在几何，因为它是一条缺乏可微性的离散曲线。不过，这一内在框架使我们能够利用源自检索度量本身的数学工具来度量 G_f^* 上的距离。

对检索度量的内在框架进行比较是有意义的，因为这些比较是从度量本身推导出来的。如果考虑不同的检索度量，那么内在框架的排序关系也会改变。不过，用于衡量新度量最短路径长度的准则仍将

[4]设想一些生活在二维欧几里得空间表面上的假想生物， $\mathbb{R}^2$。它们不知道周围存在三维空间（但能感知欧几里得距离）。这些生物是局部观察者，其视野仅限于一个二维环境。它们能够观察或测量的该表面的几何元素（本质上是长度），构成了该表面的 *intrinsic geometry* 。表面的内在属性仅取决于表面本身。

保持一致，因为它满足距离的公理性质。内在框架的另一项性质是，它适用于任意规模的排名，因为对此没有作出任何假设。此外，由于没有对相关性等级作出假设，内在框架可以处理二元或多级相关性标签。

4 信息检索评价度量的内在性质

本节将阐述内在框架的度量与尺度属性。

4.1 信息检索评价度量的度量性质

回顾度量空间的一些基本概念 [37,47]。考虑一个信息检索评价度量 f；其关联距离 d_f 若满足以下两个性质，则为伪度量：(i) 对称性，$d_f(\hat{}\mathbf{r_1}, \hat{}\mathbf{r_2}) = d_f(\hat{}\mathbf{r_2}, \hat{}\mathbf{r_1})$；以及 (ii) 三角不等式，$d_f(\hat{}\mathbf{r_1}, \hat{}\mathbf{r_3}) \leq d_f(\hat{}\mathbf{r_1}, \hat{}\mathbf{r_2}) + d_f(\hat{}\mathbf{r_2}, \hat{}\mathbf{r_3})$。若 d_f 还满足 (iii) 不可分辨者同一性，$d_f(\hat{}\mathbf{r_1}, \hat{}\mathbf{r_2}) = 0 \Leftrightarrow \hat{}\mathbf{r_1} = \hat{}\mathbf{r_2},$，则 d_f 为度量。

严格来说，必须区分评价度量 f 和距离 d_f。根据上一段，度量性质应应用于距离，例如“该距离 是伪度量”。然而，在检索场景中，由于距离是从信息检索评价度量导出的，因此人们通常会将度量性质归于检索度量，例如“该检索度量 是度量”。下述结果表明，每个检索度量都是伪度量。

命题1. 设 $(\mathbf{R} \preceq_f, d_f)$ 是信息检索评价度量 f 的内在框架，则其关联距离 d_f 是伪度量。

在检索场景中，人们创造了“信息检索评价度量”这一术语，用来指代任何信息检索评价度量。然而，下述结果表明，并非每个检索度量都是度量。

命题 2. 设 $(\mathbf{R} \preceq_f, d_f)$ 是信息检索评价度量的内在框架，若 f 是单射或一一函数[5]，则其关联距离 d_f 是度量。

因此，每个检索度量都是伪度量；只有为每个系统输出排序赋予不同值的检索度量才是度量。

[5] 在基础代数中，[36,50,51] f 是单射，如果 f 将不同的元素映射到不同的元素，形式化地说：$f(\hat{}\mathbf{r_1}) = f(\hat{}\mathbf{r_2})$ 意味着 $\hat{}\mathbf{r_1} = \hat{}\mathbf{r_2}, \forall \hat{}\mathbf{r_1}, \hat{}\mathbf{r_2} \in \mathbf{R}$。

4.2 信息检索评价度量的尺度性质

从 RTM 的角度来看，将实数一致地赋予经验域，是对待测属性的一种尺度 [90]。Stevens [91]区分了四种主要的测量尺度：名义、顺序、区间和比率。通过考察底层经验域上的一种排序关系（ $\trianglelefteq$ ），能够反映或保持这种排序关系的尺度（ φ ）称为序数尺度，形式化地表示为： $a \trianglelefteq b \Leftrightarrow \varphi(a) \leq \varphi(b)$ 。

在检索场景中，定义在排序结构（$\mathbf{R} \preceq_f$）上的每个信息检索评价度量 f 都是顺序尺度，因为相关排序关系的定义（见公式 1）仅验证了这一性质。

检索度量的内在框架（$\mathbf{R} \preceq_f, d_f$）允许考虑一个区间，即两个端点之间元素的集合，形式化表示为：$[\hat{}\mathbf{r_1}, \hat{}\mathbf{r_2}] = \{\hat{}\mathbf{r} \in \mathbf{R} : \hat{}\mathbf{r_1} \preceq_f \hat{}\mathbf{r} \preceq_f \hat{}\mathbf{r_2}\}$。区间可用其基数或区间跨度来量化，记为 $\Delta_{\hat{\mathbf{r}}_1 \hat{} \mathbf{r}_2}$ ，，形式化表示为：$\Delta_{\hat{\mathbf{r}}_1 \hat{} \mathbf{r}_2} = |[\hat{}\mathbf{r_1}, \hat{}\mathbf{r_2}]|$。区间的跨度表示一对排名之间的间隔有多少。

然后，可以在所有可能的区间集合上定义排序关系，如下所示：$[\hat{}\mathbf{r_1}, \hat{}\mathbf{r_2}] \preceq_{d_f} [\hat{}\mathbf{r_3}, \hat{}\mathbf{r_4}] \Leftrightarrow \Delta_{\hat{\mathbf{r}}_1 \hat{} \mathbf{r}_2} \leq \Delta_{\hat{\mathbf{r}}_3 \hat{} \mathbf{r}_4}$。这一排序关系 $\preceq_{d_f}$ 是一种弱序，因为任意一对区间都可比较，且传递性显然成立。

根据 RTM [72], ，定义在排序结构（$\mathbf{R} \preceq_f$）上的信息检索评价度量 f 若保持差值，即将等间距区间映射为相等的差值，则称为区间尺度 ，形式化地：

$$[\hat{\mathbf{r}}_1, \hat{\mathbf{r}}_2] \preceq_{d_f} [\hat{\mathbf{r}}_3, \hat{\mathbf{r}}_4] \Longleftrightarrow f(\hat{\mathbf{r}}_2) - f(\hat{\mathbf{r}}_1) \leq f(\hat{\mathbf{r}}_4) - f(\hat{\mathbf{r}}_3) \ .$$

以下结果刻画了区间尺度。

命题 3. 考虑一个顺序尺度 f，它是一种度量，那么 f 是区间尺度，当且仅当其取值等间距。

因此，每个检索度量都是其关联排序关系上的顺序尺度；只有取值等间距的检索度量才是区间尺度。

5 信息检索评价测量指标的内在分类体系

第 4 节中的特征描述使我们能够将检索度量分为三大类别，并提供了一条可直接使用的规则，以根据所得数值识别内在属性：

1. 每个信息检索评价度量 f 都是其内在框架上的顺序尺度和伪度量。默认情况下，任何检索度量都具有这些属性。此类别称为**顺序/伪度量**。2. 如果 f 的取值对于每一种系统输出排序都各不相同，即该检索度量是一一函数，那么 f 就是度量（不一定是区间尺度）。这些检索度量属于**顺序 / 度量**。

3. 如果 f 的取值间距相等，那么该检索度量就是区间尺度。此类别称为**区间/度量**。

快速浏览这一分类体系即可确认，其中不存在比率尺度。这是因为，内在框架旨在根据检索度量本身所包含的信息，即其相关的排序关系和距离，推导出信息检索评价度量的度量与尺度属性。比率尺度需要在排名之间增加一种运算 [64,72]，而检索度量的定义中并未包含这种运算。因此，信息检索评估指标可以是比率尺度，但必须事先规定一种作用于系统输出排名的外部运算。

6 一些示例

本节将根据第5节的分类体系，对一些常见的检索度量进行分类。每种检索度量的经验域都是其内在框架[6]。

6.1 基于集合的检索

在这种情况下，信息检索系统会返回一组响应查询的文档；用 1 表示相关文档、用 0 表示不相关文档，系统输出的一些示例如下：$\hat{\mathbf{r}}_1 = \{0,0,0,0,0\}$、$\hat{\mathbf{r}}_2 = \{1,0,0,0,0\}$ 和 $\hat{\mathbf{r}}_3 = \{1,1,0,0,0\}$。一般而言，系统输出可以用由两个二元变量构成的列联表进行总结：相关性和检索。表 1 展示了频数分布，其中两个因素同时列出。

表1. 二元情形下基于集合的检索列联表。

	相关文档	**不相关文档**
文档已检索	tp 真正例	fp 假正例
文档未检索	fn 假负例	tn 真负例

[6] 如第 4 节所述，使用此框架推导出的检索度量的内在属性基于 RTM。

一些广泛使用的评估指标包括召回率 [9], 精确率 [9], 虚警率 [9],分类准确率 [8], 漏检率和错误率，定义如下:

$$\text{recall} = \frac{tp}{(tp+fn)} \qquad \text{precision} = \frac{tp}{(tp+fp)}$$
$$\text{fallout} = \frac{fp}{(fp+tn)} \qquad \text{classification accuracy} = \frac{(tp+tn)}{(tp+fn+fp+tn)}$$
$$\text{miss rate} = \frac{fn}{(tp+fn)} \qquad \text{error rate} = \frac{(fp+fn)}{(tp+fn+fp+tn)}$$

这些都是度量，因为其解析表达式的分子至少包含以下项之一： tp、 tn、 fp 或 fn；每当系统输出中出现一个新的相关文档时，这些项都会有所不同。此外，它们也是区间尺度，因为其解析表达式的分母为常数，而每检索到一个相关文档，分子就增加一个单位。因此，它们取得的值是等间距的。

其他相关的评价测量指标包括逆召回率、逆精确率、特异度、假发现率、漏报率和F度量 [9]„ 定义如下:

$$\text{inverse recall} = \frac{tn}{(fp+tn)} \qquad \text{inverse precision} = \frac{tn}{(fn+tn)}$$
$$\text{specificity} = \frac{tn}{(tn+fp)} \qquad \text{false discovery rate} = \frac{fp}{(fp+tp)}$$
$$\text{false omission rate} = \frac{fn}{(fn+tn)} \qquad \text{F-measure} = \frac{2 \cdot \text{prec} \cdot \text{recall}}{(\text{prec}+\text{recall})}$$

F-度量是一种度量，因为它会取得不同的值（调和平均数是递增函数）。其余测量指标都是度量和区间尺度，因为它们符合上一段所述的相同性质。然而， F-度量不是区间尺度；例如，考虑一个包含15篇文档的集合，其中5篇与某个主题相关，则 $F(\hat{\mathbf{r}}_1) = 0.000$、$F(\hat{\mathbf{r}}_2) = 0.300$ 和 $F(\hat{\mathbf{r}}_3) = 0.509$。

普遍性因子或流行率 $= (tp+fn)/(tp+fn+fp+tn)$ [9] 是一个伪度量，因为对于每个系统输出，其所有取值都相同。

效用度量 [85], $utility = \alpha \cdot tp + \beta \cdot fn + \gamma \cdot fp + \delta \cdot tn$ 中， α, β, γ和 δ 是用户指定的正权重；这四个参数的不同取值构成了多种可能性。一般而言，如果四个参数各不相同，那么该效用度量就是一个度量，而某些参数组合可能会产生区间尺度。

一些面向用户的度量基于以下四个变量: （i）用户已知的相关文档总数: U； （ii）用户已知且已检索的相关文档数: R_k； （iii）用户未知但已检索的相关文档数: R_u；以及（iv）已检索文档数: A。面向用户的度量示例包括覆盖率 [58], 、检索召回率 [58], 、新颖度比率 [58], 和召回努力 [58], ，其定义如下

:

$$\text{coverage ratio} = \frac{R_k}{U} \qquad \text{retrieval recall} = \frac{(R_k + R_u)}{U}$$
$$\text{novelty ratio} = \frac{R_u}{(R_u + R_k)} \qquad \text{recall effort} = \frac{U}{A}$$

通过考察以下两种系统输出，可以看出相关性召回率和新颖度比率并非度量指标：(i)$\hat{\mathbf{r}}_5 = \{1$ 个用户已知的已检索相关文档$\}$；以及(ii)$\hat{\mathbf{r}}_6 = \{1$ 个用户已知的已检索相关文档，检索到 1 个不相关文档$\}$。相关性召回率和新颖度比率在这两种系统输出上的取值相同。召回努力也不是度量指标，因为它在以下两种系统输出上的取值相同：(i)$\hat{\mathbf{r}}_7 = \{1$ 个用户未知的已检索相关文档，检索到 1 个不相关文档$\}$；以及(ii)$\hat{\mathbf{r}}_8 = \{2$ 个已检索的不相关文档$\}$。通过考察仅相差一个已检索不相关文档的两种系统输出，可以看出覆盖率不是度量指标。

表 2 总结了这些检索度量的内在属性。

表 2. 基于集合的检索中若干检索度量的内在属性。

	ord/pseudom	**ord/metr**	**interv/metr**
召回率 [9]			✓
精确率 [9]			✓
虚警率 [9]			✓
漏检率			✓
分类准确率 [8]			✓
错误率			✓
逆召回率			✓
逆精确率			✓
特异度			✓
假发现率			✓
漏报率			✓
F-度量 [9]		✓	
普遍性因子 [9]	✓		
覆盖率 [58]	✓		
检索召回率 [58]	✓		
新颖度比率 [58]	✓		
召回努力 [58]	✓		

6.2 基于排序的检索

在这种情况下，信息检索系统会针对查询返回一个文档排序列表，$\hat{\mathbf{r}} = \langle d_1, d_2, \ldots \rangle$。文档检索完成后，在多级

相关情形中，使用相关性判定列表来确定检索到的文档与主题的相关性。设 r 为 $\hat{\mathbf{r}}$ 的一个排名位置，则每个检索到的文档 d_r 都会被赋予一个相关性等级 a_r，其中 $r \geq 1$。例如，在二元情形中，用 a_1 表示相关文档，用 a_0 表示不相关文档，以下是一些长度为四的排名示例：$\hat{\mathbf{r}}_1 = \langle \mathsf{a}_1, \mathsf{a}_0, \mathsf{a}_0, \mathsf{a}_0 \rangle$、$\hat{\mathbf{r}}_2 = \langle \mathsf{a}_0, \mathsf{a}_1, \mathsf{a}_0, \mathsf{a}_0 \rangle$和 $\hat{\mathbf{r}}_3 = \langle \mathsf{a}_0, \mathsf{a}_1, \mathsf{a}_0, \mathsf{a}_1 \rangle$。这些相关性等级可以是类别标签。为处理数值，本文采用增益函数 $gain(\cdot)$，为每个相关性等级赋予一个正实数。排名 r 处的增益记为 $g(r) = gain(\mathsf{a}_r)$，其中 $g(r) = 0$ 适用于不相关文档。例如，在二元情形中，若 $g(\mathsf{a}_1) = 1$ 且 $g(\mathsf{a}_0) = 0$，则前述排名示例可表示为 $\hat{\mathbf{r}}_1 = \langle 1, 0, 0, 0 \rangle$、$\hat{\mathbf{r}}_2 = \langle 0, 1, 0, 0 \rangle$ 和 $\hat{\mathbf{r}}_3 = \langle 0, 1, 0, 1 \rangle$。排名 r 处的累积增益是截至排名 r的总增益，其定义为$cg(r) = g(r) + g(r-1)$，其中 $r > 1$ 且 $cg(1) = g(1)$。

在这种多级情境中，排名 r [9,15,33] 处的精确率可表示如下：

$$\text{Prec@}r = \frac{cg(r)}{r} \ ,$$

该检索度量是一个伪度量，而非度量，因为满足 $\text{Prec@4}(\hat{\mathbf{r}}_1) = 0.250 = \text{Prec@4}(\hat{\mathbf{r}}_2)$。类似地，排名 r 处的召回率是一种伪度量，因为其解析表达式同样基于r处的累积增益。

在此场景中，可以通过先列出相关性等级较高的所有文档，再列出相关性等级紧接其后的所有文档，依此类推，来确定一个理想排序。设 $ig(r)$ 表示理想排序中排名 r 处的增益，$cig(r)$ 表示理想排序中排名 r 处的累积增益，L 表示排序的长度，则滑动比率 [70,58]:

$$\text{sr} = \frac{cg(L)}{cig(L)} \ ,$$

它是一个伪度量，因为 $\text{sr}(\hat{\mathbf{r}}_1) = 1.000 = \text{sr}(\hat{\mathbf{r}}_2)$。滑动比率的一种修正版 [77]:

$$\text{msr} = \frac{\sum_{r=1}^{L} \frac{1}{r} g(r)}{\sum_{r=1}^{L} \frac{1}{r} ig(r)} \ ,$$

由于它对排序顺序敏感，因此它是一种度量。不过，它不是区间尺度，因为 $\text{msr}(\hat{\mathbf{r}}_1) = 1$、$\text{msr}(\hat{\mathbf{r}}_2) = 0.5$ 和 $\text{msr}(\langle 0, 0, 1, 0 \rangle) = 0.333$。

为了介绍以下检索度量，我们需要引入指示函数 $isrel(r)$，其定义为：如果排名 r 处的文档相关，则 $isrel(r) = 1$；否则，$isrel(r) = 0$。因此，$count(r) = \sum_{k=1}^{r} isrel(k)$ 的值是排序中前 r 个文档内相关文档的数量。

设 R 为检索到的相关文档总数，则 R-精确率为：

$$\text{R-precision} = \frac{count(R)}{R} \ ,$$

是一个伪度量；考虑排序 $\hat{\mathbf{r}}_4 = \langle 1, 0, 0, 1 \rangle$，可得 R 精确率$(\hat{\mathbf{r}}_3) = 0.500 =$R 精确率$(\hat{\mathbf{r}}_4)$。

归一化召回率 [75]:

$$\mathrm{R}_{\mathrm{norm}} = 1 - \frac{\sum_{r=1}^{R} isrel(r) \cdot r - \sum_{r=1}^{R} r}{R \cdot (L - R)} \ ,$$

在二元情形下，它是区间尺度，因为每个已检索的相关文档都会为该度量贡献一个单位。

归一化精确率 [75]:

$$\mathrm{P}_{\mathrm{norm}} = 1 - \frac{\sum_{r=1}^{R} isrel(r) \cdot \ln r - \sum_{r=1}^{R} \ln r}{\ln \frac{L!}{R! \ (L-R)!}} \ ,$$

是度量，但不是区间尺度。它满足与归一化召回率相同的性质；然而，对数并非恒定的递增函数。

加权 R 精确率 [53]:

$$\text{R-WP} = \frac{cg(R)}{cig(R)} \ ,$$

是伪度量，因为它在二元情形下取得与 R 精确率相同的值。R 度量 [78], 定义为:

$$\text{R-measure} = \frac{cg(R) + count(R)}{cig(R) + R} \ ,$$

在二元情形下，其取值与 R 精确率相同，因此它是一个伪度量。在多级相关情形下，R 度量和 R-WP 也是伪度量，因为对于所有排名，只要排名前 R个文档都（至少部分）相关，它们的取值就是 1。

平均精确率 [13,46]:

$$\mathrm{AP} = \frac{1}{R} \cdot \sum_{r=1}^{L} isrel(r) \cdot \frac{count(r)}{r} \ ,$$

不是一个度量，因为 $\mathrm{AP}(\hat{}\mathbf{r_1}) = 0.250 = \mathrm{AP}(\hat{}\mathbf{r_3})$。由于 AP 不是区间尺度，因此，对于一组 Q 个查询，平均平均精确率 $\mathrm{MAP} = \frac{1}{Q} \sum_{i=1}^{Q} AP_i$以及几何平均平均精确率 $\mathrm{GMAP} = \exp \frac{1}{Q} \sum_{i=1}^{Q} \log AP_i$，根据尺度类型的允许运算 [91†]，都不应被视为合适的度量，因为它们是序数值的平均数。这一结果证实了 [73]的发现。加权平均精确率 [53]:

$$\mathrm{AWP} = \sum_{r=1}^{L} isrel(r) \cdot \frac{cg(r)}{cig(r)} \ ,$$

不是一种度量。考虑仅包含两个与查询相关的文档的排名 $\hat{}\mathbf{r_1}$ 和$\hat{}\mathbf{r_3}$, , 则 $\mathrm{AWP}(\hat{}\mathbf{r_1}) = 0.250 = \mathrm{AWP}(\hat{}\mathbf{r_3})$。

Q 度量 [78],定义如下:

$$\text{Q-measure} = \frac{1}{R} \cdot \sum_{r=1}^{L} isrel(r) \cdot \frac{cg(r) + count(r)}{cig(r) + r} \ ,$$

并非度量，因为 Q 度量$(\hat{}\mathbf{r_1}) = 0.250 =$ Q 度量$(\hat{}\mathbf{r_3})$（在二元情形下，其取值与平均准确率相同）。

排名为 r 时的倒数排名:

$$\text{RR}_r = isrel(r) \cdot \frac{1}{r} \ ,$$

不是度量，因为 $\text{RR}_4(\hat{}\mathbf{r_2}) = 0.500 = \text{RR}_4(\hat{}\mathbf{r_3})$。因此，期望倒数排名 [19,89]:

$$\text{ERR} = \frac{1}{Q} \sum_{i=1}^{Q} \text{RR}_i \ ,$$

根据尺度类型的允许运算，不应将其纳入考虑 [91] 因为它是序数值的均值。

折损累计增益 [57,52]:

$$\text{DCG}_b = \sum_{r=1}^{L} \frac{g(r)}{\max\{1, \log_b r\}} \ ,$$

不是度量，因为 $\text{DCG}_2(\hat{}\mathbf{r_1}) = 1.000 = \text{DCG}_2(\hat{}\mathbf{r_2})$。

分级排序偏置精确率 [68,82]:

$$\text{RBP}_p = \frac{1-p}{g(a_c)} \cdot \sum_{r=1}^{L} p^{i-1} \cdot g(r)$$

一般而言，它不是度量。例如，考虑排名 $\langle 1,0,0\rangle$ 和$\langle 0,1,1\rangle$，相应的得分分别为$(1+0+0)/(1-p)$和$(0+p+p^2)/(1-p)$。令这些表达式相等，可得到 $1 = p+p^2$ 的一个实数解。因此，对于这个参数 p，存在两个得分相同的排名。因此，RBP_p 不是区间尺度。不过，在某些特定情况下，它是区间尺度，因为其取值等间距；例如，$\text{RBP}_{0.5}$ [31]。

二元偏好评估指标 [12]:

$$\text{bpref} = \frac{1}{R} \cdot \sum_{r=1}^{L} 1 - \frac{r - count(r)}{R} \ ,$$

是一种伪度量，而非度量，因为对于一种排序（其中第一个排序位置上的相关文档数为一个）和另一种排序（其中第一个和第二个排序位置上的相关文档数各为一个），它取得的值相同。

累积增益（CG）的一种扩展是检索度量 XCG [55] 系列。它们考虑 XML 元素之间的依赖关系，例如重叠和近似匹配。本文通过定义 $\mathbf{xG}[r] = g(r)$，对这些测量指标进行了调整，适用于每个排序位置 r。面向用户的归一化扩展累积增益度量 [56]:

$$\mathrm{nxCG}[r] = \frac{xCG[i]}{xCI[i]} = \frac{cg(r)}{cig(i)} \ ,$$

不是一种度量，因为当只有一个相关文档时，$\mathbf{nxCG}[4](\hat{}\mathbf{r_1}) = 1.000 = \mathbf{nxCG}[4](\hat{}\mathbf{r_2})$。因此，排名为 r [56] 时的平均 nxCG:

$$\mathrm{MAnxCG}[r] = \frac{\sum_{j=1}^{r} \mathrm{nxCG}[j]}{r} = \frac{\sum_{j=1}^{r} \frac{cg(j)}{cig(j)}}{r}$$

根据尺度类型上的允许运算 [91]，不应予以考虑。此外，若将其视为一种测度，它也不是一种度量，因为 $\mathbf{MAnxCG}[4](\hat{}\mathbf{r_1}) = 0.250 = \mathbf{MAnxCG}[4](\hat{}\mathbf{r_2})$。

系统导向的努力精确率 / 增益召回率 [56]:

$$\mathrm{gr}[r] = \frac{\mathrm{xCG}[r]}{\mathrm{xCI}[L]} = \frac{cg(r)}{cig(n)}$$

不是一种度量，因为 $\mathbf{gr}[4](\hat{}\mathbf{r_1}) = 1.000 = \mathbf{gr}[4](\hat{}\mathbf{r_2})$。

最后，期望搜索长度 [23], 定义如下:

$$\mathrm{esl} = j + \frac{i \cdot s}{t+1}$$

其中，j 是最终等级之前所有等级中不相关文档的总数；t 是最终等级中相关文档的数量；i 是最终等级中不相关文档的数量；s 是根据需求类型，为满足需求而需要从最终等级中获取的相关文档数量。对于类型 2 检索，期望搜索长度不是一种度量，因为对于两个仅在某一特定等级中文档顺序不同的排序，它得到的值相同。

一般而言，在许多检索度量的非二元情形中，不同排序可能通过多种方式获得相同分数。例如，若可能的 qrel 值为：$\{a_0 = 0,\ a_1, a_2, a_3, a_4 = 1\}$，则有三个变量可以组合起来，使不同排序映射到相同分数。表 3 总结了这些检索度量的内在属性。

7 结论

如第 1 节所述，确定检索度量量表类型的方法有多种。第一步应当从理论上论证其

表 3. 基于排序的检索中一些检索度量的内在属性。

	序数/伪序数	**序数/度量**	**间隔/度量**
Prec@r [9,15]	✓		
R 精确率	✓		
滑动比率 [58]	✓		
修正滑动比率		✓	
$\mathrm{R}_{\mathrm{norm}}$ [75]			✓*
$\mathrm{P}_{\mathrm{norm}}$ [75]		✓	
R-WP [53]	✓		
R 度量 [78]	✓		
平均精度 [13,46]	✓		
AWP [53]	✓		
Q 度量 [78]	✓		
RR	✓		
折损累积增益$_b$ [57]	✓		
基于排名的精确率$_p$ [68,82]	✓		
二元偏好 [12]	✓		
归一化扩展累积增益[r] [56]	✓		
平均归一化扩展累积增益[r] [56]	✓		
分级召回率[r] [56]	✓		
期望搜索长度 [23]	✓		

(*)仅适用于二元情形。

论据，以明确检索度量背后的假设。只有在这些论据得到恰当论证后，才能评估其中一种、两种或其他替代方案是否有效。

这里获得的结果对应于表征范式，即假定 RTM 时的情形。当检索度量的经验域未被明确指定时，本文为其度量与尺度属性提供了理论基础。这些属性是从检索度量本身所包含的信息中推导出来的，即它们属于内在属性。本文介绍了一套分类体系，以及一种基于所得数值的现成规则，并依据其内在属性对一些常见的面向用户和面向系统的检索评价指标进行了分类。研究发现，基于集合的（第一代）检索度量的优势在于其形式属性，其中大多数是度量和区间尺度。因此，根据 Stevens [91] 的尺度类型允许的运算，可以对其所得数值执行涉及排序、加法或差值的运算。另一方面，基于排名的（现代）检索度量试图更准确地捕捉系统有用性的某些方面，而舍弃了其形式属性。一般而言，它们是伪度量和顺序尺度；因此，根据允许的运算，只应执行涉及其所得数值排序的运算。由此，检索度量在满足形式属性与捕捉用户对有用性的感知之间面临权衡。

在假定 RTM 的情况下，内在框架有助于确定新出现的或现有的信息检索评估度量的度量与尺度属性。此外，它还能够研究那些完全取决于检索度量本身的其他属性，这是一个值得在未来工作中探索的有趣课题。

参考文献

1. Allan, J., Aslam, J., Belkin, N., Buckley, C., Callan, J., Croft, B., Dumais, S., Fuhr, N., Harman, D., Harper, D.J. 等：信息检索与语言建模中的挑战：信息检索智能中心、马萨诸塞大学阿默斯特分校于2002年9月举办的研讨会报告。载于：《ACM SIGIR论坛》，1，页31–47。ACM，美国纽约州纽约市（2003年） 2. Amig´o, E., Gonzalo, J., Artiles, J., Verdejo, F.：基于形式约束的外部聚类评估指标比较。《信息检索》**12**(4)，461–486（2009年） 3. Amigo, E., Gonzalo, J., Mizzaro, S.：我的问题是什么？基于测量理论识别数据挖掘中的形式任务与度量指标。《IEEE 知识与数据工程汇刊》（2021年） 4. Amig´o, E., Gonzalo, J., Verdejo, F.：一种用于文档组织任务的通用评估指标。载于：第36届 ACM SIGIR信息检索研究与发展国际会议论文集，页643–652（2013年） 5. Amig´o, E., Mizzaro, S.：论信息访问评估指标的本质：统一框架。《信息检索期刊》**23**(3)，318–386（2020年） 6. Azzopardi, L., Thomas, P., Craswell, N.：衡量搜索引擎结果页面的效用：一种基于信息觅食的度量。载于：第41届 ACM SIGIR信息检索研究与发展国际会议，页605–614（2018年） 7. Baccianella, S., Esuli, A., Sebastiani, F.：序数回归的评估指标。载于：2009年第九届智能系统设计与应用国际会议，页283–287。电气电子工程师学会（2009年） 8. Belew, R.K., Belew, R.K.：了解信息：搜索引擎技术与万维网的认知视角。剑桥大学出版社（2000年） 9. Blair, D.C.：信息检索，第2版。C.J. van Rijsbergen。伦敦：Butterworths。《JASIS》**30**4630300621 10 6 374 375 1979 10 1002 ( )，–（）。DOI . /asi. . Bollmann, P.：信息检索评估指标的两条公理。载于：SIGIR，第84卷，页233–245。Citeseer（1984年） 11. Bollmann, P., Cherniavsky, V.S.：MZ度量的测量理论研究。载于：第3届 ACM 信息检索研究与发展年会论文集，页256–267。Citeseer（1980年） 12. Buckley, C., Voorhees, E.M.：利用不完整信息进行检索评估。载于：第 27 届 ACM SIGIR 国际年会信息检索研究与发展会议论文集，页25–32（2004年） 13. Buckley, C., Voorhees, E.M.：评估度量稳定性的评价。载于：《ACM SIGIR论坛》，2，页235–242。ACM，美国纽约州纽约市（2017年） 14. Busin, L., Mizzaro, S.：公理度量学：信息检索效能度量的公理化方法。载于：2013年信息检索理论会议论文集，页22–29（2013年）

15. B¨uttcher, S., Clarke, C.L., Yeung, P.C., Soboroff, I.：利用不完整且有偏的判断进行可靠的信息检索评估。载于：第30届 ACM SIGIR 国际年会信息检索研究与发展会议论文集，页63–70（2007年） 16. Carmel, D., Yom-Tov, E.：估计信息检索中的查询难度。《信息概念、检索与服务综合讲座》**2**(1)，1–89（2010年） 17. Carterette, B.：系统效能、用户模型与用户效用：用于研究的概念框架。载于：第34届 ACM SIGIR 国际会议信息检索研究与发展会议论文集，页903–912（2011年） 18. Carterette, B.A.：基于系统的信息检索实验统计分析中的多重检验。《《ACM 信息系统汇刊》（TOIS）》**30**(1)，1–34（2012年） 19. Chapelle, O., Metlzer, D., Zhang, Y., Grinspan, P.：分级相关性的期望倒数排名。载于：第18届 ACM 信息与知识管理会议论文集，页621–630（2009年） 20. Cleverdon, C.W.：克兰菲尔德测试对索引语言的重要意义。载于：第14届 ACM SIGIR 国际年会信息检索研究与发展会议论文集，页3–12（1991年） 21. Clinchant, S., Gaussier, E.：文档频率对伪相关反馈重要吗？载于：信息检索理论会议，页89–100。Springer（2011年） 22. Clinchant, S., Gaussier, E.：伪相关反馈模型的理论分析。载于：2013年信息检索理论会议论文集，页6–13（2013年） 23. Cooper, W.S.：期望搜索长度：基于检索系统弱序作用的单一检索效能度量。《美国文献学》**19**(1)，30–41（1968年） 24. Croft, W.B., Metzler, D., Strohman, T.：搜索引擎：信息检索实践，第520卷。艾迪生-韦斯利，雷丁（2010年） 25. Do Carmo, M.P.：曲线与曲面的微分几何：修订并更新的第二版。Courier Dover 出版社（2016年） 26. Fang, H.：信息检索的公理化方法。技术报告（2007年） 27. Fang, H., Tao, T., Zhai, C.：信息检索启发式方法的形式研究。载于：第27届 ACM SIGIR 国际年会信息检索研究与发展会议论文集，页49–56（2004年） 28. Fang, H., Tao, T., Zhai, C.：信息检索模型的诊断性评估。《《ACM 信息系统汇刊》（TOIS）》**29**(2)，1–42（2011年） 29. Fang, H., Zhai, C.：对信息检索公理化方法的探索。载于：第28届 ACM SIGIR 国际年会信息检索研究与发展会议论文集，页480–487（2005年） 30. Fang, H., Zhai, C.：信息检索公理化方法中的语义词项匹配。载于：第29届 ACM SIGIR 国际年会信息检索研究与发展会议论文集，页115–122（2006年） 31. Ferrante, M., Ferro, N., Fuhr, N.：迈向信息检索评估中的有意义陈述：将评估指标映射到区间尺度。《IEEE Access》**9**，136,182–136,216（2021年） 32. Ferrante, M., Ferro, N., Fuhr, N.：回应 Moffat 关于“迈向信息检索评估中的有意义陈述：将评估指标映射到区间

尺度”（2022年）。DOI 10.48550/ARXIV.2212.11735。网址 https://arxiv.org/abs/ 2212.11735 33. Ferrante, M., Ferro, N., Pontarollo, S.: 信息检索评估度量的一般理论。《IEEE 知识与数据工程汇刊》**31**(3), 409–422 (2018年) 34. Ferro, N., Peters, C.: 变化世界中的信息检索评估：CLEF二十年的经验教训，第41卷。Springer (2019年) 35. Flach, P.: 机器学习中的性能评估：优点、缺点、问题与未来方向。载于：AAAI人工智能大会论文集，01，第9808–9814页 (2019年) 36. Fraleigh, J.B.: 抽象代数入门。培生印度教育出版社 (2003年) 37. Fr´echet, M.M.: 关于泛函演算的若干问题。巴勒莫数学会报告（1884-1940）**22**(1), 1–72 (1906年) 38. Fuhr, N.: 信息检索评估中的一些常见错误及其避免方法。载于：ACM SIGIR论坛，3，第32–41页。ACM，美国纽约州纽约市 (2018年) 39. Gaudette, L., Japkowicz, N.: 序数分类的评估方法。载于：加拿大人工智能会议，第207–210页。Springer (2009年) 40. Gauss, C.F.: 曲面通论，第1卷。Typis Dieterichianis (1828年) 41. Giner, F.: 对“信息检索评估度量的一般理论”的评论。arXiv预印本 arXiv:2303.16061 (2023年) 42. Guccione, J.A.: 度量空间。布宜诺斯艾利斯大学，文本 (2018年) 43. Han, L., Roitero, K., Maddalena, E., Mizzaro, S., Demartini, G.: 关于相关性等级变换。载于：第28届ACM信息与知识管理国际会议论文集，第39–48页 (2019年) 44. Hand, D.J.: 统计学与测量理论。《英国皇家统计学会杂志：A辑（社会统计学）》**159**(3), 445–473 (1996年) 45. Harman, D.: 信息检索评估。《信息概念、检索与服务综合讲座》**3**(2), 1–119 (2011年) 46. Hauff, C., de Jong, F.: 检索系统评估：自动评估与不完整判断。载于：第33届ACM SIGIR信息检索研究与发展国际会议论文集，第863–864页 (2010年) 47. Hausdorff, F.: 集合论，第119卷。美国数学学会 (2005年) 48. Huibers, T.W.C.: 信息检索公理化理论。博士论文 (1996年) 49. Hull, D.: 在检索实验评估中使用统计检验。载于：第16届ACM SIGIR信息检索研究与发展国际会议论文集，第329–338页 (1993年) 50. Hungerford, T.W.: 代数，第73卷。施普林格科学与商业媒体 (2012年) 51. Jacobson, N.: 基础代数 I。Courier Corporation (2012年) 52. J¨arvelin, K., Kek¨al¨ainen, J.: 基于累积增益的信息检索技术评估。《ACM 信息系统汇刊》（TOIS）**20**(4), 422–446 (2002年) 53. Kando, N.: 使用多级相关性判断评估信息检索系统——关于可求平均的单值指标的讨论。IPSJ SIG会议记录**63**, 105–112 (2001年) 54. Karimzadehgan, M., Zhai, C.: 信息检索翻译语言模型的公理化分析。载于：欧洲信息检索会议，第268–280页。Springer (2012年) 55. Kazai, G.: INEX 2003指标工作组报告。载于：XML检索评估计划（INEX）：INEX 2003研讨会论文集，德国达格施图尔，2003年12月15日至17日 (2004年)

56. Kazai, G., Lalmas, M.: Inex 2005评估指标。载于：N. Fuhr, M. Lalmas, S. Malik, G. Kazai (编) XML信息检索与评估进展，第16–29页。Springer Berlin Heidelberg，柏林，海德堡 (2006年) 57. Kek¨al¨ainen, J., J¨arvelin, K.: 在信息检索评估中使用分级相关性评估。《美国信息科学与技术学会会刊》**53**(13), 1120–1129 (2002年) 58. Korfhage, R.R.: 信息存储与检索。John Wiley and Sons, Inc.，美国 (1997年) 59. Krantz, D., Luce, D., Suppes, P., Tversky, A.: 测量基础，第I卷：加法与多项式表示 (1971年) 60. Krantz, D.H.: 测量基础。第II卷：几何、阈值与概率表示 (1989年) 61. Luce, D., Krantz, D., Suppes, P., Tversky, A.: 测量基础，第III卷：表示、公理化与不变性 (1990年) 62. Maddalena, E., Mizzaro, S.: 公理度量学：信息检索效能指标的公理。载于：EVIA@ NTCIR (2014年) 63. Michell, J.: 测量尺度与统计学：范式冲突。《心理学通报》**100**(3), 398 (1986年) 64. Michell, J.: 心理测量逻辑导论。Psychology Press (2014年) 65. Moffat, A.: 有效性指标的七个数值属性。载于：亚洲信息检索研讨会，第1–12页。Springer (2013年) 66. Moffat, A.: 信息检索中的批量评估指标：测量指标、尺度与含义。IEEE Access **10**, 105,564–105,577 (2022年) 67. Moffat, A., Bailey, P., Scholer, F., Thomas, P.: 将用户期望和行为纳入搜索有效性的测量。《ACM 信息系统汇刊》（TOIS）**35**(3), 1–38 (2017年) 68. Moffat, A., Zobel, J.: 用于测量检索效能的排名偏置精确率。《ACM 信息系统汇刊》（TOIS）**27**(1), 1–27 (2008年) 69. Montazeralghaem, A., Zamani, H., Shakery, A.: 改进对数逻辑斯蒂反馈模型的公理化分析。载于：第39届ACM SIGIR信息检索研究与发展国际会议论文集，第765–768页 (2016年) 70. Pollock, S.M.: 比较信息检索系统的测量指标。《美国文献学》**19**(4), 387–397 (1968年) 71. Rahimi, R., Montazeralghaem, A., Shakery, A.: 基于语料库的跨语言信息检索公理化方法。《信息检索期刊》**23**(3), 191–215 (2020年) 72. Roberts, F.S.: 测量理论。《数学及其应用百科全书》**7** (1985年) 73. Robertson, S.: 关于gmap及其他变换。载于：第15届ACM信息与知识管理国际会议论文集，第78–83页 (2006年) 74. Robertson, S.: 关于信息检索评估的历史。《信息科学杂志》**34**(4), 439–456 (2008年) 75. Rocchio, J.: 文档检索系统的性能指标。《信息存储与检索》第83页 (1964年) 76. Rosset, C., Mitra, B., Xiong, C., Craswell, N., Song, X., Tiwary, S.: 神经排序模型正则化的公理化方法。载于：第42届ACM SIGIR信息检索研究与发展国际会议论文集，第981–984页 (2019年)

77. Sagara, Y.：排序输出检索系统的性能测量指标。《日本信息与知识学会期刊》**12**(2), 22–36 (2002年) 78. Sakai, T.：基于多级相关性的新性能指标：其在问答中的应用。载于：NTCIR (2004年) 79. Sakai, T.：度量、统计学与检验。载于：PROMISE冬季学校，第116–163页。Springer (2013年) 80. Sakai, T.：信息检索中的统计学改革？载于：《ACM SIGIR论坛》，第48卷，第3–12页。ACM，美国纽约州纽约市 (2014年) 81. Sakai, T.：论Fuhr关于信息检索评估的准则。载于：《ACM SIGIR论坛》，第54卷，第1–8页。ACM，美国纽约州纽约市 (2021年) 82. Sakai, T., Kando, N.：关于为不完整相关性评估而设计的信息检索度量指标。《信息检索》**11**(5), 447–470 (2008年) 83. Sakai, T., Oard, D.W., Kando, N.：《信息检索与访问任务评估：NTCIR的研究影响遗产》。Springer Nature (2021年) 84. Salton, G.：《自动信息组织与检索》(1968年) 85. Salton, G., McGill, M.J.：《现代信息检索导论》。mcgraw- hill (1983年) 86. Sanderson, M., et al.：基于测试文档集的信息检索系统评估。《信息检索基础与趋势》**4**(4), 247–375 (2010年) 87. Savoy, J.：检索效能评估中的统计推断。《信息处理与管理》**33**(4), 495–512 (1997年) 88. Sebastiani, F.：一种公理推导的分类算法评价度量指标。载于：2015年信息检索理论国际会议论文集，第11–20页 (2015年) 89. Sirotkin, P.：论搜索引擎评估指标。arXiv预印本 arXiv:1302.2318 (2013年) 90. Stevens, S.S.：数学、测量与心理物理学。Wiley (1951年) 91. Stevens, S.S., et al.：论测量尺度理论。Bobbs-Merrill, College Division (1946年) 92. Swets, J.A.：信息检索系统。《科学》**141**(3577), 245–250 (1963年) 93. Urbano, J., Lima, H., Hanjalic, A.：信息检索中的统计显著性检验：对第一类、第二类和第三类错误的实证分析。载于：第42届ACM SIGIR信息检索研究与发展国际会议论文集，第505–514页 (2019年) 94. Van Rijsbergen, C.J.：评价基础。《文献学杂志》**30**(4), 365–373 (1974年) 95. Vanbelle, S., Albert, A.：关于适用于序数尺度的线性加权κ系数的说明。《统计方法学》**6**(2), 157–163 (2009年) 96. Velleman, P.F., Wilkinson, L.：名义、序数、区间和比率分类具有误导性。《美国统计学家》**47**(1), 65–72 (1993年) 97. Voorhees, E.M.：TREC 2005稳健检索赛道。载于：《ACM SIGIR论坛》，第40卷，第41–48页。ACM，美国纽约州纽约市 (2006年) 98. Voorhees, E.M., Harman, D.K., et al.：TREC：信息检索实验与评价，第63卷。Citeseer (2005年) 99. Voorhees, E.M., et al.：TREC 2003稳健检索赛道概述。载于：TREC，第69–77页 (2003年) 100. Wicaksono, A.F., Moffat, A.：度量指标、用户模型与满意度。载于：第13届网络搜索与数据挖掘国际会议论文集，第654–662页 (2020年)

101. Zhang, F., Liu, Y., Li, X., Zhang, M., Xu, Y., Ma, S.：基于宝石消除游戏玩家模型的网络搜索评估。载于：第40届国际 ACM SIGIR 信息检索研究与发展大会论文集，第425–434页 (2017年)

A 附录

A.1 形式化证明

证明。 [**命题 1**]:对称性已验证，因为 $d_f(\hat{\mathbf{r}}_1, \hat{\mathbf{r}}_2) = |f(\hat{\mathbf{r}}_1) - f(\hat{\mathbf{r}}_2)| = |f(\hat{\mathbf{r}}_2) - f(\hat{\mathbf{r}}_1)| = d_f(\hat{\mathbf{r}}_2, \hat{\mathbf{r}}_1)$ is trivial。三角不等式也显而易见，只需考虑实数上的三角不等式：$|f(\hat{\mathbf{r}}_1) - f(\hat{\mathbf{r}}_2)| \leq |f(\hat{\mathbf{r}}_1) - f(\hat{\mathbf{r}}_3)| + |f(\hat{\mathbf{r}}_3) - f(\hat{\mathbf{r}}_2)|$。□

证明。[**命题 2**]:

关于度量空间的一个有趣结果 [42] 如下所述：“*Let* $(\mathbf{R}_2, d_2)$ *be a metric space and let* $f: \mathbf{R}_1 \longrightarrow \mathbf{R}_2$ *an an injective or one-to-one function, then* $(\mathbf{R}_1, d_1)$ *is a metric space, where* $d_1(\hat{\mathbf{r}}_1, \hat{\mathbf{r}}_2) = d_2(f(\hat{\mathbf{r}}_1) f(\hat{\mathbf{r}}_2)), \forall \hat{\mathbf{r}}_1, \hat{\mathbf{r}}_2 \in \mathbf{R}_1$”。

在检索场景中，$(\mathbf{R}_2, d_2) = (\mathbb{R}, |\cdot|)$是赋予通常范数（绝对值）的实数轴的度量空间。设 f为一个单射的信息检索评价度量；根据前述结果可知，$(\mathbf{R}_1, d_1) = (\mathbf{R}\ d_f)$是一个度量空间，即 d_f满足度量的三个公理。□

证明。 [**命题 3**]:

下面将说明从右向左的蕴含关系。考虑一个度量序数尺度，f，其中所取得的值等距分布。

如果 $[\hat{\mathbf{r}}_1, \hat{\mathbf{r}}_2] = \{\hat{\mathbf{r}}_1, \hat{\mathbf{r}}_2\}$，则称一个区间为 $prime$。首先将证明，函数 $F(\mathbf{x}, \mathbf{y}) = |f(\mathbf{x}) - f(\mathbf{y})|$在任意素区间上取得其最小值。

设 $[\hat{\mathbf{r}}_1, \hat{\mathbf{r}}_3] = \{\hat{\mathbf{r}}_1, \hat{\mathbf{r}}_2, \hat{\mathbf{r}}_3\}$为一个非素区间，其中 $\hat{\mathbf{r}}_1 \preceq_f \hat{\mathbf{r}}_2 \preceq_f \hat{\mathbf{r}}_3$，，则由于 f是一个顺序尺度，有 $f(\hat{\mathbf{r}}_1) \leq f(\hat{\mathbf{r}}_2) \leq f(\hat{\mathbf{r}}_3)$。这意味着 $F(\hat{\mathbf{r}}_1, \hat{\mathbf{r}}_3) = |f(\hat{\mathbf{r}}_3) - f(\hat{\mathbf{r}}_1)| \geq |f(\hat{\mathbf{r}}_3) - f(\hat{\mathbf{r}}_2)|$，即，的最小值不会在$[\hat{\mathbf{r}}_1, \hat{\mathbf{r}}_3]$处取得。此外，对于每个素区间，函数 F取相同的值。给定一个素区间，$[\hat{\mathbf{r}}_1, \hat{\mathbf{r}}_2]$,可以将其视为其连续素区间之一，$[\hat{\mathbf{r}}_2, \hat{\mathbf{r}}_3]$,，因为 $\preceq_f$是一个弱序（每一对元素都具有可比性）。这两个素区间验证了 $f(\hat{\mathbf{r}}_1) < f(\hat{\mathbf{r}}_2) < f(\hat{\mathbf{r}}_3)$，因为 f是一个度量，且 f所取得的值等距分布。因此，可以假设，对于任意素区间 $[\hat{\mathbf{r}}_1, \hat{\mathbf{r}}_2]$，$F(\hat{\mathbf{r}}_k \in \mathbb{R}^+$。

现在可以看出，等间距区间（不一定是素区间）被赋予相等的差值。考虑任意一个非素区间，$[\hat{\mathbf{r}}_1, \hat{\mathbf{r}}_m] = \{\hat{\mathbf{r}}_1, \hat{\mathbf{r}}2, \ldots, \hat{\mathbf{r}}_m\}$。由于 f 是一种度量，因此对不同元素取不同的值。因此，可以假设 $f(\hat{\mathbf{r}}_1) < f(\hat{\mathbf{r}}_2) < \cdots < f(\hat{\mathbf{r}}_{m-1}) < f(\hat{\mathbf{r}}_m)$。那么，对于 $i = 1, \ldots m-1$，每个区间 $[\hat{\mathbf{r}}_i, \hat{\mathbf{r}}_{i+1}]$ 都是素区间，因为 F 在这些区间上取到最小值。由于 $f(\hat{\mathbf{r}}_m) - f(\hat{\mathbf{r}}_1) = f(\hat{\mathbf{r}}_m) - f(\hat{\mathbf{r}}_{m-1}) + f(\hat{\mathbf{r}}_{m-1}) - \cdots - f(\hat{\mathbf{r}}_2) + f(\hat{\mathbf{r}}_2) - f(\hat{\mathbf{r}}_1)$，且对于 $1 \leq i \leq m-1$ 有 $f(\hat{\mathbf{r}}_{i+1}) - f(\hat{\mathbf{r}}_i) = k$，因此 $f(\hat{\mathbf{r}}_1) - f(\hat{\mathbf{r}}_m) = k \cdot m$，它只取决于区间的跨度 m，而不取决于所考虑的元素。因此，等间距区间被赋予相等的差值，即 f 是一个区间尺度。

最后，我们将看到另一个蕴含关系。考虑$[\hat{\mathbf{r}}_1, \hat{\mathbf{r}}_2]$,的任意一个素区间，其中$\mathbf{R}$是一个区间尺度，那么等间距区间会被赋予相等的差值，也就是说，对于$\mathbf{R}$的每个素区间，值 $|f(\hat{\mathbf{r}}_2) - f(\hat{\mathbf{r}}_1)|$ 都是常数。此外，该值应为严格正值。为证明所得值

等间距，只需验证$\mathbf{R}$ 中的不同元素被赋予不同的 f值；这一点成立，因为 f 是一个度量。□