信息检索评估的整合任务和用户特征的蓝图：测试集与评价指标

Kal Jarvelin 和 Eero Sormunen，坦佩雷大学信息技术与传播科学学院 | 传播科学，坦佩雷大学，FI-33014 坦佩雷大学，芬兰 ORCID：https://orcid.org/0000-0001-7655-8930（Jarvelin）

摘要：信息检索中传统的搜索结果评估指标，如平均准确率均值和归一化折损累计增益，往往简单地关注文档与搜索主题之间的主题相关性，并假定这种关系是单维的，且通常是二元的。它们忽略文档内容重叠，并简单地假定随着搜索者越来越深入地浏览排序列表，增益会不断累积。搜索者仍然被置于幕后。本蓝图提出一种新颖的测试集和搜索结果评估指标，其基础是多维、非二元的相关性评估，对文档重叠进行显式建模，并纳入除相关性之外影响搜索者文档可用性的属性。我们认为，特定情境下文档与搜索任务的相关性由一个或多个内容主题，以及零个或多个文档可用性属性构成。给定文档可能在不同程度上与这些内容主题相关。不同文档在内容主题方面也可能存在程度不一的重叠。对于特定搜索者及其驱动任务（例如学校学习任务或科学研究项目），文档可读性、可信度或语言等属性，体现了文档在搜索任务情境中的整体可用性。所提出的指标通过考量排序搜索结果中每篇后续文档的贡献、各主题之间估算的重叠，以及基于其属性的可用性，来评估结果质量。内容主题的数量和类型取决于具体搜索任务，因此在评估研究的不同搜索任务之间可能有所不同。属性的数量和类型则在任务情境中固定，并且在不同搜索任务之间可能保持稳定。

1. 引言

相关性是信息科学的一个基本概念。它易于直观理解，却难以精确定义。尽管如此，人们普遍认为，相关性是信息需求与信息对象之间一种多层次、多维度的关系（Borlund，2003年；Ingwersen 和 Jarvelin，2005年；Saracevic，2007）；Vakkari 和 Hakala，2000年）。信息对象（下文简称文档）可以基于工作任务情境，以不同程度满足信息需求的各个维度（下文简称主题）。

在信息检索（IR）中，相关性在评估检索方法和系统时发挥着重要作用（Croft 等，2010）。直到上个千年末，最先进水平（SOTA）的实验性信息检索

评估采用二元单维相关性判断：文档要么完全符合某一信息需求（或主题），要么完全不相关，因此忽略了文档相关性的多层次特征和多维特征。2000年前后，Sormunen（2002）构建了首个基于分级（即多层次）相关性判断的大型测试集，Jarvelin & Kekäläinen（2000；2002；Kekäläinen & Jarvelin，2002）则提出了基于分级评估的信息检索评估适用指标。Keskustalo 等人（2008）和 Ferro 等人（2016）提出了将检索者遇到不相关文档时的不悦纳入考量的指标。以上研究都没有涉及相关性的多维性。本文关注的问题之一是："在信息检索评估中，如何处理多维且分级的相关性评估？"

搜索算法会生成文档排序列表。排序中相邻的文档可能提供相似内容，无法在前面文档已经提供给用户的信息基础上增加新信息。此外，评价指标忽略了文档内容重叠，并想当然地认为，随着搜索者浏览排序列表的范围扩大，增益会不断累积。因此，本文的另一关注重点是："在信息检索评估中，如何处理文档内容重叠（重复识别之外）？"

对信息需求者而言，文档的可用性还取决于相关性之外的其他因素。一篇完全相关（主题相关——参见 Ingwersen & Jarvelin，2005年）的文档，可能因其他原因而毫无用处，例如其语言、年代、偏好，或假定读者具备的专业水平或世界观。我们将这些因素称为文档可用性属性。除去重叠的部分，相邻文档可能会增加每个搜索任务主题所累积的内容，但属性并不具备这种跨文档累积特性。每种属性都构成一个独立维度。因此，本文的第三个关注重点是："在信息检索评估中，如何处理可用性属性，并将其与多维相关性评估相结合？"

本文的第四个重点是探讨这样一个问题："如何定义一种形式化模型，在一个连贯统一的框架中处理多维分级相关性评估、文档重叠和文档可用性属性，以服务于信息检索评估？此外，我们还希望能够将最先进的信息检索评价模型作为所提出形式化模型的特例推导出来。"

Newell（1969）提出，定义一种方法需要包含三个组件：方法的领域 D、算法 A 和论证 J。D 规定了该方法适用的任务类型。算法 A 规定了该方法各步骤的输入、处理和输出。最后，论证 J 包含说服该方法使用者的论据，使其相信该方法值得信赖，能够发挥作用并得出有效且可靠的结果。提出一种新方法，就意味着并要求指出其在任务、算法和／或论证方面的新颖性。比较旧方法和拟议的新方法时，可以采用一些标准，例如简洁性、准确性、范围、系统化能力、解释力、成果丰硕性和有效性，正如 Järvelin 和 Wilson（2003年）所讨论的那样。由于指标／测试集也被称为方法，我们在定义拟议的新指标时遵循 Newell 的提议。

众所周知，当一项活动随着时间推移而经历反复评估时，被衡量的特质往往会成为评估的目的，并决定发展的重点。我们将此称为设计中的 WYSIWYG（所见即所得）原则。其推论在这里同样适用，即 WYDSIWYDU （所不见即所不懂）。因此，人们应当搁置容易完成的评估任务及其浅显的有效性指标，转而接受更具挑战性的评估任务及相应指标——即使这能证明某人的 搜索算法无效，且其方法狭隘、准备不足。

我们以 Jarvelin & Kekäläinen（2002年）提出的累积增益（CG）系列信息检索评价指标为基础，构建了称为 MDCU 的新指标，即多维累积效用指标。

2. MDCU 测试集框架的任务

2.1 目标与要求

实验性信息检索传统上以测试集为基础（Ingwersen & Järvelin，2005年；Sanderson，2010）。测试集由一个通常规模较大的文档集合、一组搜索任务（或主题），以及一个相关性语料库（通常称为相关性判断集）组成；相关性语料库指明哪些文档与哪些主题相关。相关性是单维的，通常为二元（Harman & Voorhees，2006），有时采用分级标注（Sormunen，2002）。标准的搜索任务表示包括一个简短标题 T、一段用几句话描述信息需求的文字 D，以及一段用于明确相关性标准的叙述 N。对比测试不同的搜索机制，评估其能否针对每项搜索任务找到相关文档，并按相关性从高到低对其排序。

这种广泛使用的方法存在若干局限：对相关性的处理过于简单（单维且二元）；将主题相关性作为唯一依据（忽视基于任务和情境的相关性）；假定文档在相关性方面彼此独立（对重复信息给予全额奖励）；假定用户与搜索结果的交互方式不影响价值（只要相关性相同，排名靠后的文档就与排名靠前的文档同样有价值）。从以用户为中心的信息检索研究的角度来看，这些局限受到了批评（例如 Froelich，1994年；Schamber，1994年；Harter & Hert，1997年；Borlund & Ingwersen，1998年）。在许多实证研究中，真实搜索者在评估检索出的文档时，采用了广泛且不断变化的主观相关性标准（例如 Barry，1994年；Barry & Schamber，1998年；Greisdorf，2003年；Kekäläinen & Järvelin，2002年；Maghlaughlin & Sonnenwald，2002年；Rieh，2002年；Spink & Greisdorf，1998年）。要克服这些局限，信息检索评价框架必须明确呈现（且最好允许操控）评估中搜索行为背后的任务、搜索者及信息交互情境特征。此外，还需要采用能够考虑这些特征的评价指标。

Belkin（2016）提出，应将评估信息检索系统性能的主要标准从相关性转向有用性。这意味着，应评估信息检索系统帮助人们实现驱动性任务/目标的能力。这需要建立一个基于实证的此类任务/目标分类体系。Byström & Järvelin（1995）、Kumpulainen & Järvelin（2010）、Li & Belkin（2008）、Saastamoinen

& Järvelin（2017）以及 Vakkari（2016）为构建此类分类体系并推断有助于搜索的需求提供了思路。然而，这些研究并未提出用于检索评估的表示或指标。如果希望实现新的或更广泛的评估场景，就需要开发新方法（Kekäläinen & Järvelin，2002）。 我们提出的方案是迈向 Belkin 的有用性的一步，涵盖搜索评估，但不涉及信息交互的其他方面（见 Järvelin & al.，2015）

模拟工作任务情境（Borlund，2016；Ingwersen 和 Jarvelin，2005）旨在将基于任务的因素和情境方面纳入交互式信息检索（IIR）的评估中。这种方法已被广泛用于提升交互式信息检索评估的真实感，主要影响了实验设计和交互式搜索者行为。这种方法影响了搜索任务的表示方式，并旨在引导测试对象的行为。然而，关于如何在搜索结果评估中处理多维相关性、文档重叠和属性的表示方式，并没有提出新的方案。

有许多指标可用于衡量单个搜索任务及多个搜索任务的搜索结果质量。传统指标包括召回率（R）、精确率（P）、截断精确率、平均准确率均值（MAP），参见（Sanderson，2010）。这些指标适用于单维相关性情境中的二元相关性。也存在一些推广指标，例如广义召回率 gR、精确率 gP 和广义平均准确率均值（gMAP），参见（Kekäläinen & Järvelin，2002；Robertson 等，2010）；还有一些“天生分级”的指标，例如基于累积增益（CG）的指标（Jarvelin & Kekäläinen，2002），但这些指标仍然针对单维相关性，无法处理文档重叠和可用性。TREC 交互式轨道的一项贡献是提出了实例相关性的概念，旨在解决交互式用户测试中检索到重叠信息的问题（Over，2001）。这种方法的一个局限是，搜索任务必须聚焦于独特的“可计数实例”，而这类任务并不典型。

INEX 检索实验采用了分级主题评估，并使用过滤器指出应当找到所需答案的 XML 结构元素（INEX，2014年；Stanford，2022）。评价指标根据每个检索到的答案元素与搜索任务的主题匹配程度及其结构精确性进行评分。可以认为，通过应用适当的结构过滤器，INEX 评价指标在一定程度上支持了多维主题相关性和属性的评估。然而，针对结构的评分标准仅考虑了每个搜索结果包含答案的完整程度，并没有明确的方法将多个独立属性整合到评估中。Fuhr 和 Roelleke（1997年）提出了概率关系模型，作为标准关系模型的一种泛化。该模型为文本内容和不确定的数据库属性值上的概率推理提供了一种定义明确的机制。这可以被视为一种评分机制，尤其适用于多维属性，但它并不直接支持对多个主题和文档重叠的处理。

用于累积效用计算的文档重叠判定是一个棘手的问题。传统的单维度、一次处理一个文档的相关性评分忽略了重叠问题。它不会鼓励将数量最少、但整体上能为搜索任务提供尽可能全面回答的文档排序到搜索结果的顶端

。相反，它会奖励检索模型用冗余答案填满结果。最大化评价指标值取代了原本更为根本的目标。一些研究分析了来自不同来源、由不同类型的查询或文档表示产生的搜索结果之间的重叠（Katzer 等，2017；Saracevic 等，1988）。这些研究关注的是通过不同方式检索到同一文档，而我们关注的是：与其他文档（集合）相比，不同文档如何围绕相同主题贡献互补内容。聚类和文本分类研究（Mirończuk & Protasiewicz，2018）关注文档之间的文本相似性，而当前的挑战是找到主题相似、但在文本上尽可能不同的文档，直到所有主题的主题效用最大化。另一个问题是，应当能够评估所检索文档的任何排列顺序。现代测试集的规模不允许在这种情况下进行人工评估。

文档的可用性方面在信息检索文献中受到的关注不如相关性。我们不了解任何系统性方法来处理可用性属性，包括属性的选择、测量或证据的计算。通常，属性选择基于针对文档元数据字段的临时标准，测量值取自文档元数据，证据则通过布尔函数组合。关于如何利用搜索环境中其他可用数据的讨论很少。一项值得关注的提议是“信息营养标签”（Fuhr 等，2017），可用于描述文档可用性。此类标签还可以提供文档的语言、体裁、事实性、传播性、观点、争议性、权威性、技术性和主题相关性等信息。有些属性可能是通用的、可自动计算的，并适用于所有搜索任务；另一些则取决于特定的任务与搜索者组合。

表1 co　将传统信息检索评估与本文提出的多维方法进行比较　r.

所提框架的任务是在信息检索任务中，当文档相关性具有多维性、各主题中文档的贡献可能重叠，且文档可用性属性会影响评分时，对搜索者的搜索结果进行评分。因此，该框架必须：

- 提供文档表示，用于评估 (a) 每个文档对测试集中的多维主题信息库的贡献，以及 (b) 可用性属性。在一个文档内，主题和属性的评分彼此独立。- 可计算，从而能够根据其主题相关性和可用性属性对文档效用进行总体评分。评分必须考虑每个文档在搜索结果中的位置：如果文档排名较低，或前面的文档已经针对某一特定主题贡献了大量信息，则应降低其分数。文档之间的重叠评估必须避免对可能出现在搜索结果中的所有文档序列进行智力评估；在内容主题内部及跨内容主题之间，重叠可能是部分的。- 对良好的属性值（文档在搜索任务情境中的可用性）给予奖励；对不适当的属性值降低评分。

- 支持针对任务、搜索者和情境特征开展系统化实验。 s.

表1. 传统方法与所提出的 MDCU 评估框架比较。

任务特征	信息检索评估任务的方法	
	传统方法	**MDCU**
相关性等级	二元	分级
相关性维度	一个（主题相关）	多个基于任务的主题， 多个属性
相关性判断、成本	人工，€€	人工 ，€€€
相关性依据	“至少有一个相关 文档文本中的句子”	“相关文本的数量 与相关程度相关 每个主题的可用性属性 可能会被观察到/计算得出。”
相关性特征：内容 / 结构 / 属性	是 / 否 / 否	是 / 否 /是
结果分数范围	[0, ∞)，[0, 1], %	[0, ∞)，[0, 1], %
重叠处理	不支持	多维评分提供 一些可能性
主题设计成本	€	€€€

2.2 基本文档表示

我们为 MDCU 框架提出一种由三个部分构成的文档表示：内容（包括元数据）、主题相关性列表和可用性属性列表。因此，文档列表（例如搜索结果、相关性语料库或相关性反馈列表）可以表格形式呈现，见表2。表中展示了一个抽象的六文档列表 d1 – d6，其中省略了内容部分。对于虚构的搜索任务，文档相关性通过 4 个主题（主题1 – 主题4）表示，取值为从 0（不相关）到 3（高度相关）的整数；另有 3 个虚构属性 Attr1 – Attr3，其可用性取值为 0 到 1 之间的实数。内容主题由（搜索）任务决定，相关性得分则由主题和文档决定。此处我们采用 Sormunen（2002）的四级相关性模型，因为该模型广受欢迎，但也可以使用其他任何分级方式。可用性属性由搜索者和搜索任务决定，可用性得分则由属性和文档决定。原则上，即使是高度相关的文档，其可用性也可能为零。

假设我们已经通过这些主题和属性，以某种方式构建了任务和搜索者特征的表示，并且进一步获得了描述性数值——这一点我们稍后再讨论。现在我们可以着重探讨 MDCU 框架在获得这些数值后如何运作。首先，我们将考察该框架的形式特征（即与内容无关的特征）： a) 如何表示各个方面，

b) 各变量如何相互作用以综合证据并进行评分，以及 c) 该框架在抽象情境下的表现。

表2。用于表示多维相关性量值的简单模型，取值范围为 {0, 1, 2, 3}，并包含取值范围为 [0,1] 的属性可用性分数。

	主题相关度					可用性属性评分			
文档编号	主题1	主题2	主题3	主题4	…	属性1	属性2	属性3	…
d1	0	1	3	2	…	1,0	1,0	1,0	…
d2	2	0	0	2	…	0,9	0,7	0,9	…
d3	1	0	2	0	…	1,0	0,9	1,0	…
d4	0	0	3	1	…	0,8	0,9	0,7	…
d5	1	2	0	2	…	1,0	1,0	1,0	…
d6	0	0	0	2	…	1,0	0,8	1,0	…
…	…	…	…	…	…	…	…	…	…

2.3 框架的组件

测试集包含若干组件，下面将逐一介绍，并突出说明 MDCU 框架所需的特殊特性。相关内容将在下文的插文中介绍。

组件	描述
文档集合 DC	文档包含文本（或媒体）部分，以及 0 至 n 个元数据字段， 这些字段的值可能是显式存储的，也可能是通过函数生成的，例如， lang(doc) = “法语”。
搜索任务 ST	搜索任务采用传统方法，以 TDN（Tile – Description – Narrative）结构表示。类型一搜索任务 TDN-1 均为纯文字形式， 包括对属性的要求，例如“法语学术文章”，而 TDN-2 任务有一种正式表达此类条件的方式，例如 type(doc) = “学术” & lang(doc) = “法语”。TDN-1 任务为查询构建过程和结果评分带来了更多不确定性 用于查询构建过程和结果评分。 N-组件包含对搜索任务中每个主题和属性的描述， 并解释每个等级的相关性标准和要求 相关性。如果收集具有语义指示性的词汇， 将其用于每个主题会很有帮助——这些词汇可以整理成一个本体（Jarvelin 等，2001）。
相关性语料库 RC	RC 为针对某个搜索任务评估过的所有文档提供一种表示形式，如 表2所示。DC 的其余部分则隐含地视为不相关或不可用。 对文档中的主题分别进行数值评估；主题评分 可能具有不同的取值范围；此处不失一般性地采用整数值 {0, 1, 2, 3}，从不相关到高度相关；请注意，原则上 评估尺度可能因主题而异。给定主题的内容可能 在不同文档之间重叠。 文档属性分别独立地进行数值评估。属性 评分可能具有不同的取值范围；此处不失一般性地假设为实数

	值[0, 1]。给定文档的属性之间会相互作用，而 给定属性的值不会在不同文档之间相互作用。例如，如果 ST 包含要求 type(doc) = “scholarly” & lang(doc) = “French”， 找到一个满足该要求的文档，例如 d_4，，不会影响对 后续文档 d_5 等的属性进行评分。
对搜索结果进行评分	文档之间主题的重叠以形式化方式处理：如果两篇文档 与同一主题相关，则通过形式化方法估算它们预期的主题重叠 使用参数函数，从而避免对重叠进行智力分析（即分析每一种 可能的文档组合）。如果两篇文档与同一主题相关， 分数分别为 s1 和 s2，则它们的联合得分介于 max(s1, s2) 和 s1+s2 之间， 具体取决于重叠程度。s2 相对于 s1 的贡献会按一定比例折减， 对数因子。 按主题计算的累积增益：相应地，如果 CTR_j 是主题相关性 在结果排名中截至第 j 位累积的总量，而 sj1 是 下一个文档的主题相关性，则它们的联合得分 js 的范围为 $\max(CTR_j, sj1) \le js \le CTR_j$ + sj1。即使 CTR 很高，一个与新主题相关的文档也会对 整体主题增益得分有所贡献。 文档的总效用得分是其折扣主题相关性 得分经其可用性属性因子缩减后的总和。可用性属性 通过文档的总体属性可用性，统一降低文档的主题相关性得分； 总体属性可用性计算为其属性可用性因子的乘积。 相关性得分与此值相乘。 排名为 j 的排名搜索结果的累积效用，是文档 $d_1, \ldots, d_j$ 效用分数。然而，每个文档 $d_2, \ldots, d_j$ 在扣除重叠部分后，其贡献可能低于单独计算时的效用分数， 。 理想排名基于相关性语料库，其中首先是 总效用最大的文档，其次（依次）是 对已累积的相关性贡献最多的文档——扣除重叠部分后。 归一化累积效用通过将实际值除以 每个排名对应的理想值而得。其在任意排名上的值都在范围 [0, 1] 内。

4. 正式定义

4.1. 文档表示与重叠

现在，我们将正式定义一些符号约定和表示。

记号约定 1：设 $e_1, e_2, \ldots, e_n$ 是一些元素，即整数、字符串或由它们构成的结构。序列 $E = <e_1, e_2, \ldots, e_n>$ 是由这些元素组成的元组。E[i] 表示其第 i 个元素；运算符 <head || tail> 将序列拆分为首元素和尾序列。例如，$E = <e_1 \| E'>$ 表示首元素为= e_1，，尾序列为 $E' = <e_2, e_3, \ldots, e_n>$

记号约定 2：设 $e_1, e_2, \ldots$、e_n 为任意元素，即整数、字符串或构造体，元组通过 $t = <e_1, e_2>$ 连接，更一般地，$\times_{i=1,\ldots,n} e_i = <\ldots<e_1, e_2,>\ldots,e_n> = <e_1, e_2,\ldots,e_n>$。长度为 n 的空元组通过 null-struct(n) = $<e_1, e_2, \ldots, e_n>$构造，其中所有 $e_i = 0$。我们扩展集合

成员运算符‘∈’扩展为元组成员关系‘∈*’。给定元组 $t = <e_1, e_2, ..., e_n>$，表达式 $e \in^* t$ 为真，当且仅当 $\exists\, i : t[i] = e$。t 的长度为 $length(i) = n$。

记法约定 3：令 $e_1, e_2, ...、e_n$ 为任意被视为原子的元素，结构由 $s= (e_1, e_2)$ 构造，更一般地由 $(...(e_1, e_2), ..., e_n) = (e_1, e_2, ..., e_n)$ 构造。

记法约定 4：令 $E = <e_1, e_2, ...、e_n>$ 和 $F = <f_1, f_2, ...、f_n>$ 为任意数值元组。它们的逐项求和为 $E \oplus F = <e_1+f_1, e_2+f_2, ...、e_n+f_n>$。同样，乘积记为 $E \otimes F$，除法记为 $E \oslash F$。

记法约定 5：令 $dc_1, dc_2, ...、dc_n$ 为任意文档，其内容和结构在此不作分析（即视为原子的）。文档集合是集合 $DC = \{ dc_1, dc_2, ...、dc_n\}$。

定义 1：文档表示是一个三元组 $d = (dc, TRel, Attr) = (dc, <r_1, r_2, ...、r_n>, <a_1, a_2, ...、a_m>)$，其中 dc 是文档的内容（在此不作分析），$TRel = <r_1, r_2, ...、r_n>$ 是搜索任务的主题相关性得分元组，该搜索任务由 dc 中的主题 1 ... n 组成；$Attr = <a_1, a_2, ...、a_m>$ 是搜索任务的属性可用性因子元组，该搜索任务由 dc 中的可用性因子 1 ... m 组成。

定义 2：设 $d = (dc, TRel, Attr)$ 为一种文档表示。我们分别用 $cont(d)= dc$、$trels(d) = TRel$ 和 $attrs(d) = Attr$ 表示 d 的组件。我们也允许使用简写记法 $d[1] = cont(d)$、$d[2] = trels(d)$ 和 $d[3] = attrs(d)$。

定义 3：对于给定的搜索任务 ST，其相关性语料库 RC 是一组文档表示 $RC_{ST} =\{d_1, d_2, ..., d_k\}$，对应于已针对 ST 进行明确评估的 k 个文档 $d_{1...k}$。对于所有 $d \in RC_{ST}$，条件 $cont(d) \in DC$ 均成立，即 RC 中表示的文档属于文档集合 DC。对于任何未评估的文档 $dc \in DC$ 但 $\neg\exists\, d \in RC$：$cont(d) = dc$，当 ST 具有 n 个主题和 m 个可用性因子时，该文档表示为元组 $d =(dc, null\text{-}struct(n), null\text{-}struct(m))$。

定义 4：给定查询 Q 和文档集合 DC，搜索结果是一个文档表示元组 $SR = <d_1, d_2, ..., d_k>$。检索函数 $retrieve(Q, DC)$ 的定义不作限定。满足以下条件：$\cup_{d \in SR} \{cont(d)\} \subseteq DC$。

定义 5：对于 n 个主题，累积相关性表示为 $crr = <r_1, r_2, ..., r_n>$，其中 $r_i \in \Re$。评估搜索结果质量时，crr 会针对搜索任务中的每个主题，逐排名累积增益。最初，在评估第一个文档之前，crr = 空结构 (n)。

定义 6：给定文档表示 $d = (dc, <r_1, r_2, ..., r_n>, <a_1, a_2, ..., a_m>)$，所有 $a_i \in [0, 1]$，文档可用性因子是各个可用性属性因子的乘积，定义为 $attr\text{-}fact(d) = \prod_{a \in^* attrs(d)} a$。

定义 7：给定累积相关性表示 *crr* = <r_1, r_2, ..., r_n>，*b*是用于控制重叠的一个小整数，作为对数底数；给定文档表示 d = (dc, drel, dattr) = (dc, <r_1, r_2, ...,r_n>， <a_1, a_2, ..., a_m>)，它们的组合相关性为：

$$contrib(crr, \text{drel}, \text{dattr}) = \times_{i=1,\ldots,n} (crr[i] + v * \text{drel}[i] / max(1, log_b\ crr[i]))$$

这会根据主题 i 的累积相关性质量的对数（最小值为 1），以及文档可用性因子 v = attr-fact(d)，折减 d 在主题 i（i = 1 ... n）上的贡献。结果是一个新的累积相关性表示 crr'，其组件 r'_i 通过相应 d 组件的贡献递增。请注意，只有当累积相关性质量超过对数底数 b 时，折减才会开始。将 b 设得较小（≤ 1.5）即可模拟较高的重叠度。例如，考虑以下包含四个主题和三个属性的文档表示 d_1 和 d_2，以及 crr = <0, 0, 0, 0>。各主题的相关性得分分别为 0 ... 3，重叠参数 b = 2，属性可用性因子的范围为 [0 ,1]。

d_1 = (dcont1, <0, 1, 3, 2>, <1, 1, 1>)　　whence v_1 = *attr-fact*(d_1) = 1

d_2 = (dcont2, <2, 0, 0, 2>, <0.9, 0.7, 0.9>)　　whence v_2 = *attr-fact*(d_2) = 0.57

Now　　*contrib*(*crr*, d_1[2], d_1[3]) = *crr'* = *contrib*(<0, 0, 0, 0>, <0, 1, 3, 2>, <1, 1, 1>) = <0, 1, 3, 2>,

这就是 d_1 的相关性评分——文档不会与自身重叠，而 v_1 = 1

but　　*contrib*(*crr'*, d_2[2], d_2[3]) = *crr''* = <1.14, 1.0, 3.0, 3.13>,

其中 crr''[1] = 1.14 归因于 v_2, < 1，而 crr''[2] 和 crr''[3] 保留了它们在*crr'*.中已获得的分数。最后，crr''[4] = 3.13 估计主题四的重叠程度足以使 d_1[2] [4] + d_2[2][4] 减少近一个百分点，从 4 降至 3.13。

定义 8：给定搜索结果 SR，其按排名累积的相关性为

cum-rel(crr, SR, OSR) =OSR 若 SR = < >cum-rel(crr ⊕ ddr, SR', <OSR, < d [1], ddr, d[3]>>)，否则其中 SR = <d || SR'> 且 ddr = contrib(crr, d[2], d[3]),

在第一次 调用 cum-rel 时，操作数 crr = null-struct(n)，其中 n = length(d[2][1])，且 OSR = <>。

表3展示了基于表2中的数据进行的主要计算，并按从上到下的顺序将这些数据视为排名搜索结果。第2列显示独立的总体文档得分（来自表2）。主题下方的各列报告了每篇文档在各主题上的相关性，并考虑了按给定顺序访问文档时估算的重叠（b = 2）。第7列和第8列分别给出了每篇文档在各主题上的贡献以及可用性因子。最后两列显示了应用可用性

因子后的文档得分，以及搜索引擎结果页面的累积相关性得分。 请注意，累积相关性(空结构(4)，搜索引擎结果页面，<>)= 搜索引擎结果页面-主题相关性 = <3.63, 3.00, 5.696, 6.242> 以及$\sum_i$ 搜索引擎结果页面-主题相关性[i] = 18。57 [注意：此分数不包含可用性折扣]。从这一评分可以看出，搜索引擎结果页面在主题四上的表现最强，在主题二上的表现最弱。如果将重叠参数设为b =1.1，以模拟文档之间的高重叠度，则得到搜索引擎结果页面-主题相关性 = <2.26, 3.00,3.42, 2.81> 以及$\sum_i$ 搜索引擎结果页面-主题相关性[i] = 11. 49.

表2。基于表2数据，包含多维相关性评分组成部分的简化版搜索引擎结果页面。相关性评分 {0, 1, 2, 3}，可用性因子 [0, 1], and b = 2。

SERP	**Orig**	主题				贡献可用性因子		**Doc**	**Cum**
文档编号	总相关性	r_1	r_2	r_3	r_4	$\sum r_i$	v_i	分数	**Rel**
D_1	6	0,000	1,000	3,000	2,000	6,000	1,000	6,000	6,00
D_2	4	2,000	0,000	0,000	2,000	4,000	0,567	2,268	8,27
D_3	3	1,000	0,000	1,262	0,000	2,262	0,900	2,036	10,30
D_4	4	0,000	0,000	1,434	0,500	1,934	0,504	0,975	11,28
D_5	5	0,631	2,000	0,000	0,922	3,553	1,000	3,553	14,83
D_6	2	0,000	0,000	0,000	0,820	0,820	0,800	0,656	15,49
Sum	24	3,631	3,000	5,696	6,242	18,569	…	…	…

基于主题、效用和重叠的文档及搜索引擎结果页面评分，能够提供有关查询性能的重要数据，但不同搜索任务之间的可比性存在问题，因为与累积增益（Järvelin & Kekäläinen，2002年）的情况一样，这种评分并未归一化。为解决这一问题，我们定义了考虑重叠的理想多维评分及其归一化形式。

4.2. 理想排名——动态重叠

对于搜索任务 ST 和相关性语料库 RC，理想排名是指在避免因重叠结果而给予奖励的前提下，按增益累积效果对文档进行的最佳排序。原则上，每次都应从 RC 中剩余的文档里，选择贡献最大的文档排在下一位。由于主题内容已经累积，不能直接使用主题的面值。在构建理想排名的每一步中，我们都会选择能够带来最大剩余贡献的文档。

定义 9：给定相关性语料库 RC = { $d_1, d_2, ..., d_k$}，其理想排序为 ideal-ranking(RC, crr) =<id || ideal-ranking(RC – {d}, cum-rel(crr, <d>))>，其中 d ∈ RC ∧ ∀ d’ ∈ RC：contrib(crr, d) ≥ contrib(crr, d’)，如果 |RC| ≥ 2<id>，否则，其中 id = <d[1], cum-rel(crr, d), d[3]>

I在 *ideal-ranking* 的第一个调用中，操作数 *crr* = 空结构(*n*)，其中 *n* = 长度(d[2]）。如果相关性语料库很大，构建理想排序的计算开销会很大。当文档集、其主题或属性发生变化时，也必须重新构建。

表4 (a)。相关性语料库 RC'，10 份文档。相关性评分 {0, ..., 3}，可用性因子 [0, 1]。

RC'	主题相关度					可用性属性评分			
文档编号#	The-1	The-2	The-3	The-4	…	Attr1	属性2	Attr3	…
d1	0,00	1,00	3,00	2,00	…	1,00	1,00	1,00	…
d2	2,00	0,00	0,00	2,00	…	0,90	0,70	0,90	…
d3	1,00	0,00	2,00	0,00	…	1,00	0,90	1,00	…
d4	0,00	0,00	3,00	1,00	…	0,80	0,90	0,70	…
d5	1,00	2,00	0,00	2,00	…	1,00	1,00	1,00	…
d6	0,00	0,00	0,00	2,00	…	1,00	0,80	1,00	…
d7	0,00	0,00	0,00	0,00	…	0,00	0,00	0,00	…
d8	1,00	1,00	1,00	0,00	…	0,30	1,00	1,00	…
d9	0,00	0,00	0,00	2,00	…	0,90	0,90	0,90	…
d10	3,00	3,00	3,00	1,00	…	1,00	1,00	1,00	…

表4 (b). 基于 RC' 相关性评分 {0, 1, 2, 3}、效用因子 [0, 1]，和 b = 1 的理想排名。5.

理想 iOAr	主题相关度				**Face**	**Disc**	属性	**Doc**	
文档编号 The-1		The-2	The-3	The-4	值	贡献因子		分数	**iOAr**
排名1 d10	3,00	3,00	3,00	1,00	10,00	10,00	1,00	10,00	10,00
排名2 d1	0,000	0,369	1,107	2,000	6,00	3,48	1,00	3,48	13,48
排名3 d5	0,369	0,668	0,000	0,738	5,00	1,77	1,00	1,77	15,25
排名4 d3	0,334	0,000	0,574	0,000	3,00	0,91	0,90	0,82	16,07
排名5 d2	0,619	0,000	0,000	0,615	4,00	1,23	0,57	0,70	16,77
排名6 d4	0,000	0,000	0,861	0,369	4,00	1,23	0,50	0,62	17,39
排名7 d6	0,000	0,000	0,000	0,522	2,00	0,52	0,80	0,42	17,81
排名8 d9	0,000	0,000	0,000	0,489	2,00	0,49	0,73	0,36	18,16
排名9 d8	0,277	0,291	0,237	0,000	3,00	0,80	0,30	0,24	18,40
排名10 d7	0,000	0,000	0,000	0,000	0,00	0,00	0,00	0,00	18,40

表4 (a) 给出了包含10篇文档的相关性语料库 RC'，其中 d1 – d6 与表2相同。表4 (b) 展示了它们的理想排名。每篇文档对各主题及总计的贡献均已给出。

独立值，并根据基于适中参数估计的重叠进行折扣b = 1.5。右侧的三列分别包含属性因子v、每个排名处的文档得分和每个排名处的累积增益。

请注意，在所有主题中使用单一重叠参数 b 和单一属性因子 v，是一种简化设计选择。实验可能表明，使用多个系数值得引入额外的复杂性。

4.3 归一化增益

搜索任务搜索结果页可能产生的累积增益，会因相关性语料库的规模、假定的文档重叠情况以及属性因素的严格程度而有很大差异。不同搜索任务的增益无法直接比较，某些搜索任务可能会主导总体情况（Jarvelin & Kekäläinen，2002年）。当需要比较不同搜索任务时，可以相对于理想结果对搜索结果增益进行归一化。给定搜索结果 SR，以及其理想排序为 IR = ideal-ranking(RC, null-struct(n)) 的相关性语料库 RC，可以用一种简单的方式计算归一化增益。二者都是由文档表示 d = (dc, TRel, Attr) 构成的元组，其中每个主题相关性分量都会针对重叠进行折扣，并在整个排序中累积。接下来需要提取两个排序的累积相关性向量，并按排名逐位将观测性能除以理想性能。可以分别针对每个主题执行此操作，也可以针对它们的任意组合执行。还可以针对任意属性子集执行此操作。下面定义用于将基于主题的相关性数值聚合为总体得分的函数，并根据属性因素进行折扣。首先，我们定义 doc-gain 和 gain-vector 函数，分别用于提取单个文档的跨主题增益，以及基于文档排序构建增益向量。

定义 10：设 d = (dc, Trel, 属性) 为任意文档，R = <d_1, d_2, ...，d_k> 为包含 k 个文档的任意排序。d 提供的增益为

$$\textit{doc-gain}(d) = \sum_{i=1,,\ldots,n} \text{Trel}[i] * \textit{attr-fact}(d)$$

T基于带有可用性折扣的跨主题相关性，R 的累积增益向量为

$$\textit{gain-vector}(R, k) = \\ < cg_1, cg_2, \ldots, cg_k > \\ \text{where } cg_i = \sum_{j=1,\ldots,i} \textit{doc-gain}(d_j)$$

例如，在表4 (a)中，文档 d4 提供的增益为文档增益(d4) = (0.00+0.00+3.00+1.00) * (0.80*0.90*0.70) = 4*0.5 = 2。当该文档在表4 (b)中排名第 6 时，此时应用重叠折扣，得到文档增益(d4) = (0.00+0.00+0.861+0.369) * 0.5 = 0.62。表4 (b)最右侧一列给出了样本理想排序的增益向量，排名第 10 时以 18.40 分收尾。

最后，我们将增益向量归一化定义为两个向量逐组件进行除法运算。搜索结果提供分子组件 < sg_1, sg_2, ...，$_k$ > 而理想排名提供分母组件 < ig_1, ig_2, ...，$_k$ >。

定义 11：设 IR 为理想排名，SR 为待归一化的观测排名搜索结果，二者长度均为 k，ig = 增益向量(IR，k) 和 sg = 增益向量(SR，k) 分别为对应的增益向量。归一化后的增益向量为：

$$\textit{norm-vector}(\text{sg}, \text{ig}) = \text{sg} \oslash \text{ig} = \langle \text{sg}_1/\text{ig}_1, \text{sg}_2/\text{ig}_2, \ldots, \text{sg}_k/\text{ig}_k \rangle .$$

表 5 给出了归一化增益向量计算的示例。

表 5. 归一化增益向量的计算：给定搜索增益向量和理想增益向量，二者均基于主题相关性评分 {0, 1, 2, 3}、可用性因子 [0, 1]，和 b = 1.5。

SR	搜索累积增益	理想 R	iCG	nCG
d1	6,00	**ir1 d10**	10,00	0,60
d2	7,80	**ir2 d1**	13,48	0,58
d3	8,99	**信息检索3 d5**	15,25	0,59
d4	9,63	**信息检索4 d3**	16,07	0,60
d5	12,70	**信息检索5 d2**	16,77	0,76
d6	13,16	**信息检索6 d4**	17,39	0,76
d7	13,16	**信息检索7 d6**	17,81	0,74
d8	13,46	**信息检索8 d9**	18,16	0,74
d9	13,84	**ir9 d8**	18,40	0,75
d10	16,84	**ir10 d7**	18,40	0,92

5. 讨论

MDCU 框架可依据若干标准（例如，Järvelin 和 Wilson，2003年）加以论证。我们将重点关注其目标和范围、其在各种条件下的表现、测试集构建中的问题、有效性和成本、对实验的支持，以及与信息检索其他领域的整合。

5.1. MDCU 或信息检索评估的目标和范围

信息检索和信息检索评估的目标。传统方法中，信息检索的实验目标是为搜索请求找到主题相关的文档。信息检索评估提供工具和指标，用于评估搜索引擎在这方面的表现。作为主要实验变量，更准确的搜索引擎更具优势。我们提出了更高要求的目标：信息检索应尽可能找到简洁的答案，帮助人们克服问题情境。因此，请求可能需要涉及一个人任务（或其他参与活动）的多个方面、此人已有或缺乏的知识，以及其个人能力、技能和兴趣。忽略这些方面，意味着对基于任务的评估只是敷衍了事。修订后的目标也适用于信息检索作为现代问答系统或交互式人工智能系统的检索组件的情形。

评估框架的范围。 我们提出了一个抽象框架，用于将任务、情境和人员相关方面纳入信息检索评估。之所以称其为抽象框架，是因为我们尚未针对任何特定评估案例明确相关性的主题或属性 –，也没有测试集。不过，这些内容可以具体化，例如从以下草图开始：

- 人物：苏黎世的一名高中生，对绿色运动持中度倾向- 任务：撰写一篇关于应对气候变化措施的文章，用于课堂展示- 情境：初步收集材料，对主题所知甚少- 文档集合：瑞士境内涵盖5种语言的多语种媒体档案- 主题：污染者、污染物、后果、绿色技术、谈判、协议……- 属性：是否为专题文章、是否带有绿色立场偏见、是否为法语、是否由 Facebook 好友分享、是否面向非专业人士……

这些主题和属性涵盖了该案例中与主题、任务、情境和个人相关的方面。我们已经表明，该框架能够以可计算的形式将它们纳入信息检索评估。因此，MDCU 框架的范围比传统评估模型更广；传统模型不会促使人们同样广泛地探究成功或失败的原因。概念上较为贫乏的模型为分析成功提供的依据较弱，也会使分析受到更多限制。

主要区别。传统信息检索评估方法关注搜索结果中各项目序列的相关性，并提供若干指标来衡量结果质量。将表3视为搜索引擎通过传统视角返回的 SERP S1：附录1以多种方式将 S1 归并为单维度的平均相关性和相关性分数向量——通过两个阈值将其二值化，以计算 P@R（排名精确率），以及若干基于累积增益的向量：iCG（理想累积增益）、S1 CG、iDCG（理想折损累积增益）和 S1 DCG。这些内容见图1(a)–(b)。每个指标都形成一系列列，每个排名（文档）对应一个读数。虽然这些读数说明了 S1 的增益，但我们无法了解 S1 的效用、重叠程度或代表不足的主题。

图 1（c）–（d）展示了基于 MDCU 的 SERP1 评分。左侧的朴素评分基于每个项目的聚合相关性乘以其可用性。我们可以看到，由于可用性较低，d4 和 d8 的分数不高。右侧展开了相关性维度。我们可以看到，单个项目的分数基于不同的维度（例如 d3 与 d5），这表明它们之间的重叠较少；而 d6 和 d9 可能是重复且质量较差的项目。与完整的 SERP1 相比，项目 d1 和 d10 可能组成更优的简洁双文档组合。MDCU 提供了用于评分和比较搜索引擎结果页面质量的指标。

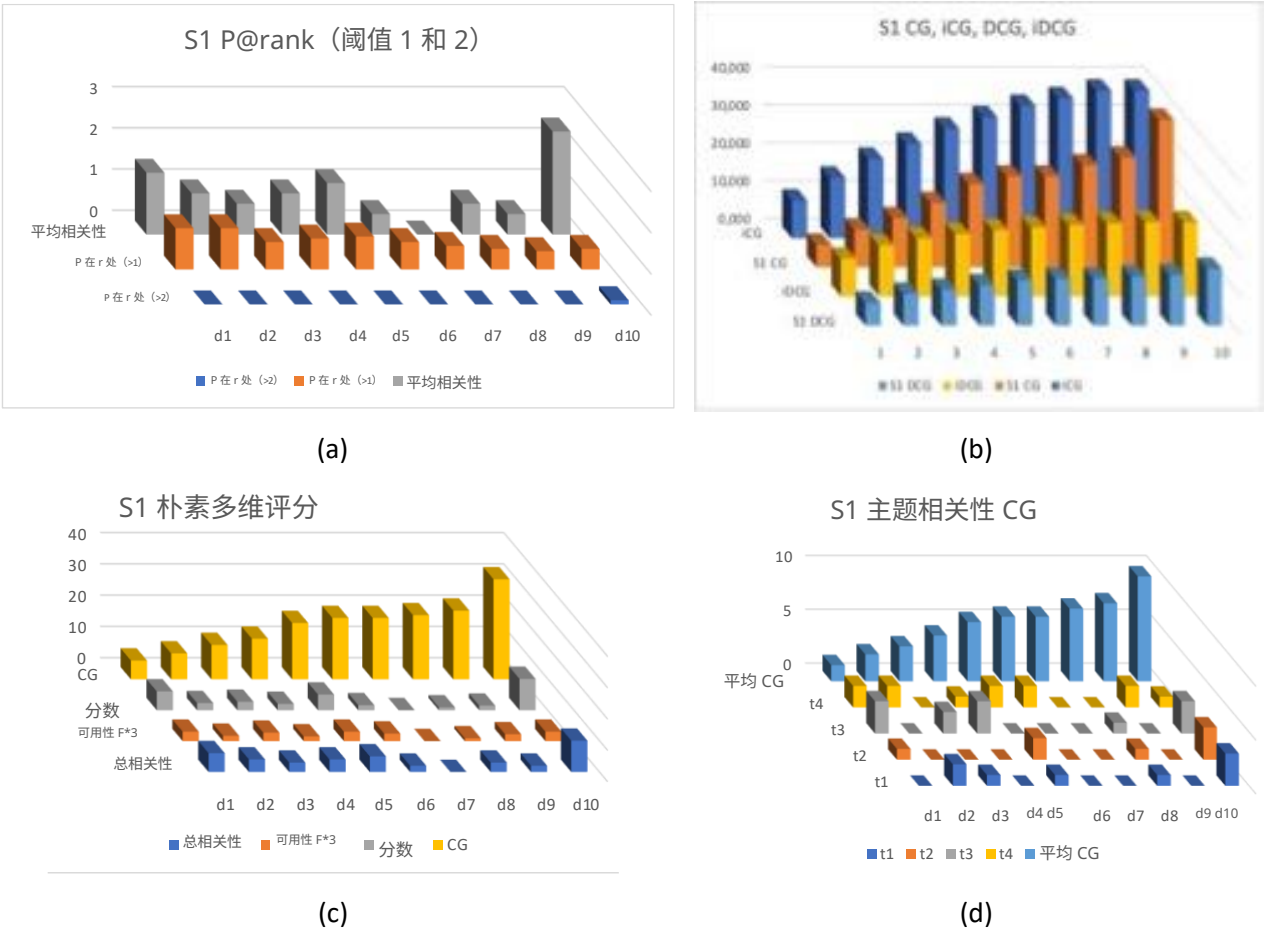

图 1. (a)-(b) 通过单维指标展示 SERP S1 的质量，以及**(c)-(d)**与多维指标进行对比。**注：** 在 (c) 中，可用性系列的值乘以 3 以便于展示。

图 2–4 图例：

图 2–4 展示了该框架基于抽象合成评估数据的行为：相关性主题和属性与表 2–5 中的一样，都是抽象的。这些值是合成的，但完全可能出现。附录 I 给出了生成这些图表的表格。图表标题经过编码，用于指示所展示的评估类型：代码中的各个位置分别对应相关性重叠、属性和基于排名的折扣。在每个位置，大写字母 O/A/R 表示应用了相应类型的折扣，小写字母则表示未应用。例如，“OAr”表示应用了重叠折扣和属性折扣，但关闭了基于排名的（= 归一化折损累计增益）折扣。

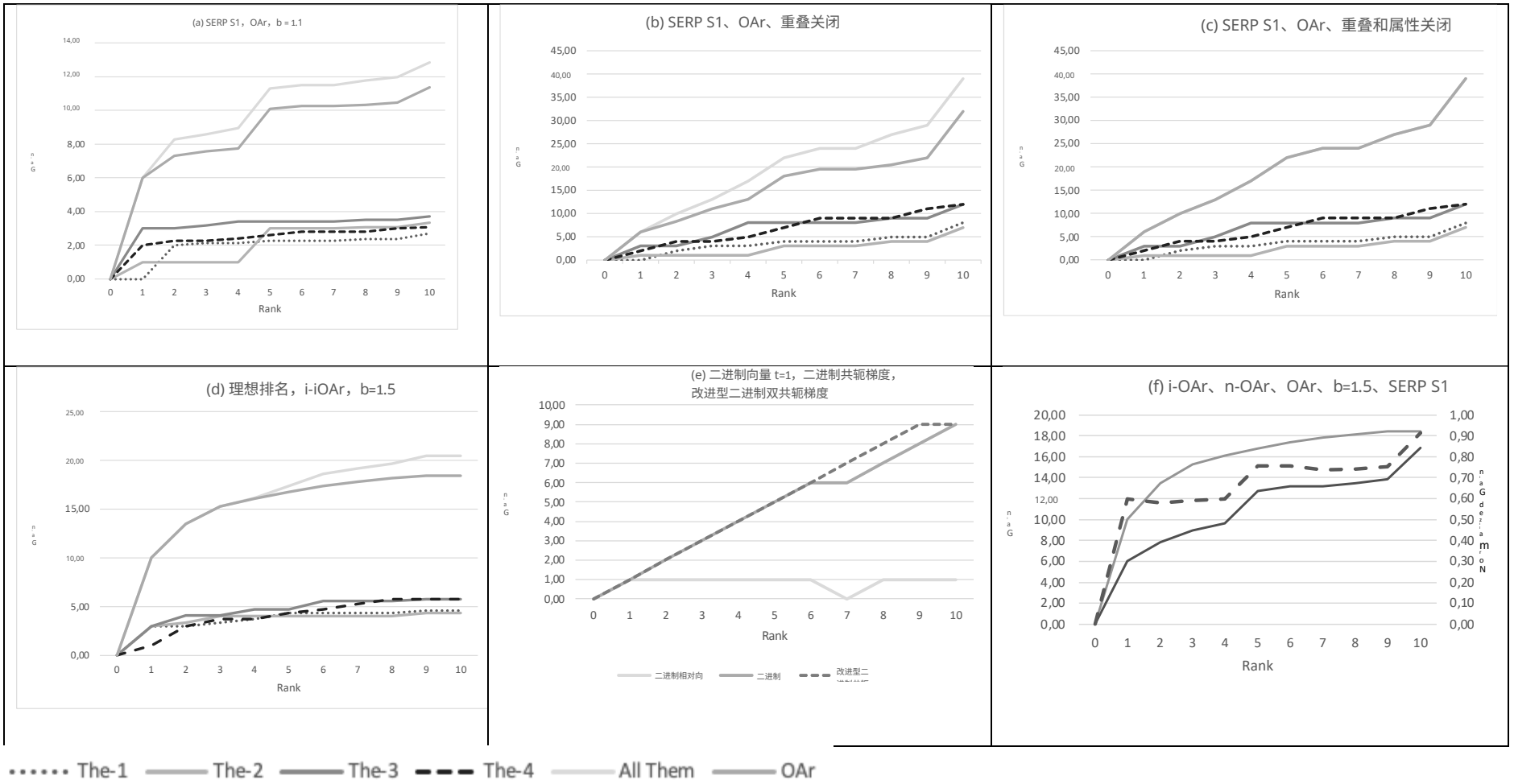

图2。（a）-（c）带有/不带有重叠折扣和属性降级的有效性图，以及**（d）-（f）**各种理想曲线和参考曲线，主要基于 SERP S1。

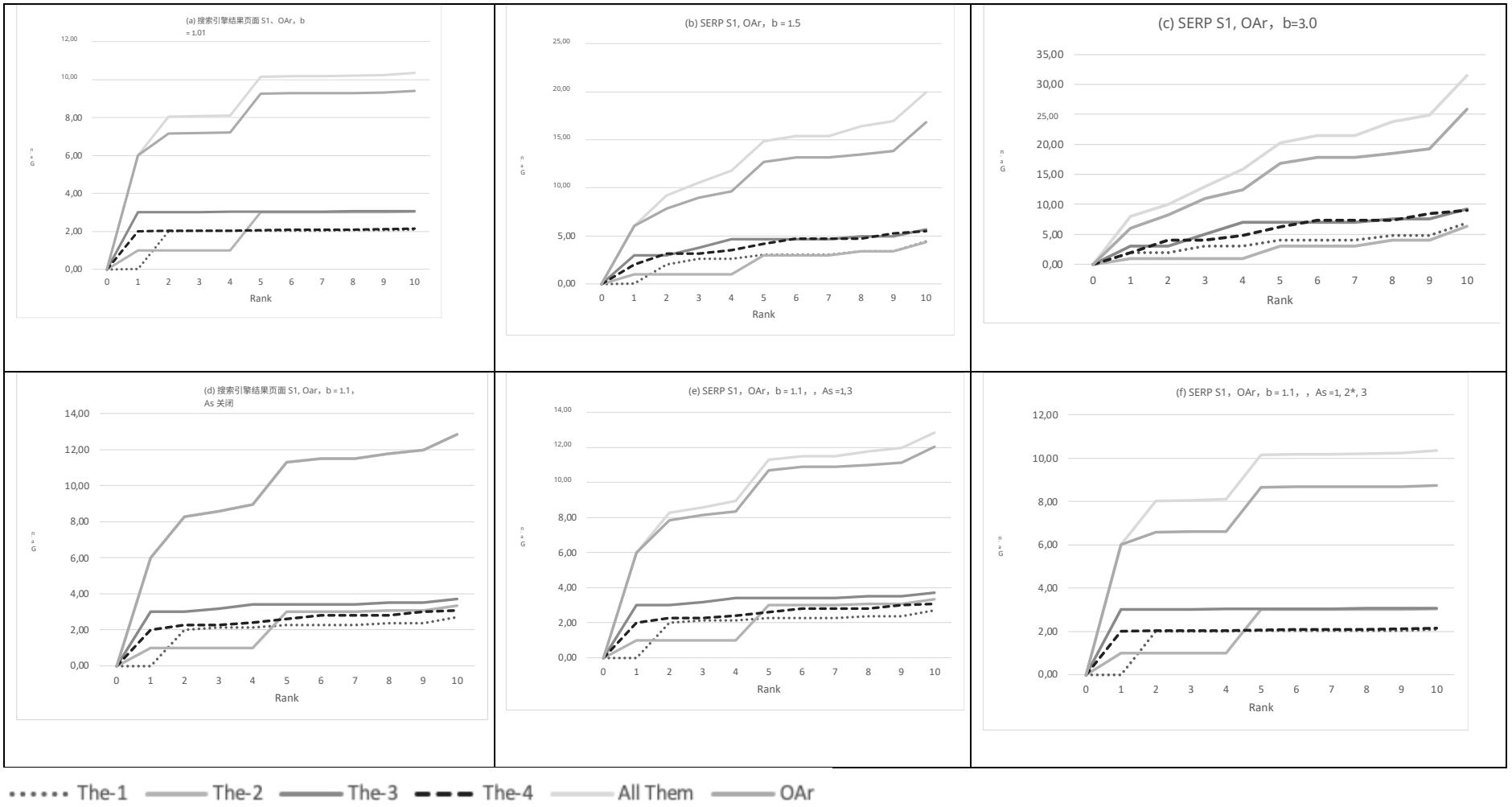

图3. (a)-(c) 重叠折扣的影响，不进行属性降级，以及**(d)-(f)**变化的属性值，均基于 SERP S1。

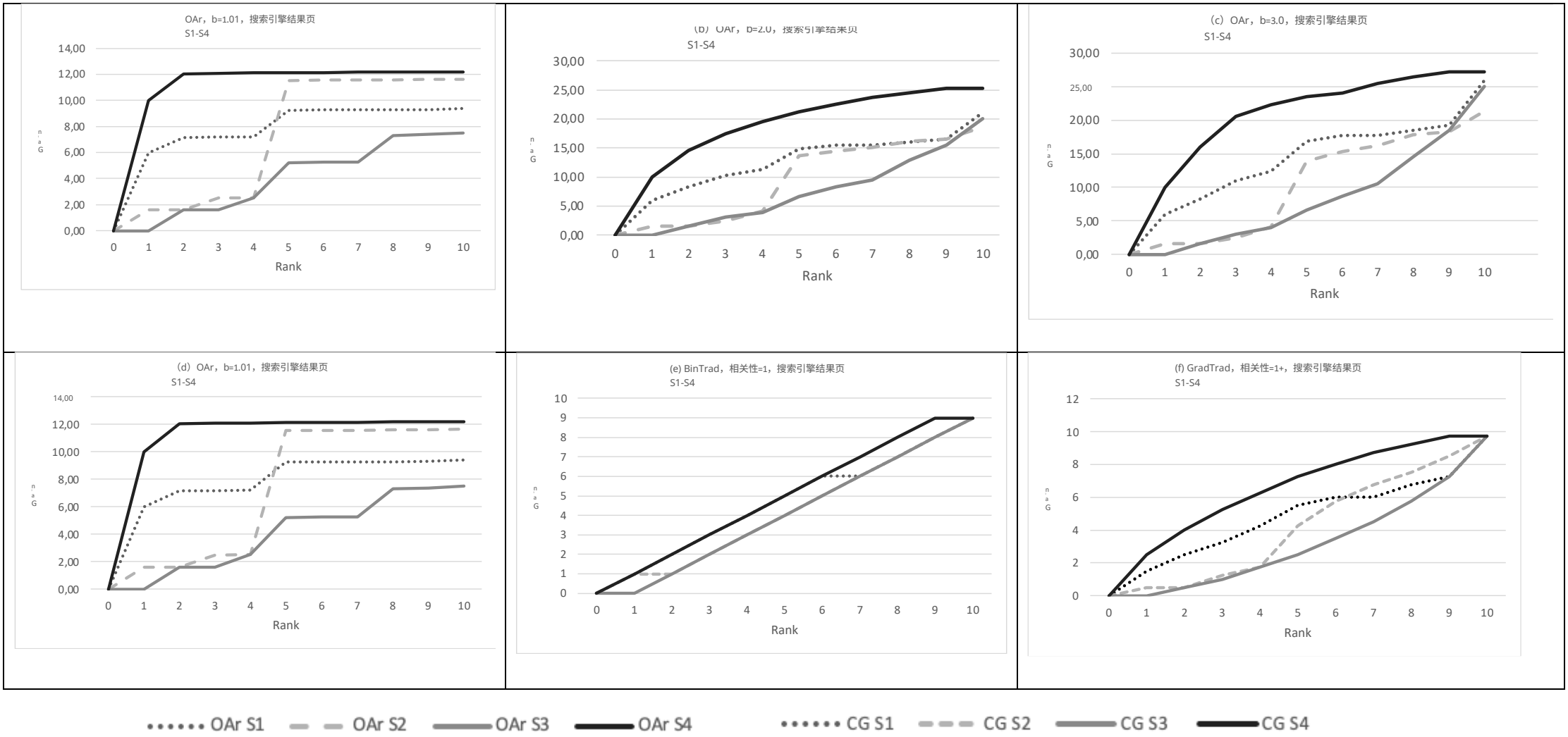

图4. (a)-(c) 重叠折扣对搜索引擎结果页面质量的影响，以及**(d)-(f)**与传统指标的对比。

5.2. MDCU 的行为

指标的行为与响应性。MDCU 的有效性得到了其在各种条件下表现符合预期这一行为的支持。图2（a）–（c）展示了 SERP S1 在中度重叠和属性折扣分别为（a）启用以及（b–c）停用时的表现。每幅图中的上方曲线表示跨主题增益，下方四条曲线表示单个主题。总体而言，我们注意到，曲线的水平和形状会根据 MDCU 折扣的启用或停用情况而变化。图2（d）–（f）展示了 S1 的理想 MDCU，并将其与几条参考曲线进行了对比。在（e）中，采用二元增益时，没有迹象表明存在用于停止的重叠——S1 似乎表现得非常好。在（f）中，S1 显著低于理想值（深色实线），其归一化曲线（虚线）位于理想增益的 0.5–0.95 之间（右侧为纵轴）。

图3 (a)–(c) 展示了重叠对 SERP S1 性能的影响：在高度重叠 (a) 的情况下，前 1–2 篇文档几乎贡献了每个主题及总体上的全部非重叠增益；而对重叠采取宽松处理 (b-c)，则会奖励排名较低的重叠结果。当忽略重叠时，主题 3 和 4 的表现似乎更好 (b > 5)。图3 (d)–(f) 展示了 SERP S1 性能的更多示例：(d) 不考虑属性时采用中度重叠；(e-f) 在修改后的属性下，分别采用高度重叠和宽松重叠（见附录）。

图4 比较了附录中给出的四个 SERP S1–S4 在不同条件下的表现。这些 SERP 包含相同的 10 条记录，但排列顺序不同。它们可能是针对同一搜索任务的不同搜索引擎所返回的结果。上排 (a)-(c) 展示了重叠折扣对搜索结果页面评分的影响。我们注意到，S1 和 S2 的优劣排序会随着重叠权重的变化而改变。下排 (d)-(f) 将 MDCU 与传统指标进行对比。我们注意到，基于宽松二元相关性的累积增益 (e) 无法清晰地区分各 SERP；而基于分级相关性的曲线与采用宽松重叠的 MDCU 曲线相似，但与严格重叠情形 (d) 下的曲线不同。这是合乎逻辑的。

这些示例表明，该框架能够响应核心参数 b 和属性因素的变化，其响应也符合逻辑。对于每个主题和整个搜索结果而言，一旦预计不会再出现新的非重叠结果，MDCU 曲线就会变为水平。这可能远早于最后一个相关但与其他结果重叠、且会被传统指标计分的结果出现的时间。

传统指标无法指出从任务角度看相关内容未能得到妥善排序时，问题出在哪里。MDCU 框架至少可以指出哪些相关性主题或属性与排序不成功有关。这些指标倾向于选择能以简洁形式最大化增益的搜索结果——不会因重复提供边际信息而给予奖励。从概念上看，该框架比传统方法更为宽泛。它将重叠、属性和相关性主题等新概念引入信息检索评估，为将与用户和任务相关的因素纳入信息检索评估提供了一种系统方法。

另一方面，所提出的框架具有通用性，因此可以将传统评估框架推导为其特例。将所有属性设为 1，即可关闭这些属性。将重叠折扣参数 b 设为较高值，即可忽略重叠（>5）。通过对（或许经过筛选的）主题相关性得分求和，并设置二元／分级相关性的阈值，即可得到单维相关性。

5.3. 构建 MDCU 框架

构建测试集 – 情境中的搜索任务。所提出的框架支持设计基于任务、情境和个人的搜索任务。相关性主题是表示个人在搜索任务之外、更广泛任务各个维度的一种自然方式。对真实情境中的 TBII（见第 2.2 节）进行分析有助于此。可在特定实验中对这些主题或整个搜索任务进行分类和选择，或将其排除，以体现个人所处的情境以及对更广泛任务的了解。遵循搜索任务描述的“标题—描述—叙述”（TDN）结构，叙述部分可按主题包含词汇，从而在构建相关性语料库时支持算法辅助的重叠评估。TBII 研究和搜索引擎日志有助于明确这些属性。搜索任务越复杂，MDCU 框架似乎就越合适。

构建测试集 – 相关性语料库。可通过扩展分析，从传统相关性语料库中导出相关性语料库。所有至少具有边缘相关性的文档都将针对与任务、情境和个人相关的方面重新分析。相关性主题的评分方式可与最先进水平的测试集构建方法类似——如今借助主题所提供的结构，评分更加可靠——可通过人工判断、众包或统计／算法辅助来完成。可用性属性的评分通常可由算法完成（语言、难度、偏见、难度等），但在某些情况下，需要明确数据源及其分析功能（例如社交媒体或推荐系统数据）。

局限性——有效性。由于依赖过于简单的假设，且未引入新的概念，传统方法的评估受到限制。MDCU 中的假设颇为大胆，尤其是关于按累积增益进行折扣的假设。不过，该框架允许开展假设分析实验，调整重叠假设。对于输入的变化，其评分行为看起来也合乎逻辑。此外，这些假设在各种情形下的有效性都可以通过实证检验。毕竟，哪种假设更有效：(a)“无论文档在结果中的位置如何，所有文档对累积增益的贡献都相同”，还是 (b)“已收集的增益越多，剩下可收集的新颖增益就越少”？因此，只要能够承担额外的工作量，且评估者输入的数值准确，MDCU 框架看起来就是有效的。

局限性——成本。MDCU 所需的测试集显然比传统测试集更复杂。其构建需要付出更多努力。作为回报，这些集合能够产出更令人振奋的研究成果。我们既没有成本估算，也没有令人信服的大规模实证案例可供展示。因此，标题中的“蓝图”一词

可能显得不够严谨。然而，信息检索也是一个数十亿美元规模的产业，因此相对成本是微不足道的。此外，分级相关性评估最初也因成本原因而遭到反对，但如今已多年属于信息检索评估中的最先进水平。可以通过测量文档（段落）与搜索任务主题规范之间的距离（在 N 组件中），来支持主题相关性评分。同样，也可以借助对相关上下文的用户档案进行机器学习来辅助属性评分。同时，还可以通过对各种类型文档之间重叠情况开展实证研究，进一步完善重叠处理。

局限性——无需使用：有人可能会认为不需要 MDCU，因为多维相关性问题可以视为基于主题的文档流的融合问题。的确，这是开发搜索与排序算法的一种方式，但与评估没有直接关系。

5.4. 使用 MDCU 框架进行实验

支持系统化实验。要实现帮助人们完成任务这一目标，信息检索评估必须纳入用于表征任务、情境和个人特征的一些概念。否则就只是空想。MDCU 框架为信息检索评估引入了新概念及其操作化方式，并推动围绕新问题开展系统化实验。以下是一些示例：

- [重叠估算／折扣：] 不同类型的集合中，重叠问题有多显著？重叠程度有多高？在不同条件下，评估结果受到的偏差有多严重？- [属性折扣：] 用于将文档标注为难懂、有偏见或带有情绪色彩的文本分类方法，能否有效地用于降低搜索结果中较不实用文档的排名？- [识别相关性主题：] 搜索者在主题内部及跨主题的知识程度如何影响搜索与结果？在某次会话中，为了学习，搜索者在搜索其他主题时会接触到哪些主题或概念？- [相关性主题分类：] 哪些类型的主题或搜索键最适合作为通往其他主题的入口？哪些主题最容易获得较高排名或较早积累？

5.5. MDCU 与信息检索的其他领域

支持与其他研究领域的融合。信息检索这一广泛领域内部或与之密切相关的若干研究领域，都可能受益于符合 MDCU 标准的高效搜索引擎。例如

- [问答系统：] 开发查找查询简洁答案的方法具有互惠优势：问答方面所需的自然语言处理更少，同时也为确定问答类型信息检索任务的相关性主题和属性提供了实证依据。

- [推荐系统:] 关于推荐系统如何组织用户交互的数据，可用于确定相关性主题和属性，以及推荐系统类型的信息检索任务中常见的属性值；从文本产品评论中生成简明摘要的方法。- [作为学习的搜索:] 假设有一个涉及三个概念 A、B、C 的搜索任务，且分别有其表达式 {..., a_i, ...}、{..., b_j, ...}、{..., c_k, ...} 可用。进一步假设搜索者对概念 A 只略有了解，知道一些表达式。观察搜索者的交互后，分析其在所见文档中接触了哪些概念，以及这些概念后来是否被使用或记住。

还有一些限制更多、结构更明确的信息环境，例如电子商务；与一般的网络信息检索相比，其中的相关性主题更加明显且持久。这一点我们留待以后讨论。

6. 结论

我们认为，信息检索评估并未很好地服务于信息检索研究与开发，因为它在相关性、文档内容重叠、搜索结果中的增益累积，以及文档可用性属性的作用等方面所作的假设存在问题。如果所有搜索引擎在最先进的基于测试集的评估中都表现良好，那么搜索任务就过于简单。该评估并未检验检索方法是否能够帮助搜索者完成其背后驱动的更大任务。因此，需要设定要求更高的目标。这些目标需要一个新的评估框架。我们提出

在执行任务时，针对特定情境，为个人或软件应用提供简洁、不重叠且有帮助的答案。

作为信息检索的目标，以及

通过检索和排序文档、组件或信息，衡量搜索引擎实现这一信息检索目标的效果。

作为信息检索评估的目标。这要求在信息检索中明确任务–情境–检索者特定因素。因此，我们提出了一种新的评估框架，使用多维分级主题相关性和属性来表示这些因素。遵循 Newell（1993）的做法，我们介绍了评估任务、其中一些算法的定义及其论证。该框架提供了称为 MDCU 指标（即多维累积增益指标）的新型评价指标，可在评估中控制任务、情境、检索者因素以及文档重叠。我们通过示例正式定义并论证了这些算法，示例表明，与传统信息检索评估相比，这些算法的行为符合预期。我们还讨论了其局限性：构建所需测试集的成本显然高于构建同等规模传统测试集的成本。作为回报，该框架促进了要求更高且更有效的信息检索评估。

本文的贡献如下：

- 提出将任务、情境和搜索者属性作为因素纳入信息检索评估。- 提出避免对文档相关性和重叠作出简单化假设。- 提出一种通用方法：将传统信息检索评估视为特殊情况；为其他将搜索引擎作为组件的信息检索子领域提供服务。

参考文献

Barry，C.L.（1994年）。用户定义的相关性标准：一项探索性研究。《美国信息科学学会杂志》，45(3)：145-159。

Barry, C.L. 和 Schamber, L.（1998年）。用户相关性评估的标准：跨情境比较。《信息处理与管理》，34(2-3)：219-236。

Belkin, N.J.（2016）。《人与信息交互》。SIGIR论坛，49(2)：13 –27. https://doi.org/10.1145/2888422.2888424

Borlund，P.（2003年）。信息检索中的相关性概念。《美国信息科学学会杂志》与技术，54(10)：913-925。

Borlund, Pia（2016）。模拟工作任务情境在交互式信息检索评估中的应用研究。《文献学杂志》，72(3)：

Borlund, P. & Ingwersen, P.（1998）。相对相关性度量与排序半衰期：交互式信息检索的绩效指标。载于：Croft, B.W. 等（编），*21st* 信息检索研究与开发ACM SIGIR会议论文集。第324-331页。纽约：ACM出版社。

Byström, K. & Järvelin, K.（1995）。任务复杂度影响信息寻求与使用。《信息处理与管理》31(2)：191 – 213。

Croft, W. B., Metzler, D., & Strohman, T.（2010）。搜索引擎：实践中的信息检索（第第520卷，第131-141页）。雷丁：Addison-Wesley。

Ferro, N.、Silvello, G.、Keskustalo, H.、Pirkola, A.、Jarvelin, K.（2016）。信息检索的 Twist 度量评估：将用户付出的努力纳入考量。《信息科学与技术协会期刊（JASIST）》，67(3)：620-648。数字对象标识符：10.1002/asi.23416。

Fuhr, N. 等（2017）。在线文档的信息营养标签。SIGIR论坛，51 (3)：46-66。ISSN 0163-5840

Fuhr, N.、Rölleke T.（1997年）。用于整合信息检索与数据库系统的概率关系代数。ACM 信息系统汇刊（TOIS）

Greisdorf, H.（2003年）。相关性阈值：用户如何评估的多阶段预测模型信息。《信息处理与管理》39（2003年），403-423。

Harman, D. K., & Voorhees, E. M.（2006年）。TREC：概述。《信息科学与技术年度评论》与技术，40，113。

Harter, S.P. & Hert, C.A.（1997年）。信息检索系统评估：方法、议题与方法。载于：Williams, M.E.（编），《信息科学与技术年度评论》（ARIST）32（），3-94。Medford：ASIS & Information Today。

INEX, ***IN****itiative for the* ***E****valuation of* ***X***ML 检索，（2014年）。https://inex.mmci.uni-saarland.de/index.html（引自2022年11月26日）。

Ingwersen, P. 与 Järvelin, K.（2005年）。转向：情境中的信息寻求与检索整合 . 海德堡：施普林格，448 页。

Järvelin, K. 等（2015）。基于任务的信息交互评估：程序理论的视角
ACM 信息系统汇刊（ACM TOIS）33(1)，第3篇文章，情境搜索与推荐特刊
与推荐，30页。数字对象标识符：http://dx.doi.org/10.1145/2699660Järvelin, K. & Kekäläinen, J.（2000）。面向高度相关文档的信息检索评估方法。载于：Belkin, N.J.、Ingwersen, P. & Leong, M.-K.（编）SIGIR 2000：第23届年度研究会议论文集

与信息检索领域的发展（ACM SIGIR 23）。纽约：ACM出版社，第41-48页。Järvelin, K. & Kekäläinen, J.（2002年）。信息检索技术的累积增益评估。ACM 信息系统汇刊（ACM TOIS），20(4)：422-446。Järvelin, K. & Kekäläinen, J. & Niemi, T.（2001）。ExpansionTool：基于概念的查询扩展与构建。信息检索 4(3/4)：231-255。Järvelin, K., & Wilson, T. D.（2003年）。论信息寻求与检索研究的概念模型。信息研究，*9*(1)，9-1。Katzer, J. 等人，（2017年）。文档表示

0352Kekäläinen, J. & Järvelin, K.（2002a）。在交互和多

Vakkari, P.（编）第4届CoLIS会议论文集。科罗拉多州格林伍德村：Libraries Unlimited，第253-270页

Kekäläinen, J. & Järvelin, K.（2002b）。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。Keskustalo, H. & Järvelin, K. & Pirkola, A. & Kekäläinen, J.（2008）。基于显式负向高阶相关性的用户表现直觉支持型可视化。载于：Sung Hyon Myaeng 等（编），*31st* ACM SIGIR国际信息检索研究与发展年度会议论文集（ACM SIGIR 2008），新加坡，2008年7月20日至24日。纽约：ACM出版社，第675-682页。Kumpulainen, S. & Järvelin, K.（2010）。分子医学中的信息交互：多种渠道的综合运用。载于：Belkin, N. 等（编），情境中的信息交互会议论文集（IIiX 2010），新泽西州新不伦瑞克，2010年8月。纽约：ACM，第95-104页。ISBN：978-1-4503-0247

0. 数字对象标识符：10.1145/1840784.1840800Lehtokangas, R. & Järvelin, K. (2001). 报纸文章中的文本表达一致性。 期刊 o

文献资料，57(4)：535-548。Li, Y. & Belkin, N.J. (2010) 工作任务与交互式信息搜索行为之间关系的探索。 《美国信息科学与技术学会会刊》，61：1771-1789。Maghlaughlin, K.L. & Sonnenwald, D.H. (2002). 用户对相关性标准的看法：对相关、部分相关和不相关判断的比较。《美国信息科学学会杂志》 53(5), 327-342。Mirończuk, M. M., & Protasiewicz, J. (2018)。文本分类最先进水平要素的最新综述。 《专家系统与应用》，*106*，36-54。Newell, A. (1993)。启发式编程：非结构化问题。载于 《Soar论文集》（第1卷）：集成智能研究（第3-54页）。Over, P. (2001). TREC交互式检索专题：一份注释书目。 《信息处理与管理》37(3), 369-381。Rieh, S.Y.

息科学学会杂志》 53(2), 145-161。

Robertson, S. E., Kanoulas, E. & Yilmaz, E. (2010). 将平均准确率扩展至分级相关性判断。 载于：

ACM SIGIR国际会议论文集：关于

信息检索研究与开发（SIGIR '10）。纽约，纽约州，ACM，第603-610页。数字对象标识符：https://doi.org/10.1145/1835449.1835550

Saastamoinen, M. & Järvelin, K.（2016）。不同类型和复杂程度工作任务中的搜索任务特征。《美国信息科学与技术学会杂志》（JASIST）68(5)：1111–1123。数字对象标识符：10.1002/asi.23766。（在线发表于2016年12月21日）。

Sanderson, M.（2010）。《基于测试集合的信息检索系统评估》，《信息检索基础与趋势》：

Saracevic, T.（1975）。相关性：对信息科学中这一概念的思考进行回顾并提出框架。《美国信息科学学会杂志》26(6)：321-343。

Saracevic, T.（2007）。相关性：对文献进行回顾，并提出一个思考信息科学中这一概念的框架。第二部分：相关性的本质与表现形式。《美国信息科学与技术学会会刊》，58（13）：1915-1933。

Saracevic, T., Mokros, H., Su, L., & Spink, A. (1991). 在线搜索中用户与中介之间交互的性质。

Sormunen, E.（2002年）。TREC的宽松相关性标准——能否容忍可忽略文档？载于：Beaulieu, M., Baeza-Yates, R. Myaeng, S. H. & Järvelin, K.（编），《第25届ACM SIGIR国际信息检索研究与发展年度会议论文集》。纽约：ACM：324–330。

Sormunen, E. & Kekäläinen, J. & Koivisto, J. & Järvelin, K.（2001年）。文档文本特征会影响扩展结构化查询对最相关文档的排序。《文献学杂志》57：358- 376。

Spink, A. & Greisdorf, H. (2001). 区域与层级：测量并绘制用户相关性判断图。《美国信息科学学会杂志》52(2), 161-173。

Stanford (2022). XML检索评估。 https://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-xml-retrieval-1.html (引自

Vakkari, P. (2016). 作为学习的搜索：基于文献的系统化。《信息科学杂志》42(1): 7-18。

Vakkari, P. (2018). 复杂任务中的信息搜索过程。载于《2018年人类信息交互与检索会议论文集》（CHIIR '18）。计算机协会，纽约，美国，1。https://doi.org/10.1145/3176349.3176570

Vakkari, P. & Hakala. N. (2000). 任务执行中相关性标准和问题阶段的变化。《文献学杂志》, 56: 540-562

附录：文档列表

SERP S1

I:oAr 文档编号	主题相关性				可用性属性		
	主题1	主题2	主题3	主题4	属性1	属性2	属性3
d1	0,00	1,00	3,00	2,00	1,00	1,00	1,00
d2	2,00	0,00	0,00	2,00	0,90	0,70	0,90
d3	1,00	0,00	2,00	0,00	1,00	0,90	1,00
d4	0,00	0,00	3,00	1,00	0,80	0,90	0,70
d5	1,00	2,00	0,00	2,00	1,00	1,00	1,00

d6	0,00	0,00	0,00	2,00	1,00	0,80	1,00
d7	0,00	0,00	0,00	0,00	0,00	0,00	0,00
d8	1,00	1,00	1,00	0,00	0,30	1,00	1,00
d9	0,00	0,00	0,00	2,00	0,90	0,90	0,90
d10	3,00	3,00	3,00	1,00	1,00	1,00	1,00

采用单维相关性的 SERP S1

文档编号	平均 相关性	P @ R (>1)	P @ R (>2)	总 相关性	总累积增益	平均 CG	
d1	1,50	1,00	0	6	6	1,5	
d2	1,00	1,00	0	4	10	2,5	
d3	0,75	0,67	0	3	13	3,25	
d4	1,00	0,75	0	4	17	4,25	
d5	1,25	0,80	0	5	22	5,5	
d6	0,50	0,67	0	2	24	6	
d7	0,00	0,57	0	0	24	6	
d8	0,75	0,50	0	3	27	6,75	
d9	0,50	0,44	0	2	29	7,25	
d10	2,50	0,50	0,1	10	39	9,75	

理想排名

:oAr	主题相关性				可用性属性		
文档编号	主题1	主题2	主题3	主题4	属性1	属性2	属性3
ir1 d10	3,000	3,000	3,000	1,000	1,00	1,00	1,00
ir2 d1	0,000	0,369	1,107	2,000	1,00	1,00	1,00
ir3 d5	0,369	0,668	0,000	0,738	1,00	1,00	1,00
信息检索4 d3	0,334	0,000	0,574	0,000	1,00	0,90	1,00
信息检索5 d2	0,619	0,000	0,000	0,615	0,90	0,70	0,90
信息检索5 d4	0,000	0,000	0,861	0,369	0,80	0,90	0,70
信息检索4 d6	0,000	0,000	0,000	0,522	1,00	0,80	1,00
信息检索5 d9	0,000	0,000	0,000	0,489	0,90	0,90	0,90
信息检索5 d8	0,277	0,291	0,237	0,000	0,30	1,00	1,00
ir5 d7	0,000	0,000	0,000	0,000	0,00	0,00	0,00

S1 但修改属性评分

可用性属性 图2 e	可用性属性 图2 f	未使用可用性属性

属性1	属性2	属性3	属性1	属性2	属性3	属性1	属性2	属性3
1,00	1,00	1,00	1	1,00	1	1	1	1
0,90	0,35	0,90	0,9	1,00	0,9	1	1	1
1,00	0,90	1,00	1	1,00	1	1	1	1
0,80	0,45	0,70	0,8	1,00	0,7	1	1	1
1,00	1,00	1,00	1	1,00	1	1	1	1
1,00	0,40	1,00	1	1,00	1	1	1	1
0,00	0,00	0,00	0	1,00	0	1	1	1
0,30	0,50	1,00	0,3	1,00	1	1	1	1
0,90	0,90	0,90	0,9	1,00	0,9	1	1	1
1,00	0,50	1,00	1	1,00	1	1	1	1

搜索引擎结果页面 S2 两半交换

II:OAr	带重叠折扣的主题相关性				可用性属性		
文档编号	主题1	主题2	主题3	主题4	属性1	属性2	属性3
Doc6	0,000	0,000	0,000	2,000	1	0,80	1
Doc7	0,000	0,000	0,000	0,000	0	0,00	0
Doc8	1,000	1,000	1,000	0,000	0,3	1,00	1
Doc9	0,000	0,000	0,000	2,000	0,9	1,00	0,9
Doc10	3,000	3,000	3,000	0,792	1	1,00	1
Doc1	0,000	0,792	2,377	1,402	1	0,30	1
Doc2	1,585	0,000	0,000	1,205	0,9	0,40	0,9
Doc3	0,639	0,000	1,186	0,000	1	0,90	1
Doc4	0,000	0,000	1,629	0,549	0,8	0,35	0,7
Doc5	0,601	1,402	0,000	1,060	1	1,00	1

SERP S3 最差

II:OAr	带重叠折扣的主题相关性				可用性属性		
文档编号	主题1	主题2	主题3	主题4	属性1	属性2	属性3
Doc7	0,00	0,0	0,000	0,00	0,00	0,00	0,00
Doc6	0,00	0,00	0,00	2,00	1,00	0,80	1,00
Doc9	0,00	0,00	0,00	2,00	0,90	0,90	0,90
Doc8	1,00	1,00	1,00	0,00	0,30	1,00	1,00
Doc3	1,00	0,00	2,00	0,00	1,00	0,90	1,00
Doc2	2,00	0,00	0,00	1,58	0,90	0,70	0,90

Doc4	0,00	0,00	3,00	0,64	0,80	0,90	0,70
Doc5	0,79	2,00	0,00	1,20	1,00	1,00	1,00
Doc1	0,00	1,00	1,84	1,10	1,00	1,00	1,00
Doc10	2,10	2,38	1,60	0,51	1,00	1,00	1,00

搜索引擎结果页面
S4 最佳

II:OAr	**带重叠折扣的主题相关性**				**可用性属性**		
文档编号	主题1	主题2	主题3	主题4	属性1	Attr2	属性3
Doc10	3,00	3,00	3,00	1,00	1,00	1,00	1,00
Doc1	0,00	1,00	3,00	2,00	1,00	1,00	1,00
Doc5	1,00	1,58	0,00	2,00	1,00	1,00	1,00
Doc2	1,58	0,00	0,00	1,37	0,90	0,70	0,90
Doc4	0,00	0,00	1,84	0,59	0,80	0,90	0,70
Doc8	0,64	0,64	0,53	0,00	0,30	1,00	1,00
Doc3	0,60	0,00	1,03	0,00	1,00	0,90	1,00
Doc6	0,00	0,00	0,00	1,13	1,00	0,80	1,00
Doc9	0,00	0,00	0,00	1,05	0,90	0,90	0,90
Doc7	0,00	0,00	0,00	0,00	0,00	0,00	0,00