# 查询性能预测逐点评估的可行性与稳健性

苏查娜达塔[1], 德巴西斯甘古利[2], 德里克 格林[3] 和 曼达尔 米特拉[4]

[1]爱尔兰都柏林大学[2]英国格拉斯哥大学[3]
爱尔兰都柏林大学[4]印度加尔各答印度
统计学院

arXiv:2304.00310v1 [cs.IR] 1 Apr 2023

**摘要**

尽管各个查询的检索效果彼此独立，查询性能预测（QPP）系统的评估一直是通过测量整个查询集合上的秩相关来进行的。这种列表式方法有若干缺点，尤其是无法满足评估单个查询的QPP这一常见需求。本文提出一种逐点查询性能预测框架，通过衡量每项预测与相应真实值之间的偏差，再对查询集合上的结果进行汇总，从而评估QPP系统对单个查询的预测质量。实验表明，在多种不同的目标指标和检索模型下，这种新方法能够降低QPP评估结果的方差。

## 1. 引言

查询性能预测（QPP）方法旨在自动估计查询的检索效果，而无需使用任何真实相关性信息（例如[1, 2]）。在实践中，QPP方法能够根据查询的初始性能估计值，动态调整其处理步骤。尽管在许多下游任务（例如自适应查询处理 [3]）中，独立估计各个查询的性能是一项常见要求，但信息检索研究社区所采用的标准QPP评估方法此前一直采用**列表式** 方法，而不是**逐点** 方法。尽管后者事实上更适合用于下游应用。具体而言，列表式方法首先对一个查询集合 $\mathcal{Q}$ 进行处理，根据QPP估计分数 $\phi(Q) \forall Q \in \mathcal{Q}$将其转换为有序集合。随后，它计算查询真实排序与由其平均精度（AP）值 [4] 或其他信息检索指标（如nDCG） [5]所诱导的排序之间的秩相关指标，例如Kendall's $\tau$。

*QPP++ 2023*：查询性能预测及其在新任务中的评估，与第*45*届欧洲信息检索会议（*ECIR*）同期举行，时间为*2023*年*4*月*2*日至*6*日，地点为爱尔兰都柏林
suchana.datta@ucdconnect.ie（S. Datta）；debasis.ganguly@glasgow.ac.uk （D. Ganguly）；derek.greene@ucd.ie（D. Greene）；mandar@isical.ac.in（M. Mitra）https://gdebasis.github.io/ （D. Ganguly）；http://derekgreene.com/ （D. Greene）；https://www.isical.ac.in/mandar-mitra（M. Mitra）0000-0001-9220-6652（S. Datta）；0000-0003-0050-7138 （D. Ganguly）；0000-0001-8065-5418（D. Greene）；0000-0001-9045-9971（M. Mitra）©2023 本文作者保留版权。允许根据知识共享署名4.0国际许可协议（CC 用。CEUR 研讨会论文集（CEUR-WS.org）

列表式QPP方法的一个主要缺点是，评估以相对方式进行，因此一个查询的性能是相对于其他查询来衡量的。然而，单个查询的下游性能估计也需要独立于其他查询进行评估。相比之下，逐点方法衡量各个查询的有效性，然后在需要时对完整查询集合上的结果进行聚合。这类似于通过计算各个查询的平均精度值，再对其进行聚合来衡量检索效果指标MAP。逐点评估还使我们能够对某一方法开展逐查询分析，这通常可以带来有价值的洞见。例如，Buckley [6]发现，通过开展广泛的逐主题检索分析，他们能够识别出大多数信息检索系统无法检索到相关文档的查询。然而，列表式评估方法不利于开展这种细致的逐查询分析。

列表式方法的另一个缺点是，它们可能对评估所用的配置设置过于敏感。其中最重要的两项配置是：i）目标检索评估指标，它会为查询集合确定真实排序；ii）用于获取排名前 $k$ 个文档以进行 QPP 估计的检索模型。事实上，这些配置的变化可能导致所报告的秩相关度量出现较大的标准差，也可能显著改变各种 QPP 系统的相对排名 [7]。为了解决列表式方法的局限性，我们提出了一种新的 QPP 评估框架，即**聚合逐点绝对误差（APAE）**；研究表明，该框架不仅与现有列表式方法一致，而且对查询性能预测实验设置的变化也更加稳健。

## 2. 逐点 QPP 评估框架

**与列表式真实值的相关性** 在介绍我们提出的新 QPP 评估框架 APAE 之前，我们先引入所需的符号。形式上，QPP 估计是如下形式的函数： $\phi(Q,\ M_k(Q)) \mapsto \mathbb{R}$，其中 $M_k(Q)$ 是信息检索模型 $M$ 针对查询 $Q \in \mathcal{Q}$ 检索到的排名前 $k$ 的文档集合， $Q \in \mathcal{Q}$ 属于一个基准查询集。

对于列表式评估，针对每个 $Q \in \mathcal{Q}$，我们首先计算目标信息检索评估指标 $\mu(Q)$ 的值，以反映检索结果列表 $M_k(Q)$ 的质量。接下来，使用这些 $\mu(Q)$ 分数，为集合 $\mathcal{Q}$ 构建真实标注排序；换言之，按照查询的 $\mu(Q)$值由高到低（或由低到高）排列，即，

$$\mathcal{Q}_\mu = \{Q_i \in \mathcal{Q} : \mu(Q_i) > \mu(Q_{i+1}), \forall i = 1, \ldots, |\mathcal{Q}| - 1\}\} \tag{1}$$

同样，该评估框架也会生成查询的预测排序，此时查询按照查询性能预测的估计分数排序，即，

$$\mathcal{Q}_\phi = \{Q_i \in \mathcal{Q} : \phi(Q_i) > \phi(Q_{i+1}), \forall i = 1, \ldots, |\mathcal{Q}| - 1\} \tag{2}$$

随后，列表式评估框架会计算这两个有序集合 $\gamma(\mathcal{Q}_\mu,\ \mathcal{Q}_\phi)$，之间的秩相关，其中 $\gamma : \mathbb{R}^{|\mathcal{Q}|} \times \mathbb{R}^{|\mathcal{Q}|} \mapsto [0, 1]$ 是一种相关性度量，例如肯德尔 $\tau$。

**单个真实标注** 与列表式评估不同，后者的真实标注表现为一个有序查询集，逐点 QPP 评估涉及进行 $|\mathcal{Q}|$ 独立的比较

**表1**用于衡量变化的查询性能预测配置（查询性能预测方法、信息检索指标和模型）。

| | |
|---|---|
| 查询性能预测方法 | AvgIDF [8], Clarity [9], NQC [10], WIG [11], UEF(Clarity)、UEF(NQC)、UEF(WIG) [2] |
| 信息检索指标 | AP@100、nDCG@100、P@10、Recall@100 |
| 信息检索模型 | LMJM ($\lambda = 0.6$)、LMDir ($\mu = 1000$)、BM25 ($k,\ b$) $= (0.7,\ 0.3)$ |

比较。每次比较都在查询 $Q$'s 的预测 QPP 得分 $\phi(Q)$ 与其检索效果指标 $\mu(Q)$之间进行，即

$$\eta(\mathcal{Q}, \mu, \phi) \stackrel{\text{def}}{=} \frac{1}{|\mathcal{Q}|} \sum_{Q \in \mathcal{Q}} \eta(\mu(Q), \phi(Q)) \tag{3}$$

与秩相关 $\gamma$ 不同，这里 $\eta$ 是形式为 $\eta$： $\mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$ 的逐点相关函数。通常可以将 $\eta$ 视为一个距离函数的倒数，该函数用于衡量预测值与对应真实值的偏差程度。与真实值评估指标不同，大多数查询性能预测估计值（例如 NQC、WIG 等）并不限制在 $[0, 1]$ 内。因此，要使用距离度量，每个查询性能预测估计值 $\phi(Q)$ 都必须归一化至单位区间。随后， $\eta$ 可定义为 $\eta(\mu(Q)\ \phi(Q)) \stackrel{\text{定义为}}{=} 1 - |\mu(Q)\ - \phi(Q)/\aleph|$，其中 $\aleph$ 是一个足够大的归一化常数，以确保分母为正。

**选择用于逐点 QPP 评估的信息检索指标** 通常，无监督QPP估计器与目标信息检索指标无关 $\mu$。例如，NQC 分数可以被视为对平均准确率@100值的近似，但也可以解释为对任何其他指标（如 nDCG@20 或 P@10）的近似。因此，一个问题是：应使用哪个指标来计算公式 3中的各项相关性。当然，对于 $\mu$的不同选择（例如 AP 或 nDCG），结果可能会有显著差异。列表式QPP评估也存在同样的情况，相关报告见 [7]。为减小此类差异的影响，我们现在提出一种简单而有效的解决方案。

**与指标无关的逐点QPP评估** 对于一组评估函数 $\mu \in \mathcal{M}$（例如 $\mathcal{M} = \{$平均准确率@100、nDCG@20、 $\ldots\}$），我们使用聚合函数来计算 QPP 估计值相对于各指标的总体逐点相关（公式 3）。形式化地说，

$$\eta(Q, \mathcal{M}, \phi) = \Sigma_{\mu \in \mathcal{M}}(1 - |\mu(Q) - \phi(Q)/\aleph|), \tag{4}$$

其中 $\Sigma$ 表示聚合函数（并不表示求和）。具体而言，我们选择此类函数中最常用的几种作为 $\Sigma$ 的选项：‘最小值’、‘最大值’和‘平均值’，即 $\Sigma \in \{$ 平均值、最小值、最大值 $\}$。接下来，我们计算给定查询集合 $\mathcal{Q}$中这些值的平均值，也就是说，将方程 $\eta$（$Q$、 $\mathcal{M}$、 $\phi$）中的值代入方程 4，再代入方程 3的求和式中。

## 3. 实验

一个 QPP 实验环境 [7] 涉及三项配置选择：i) 用于预测查询相对性能的**查询性能预测方法**本身；ii) 用于获取**信息检索指标** 的

根据在一组由特定信息检索模型检索出的前 $k$ ($k = 100$ 个文档上测得的查询性能，得到真实排序）；iii) 一个特定的**信息检索模型**。表 1 总结了实验中使用的信息检索模型和指标，以及相关的超参数值。实验旨在探究以下两个关键研究问题：

- **研究问题1**：APAE 与标准列表式相关性指标是否一致?
- **研究问题2**：对于 QPP实验环境的变化，APAE的鲁棒性如何?

若**研究问题1** 的答案为肯定，则表明我们提出的指标 APAE 与现有用于 QPP评估的指标一致；若**研究问题2** 的答案为肯定，则说明由于 APAE 在不同实验设置下具有更高的稳定性，因此优于现有方法。

我们在 TREC Robust数据集上开展 QPP实验，该数据集包含249个主题。按照QPP实验的标准做法， [5, 12], 我们报告对数据随机选取的30组等规模训练测试划分进行聚合后的结果。每个划分中的训练集用于调整 QPP方法的超参数。

**列表式评估与逐点评估之间的一致性** 首先，我们考察 APAE 与三个标准列表式QPP评估指标的一致性：Pearson $r$、Spearman $\rho$ 和 Kendall $\tau$；以及一种逐点方法——缩放绝对秩误差（sARE）[13]。由于 sARE 是一种误差度量，我们计算 APAE 与 $1 -$ sARE度量之间的相关性（为简便起见，表 2中将其称为 sARE）。我们通过将聚合函数 avg、min 和 max 代入式 $\Sigma$ 4，来试验 APAE 的三个不同实例，分别记为$\eta_{\text{avg}}(\mathcal{M})$、$\eta_{\text{min}}(\mathcal{M})$ 和 $\eta_{\text{max}}(\mathcal{M})$。

表 2 中的结果**肯定地**回答了 RQ1。此处报告的每个数值对应于：按照通过某个标准指标（包括 $\tau$）计算的有效性对 QPP系统进行排序后，其相对排名与 APAE 之间的秩相关（Kendall $r$、 $\rho$、 $\tau$ 或 sARE）以及 APAE 之间的秩相关，即 $\eta_{\text{avg}}$（$\mathcal{M}$）、 $\eta_{\text{min}}$（$\mathcal{M}$）和 $\eta_{\text{max}}$（$\mathcal{M}$）。标准列表式指标与所提出的逐点指标之间较高的相关值表明，APAE 可以替代标准的列表式评估。值得注意的是，平均聚合函数取得了最佳结果，因此在后续实验中，我们使用 $\eta_{\text{avg}}$（$\mathcal{M}$）作为逐点评估指标。

**表 2**

我们提出的逐点评估指标 APAE 与标准的基于列表的指标——皮尔逊 $r$、斯皮尔曼 $\rho$、肯德尔 $\tau$ 和 sARE——之间的相关性。使用肯德尔 $\tau$ 计算每一对查询性能预测系统排名之间的秩相关（分别采用基于列表和逐点的度量进行评估）。较高的数值表明，逐点评估可以有效地替代标准的基于列表的度量，因为它们对不同查询性能预测方法有效性的相对排序相当相似。

| | $\eta_{\text{avg}}$（$\mathcal{M}$） | | | | $\eta_{\text{min}}$（$\mathcal{M}$） | | | | $\eta_{\text{max}}$（$\mathcal{M}$） | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $r$ | $\rho$ | $\tau$ | sARE | $r$ | $\rho$ | $\tau$ | sARE | $r$ | $\rho$ | $\tau$ | sARE |
| BM25 | 0.810 | 0.810 | 0.905 | 0.887 | 0.778 | 0.778 | 0.794 | 0.813 | 0.802 | 0.810 | 0.794 | 0.794 |
| LMDir | **0.905** | **0.810** | **0.905** | **0.887** | 0.778 | 0.794 | 0.794 | 0.810 | 0.769 | 0.782 | 0.794 | 0.796 |
| LMJM | 0.810 | 0.810 | 0.810 | 0.846 | 0.794 | 0.794 | 0.782 | 0.786 | 0.794 | 0.769 | 0.810 | 0.846 |

**表3**所提出的点式查询性能预测指标 APAE 相对于列表式方法，在不同信息检索指标和信息检索模型配对下的稳定性。红色单元格表示每组中的最低值，而每列中的最低值以粗体显示。

| 模型 | 指标 | AP@100 | 召回率@10 | 前100项召回率 | 归一化折损累积增益@10 | nDCG@100 |
|---|---|---|---|---|---|---|
| LMJM | 平均准确率 | 0.497 | 0.813 | 0.429 | 0.783 | 0.429 |
| BM25 | 平均准确率 | 0.897 | 0.722 | 0.722 | 0.793 | 0.793 |
| LMDir | 平均准确率 | 0.897 | 0.786 | 0.786 | 0.823 | 0.905 |
| LMJM | AP@100 | | **0.328** | 0.811 | 0.363 | 0.783 |
| BM25 | AP@100 | | 0.783 | 0.784 | 0.714 | 0.642 |
| LMDir | AP@100 | | 0.823 | 0.901 | 0.834 | 0.789 |
| LMJM | R@10 | | | 0.624 | 0.893 | 0.503 |
| BM25 | R@10 | | | 0.803 | 0.982 | 0.894 |
| LMDir | R@10 | | | 0.903 | 0.864 | 0.864 |
| LMJM | 前100项召回率 | | | | 0.852 | 0.804 |
| BM25 | 前100项召回率 | | | | 0.786 | 0.890 |
| LMDir | 前100项召回率 | | | | 0.738 | 0.738 |
| LMJM | 归一化折损累积增益@10 | | | | | 0.537 |
| BM25 | 归一化折损累积增益@10 | | | | | 0.904 |
| LMDir | 归一化折损累积增益@10 | | | | | 0.868 |

（a）不同信息检索目标指标对之间，7 个不同QPP系统相对排名的相关性。使用基线基于列表的指标——肯德尔 $\tau$——对 QPP系统进行评估。

| 模型 | 指标 | AP@100 | 召回率@10 | 前100项召回率 | 归一化折损累积增益@10 | nDCG@100 |
|---|---|---|---|---|---|---|
| LMJM | 平均准确率 | 0.904 | 1.000 | 0.715 | 1.000 | 0.792 |
| BM25 | 平均准确率 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| LMDir | 平均准确率 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| LMJM | AP@100 | | 0.905 | 0.811 | 0.669 | 1.000 |
| BM25 | AP@100 | | 1.000 | 1.000 | 1.000 | 1.000 |
| LMDir | AP@100 | | 1.000 | 1.000 | 1.000 | 1.000 |
| LMJM | R@10 | | | 0.603 | 0.905 | **0.542** |
| BM25 | R@10 | | | 1.000 | 1.000 | 1.000 |
| LMDir | R@10 | | | 1.000 | 1.000 | 1.000 |
| LMJM | 前100项召回率 | | | | 0.654 | 1.000 |
| BM25 | 前100项召回率 | | | | 1.000 | 1.000 |
| LMDir | 前100项召回率 | | | | 1.000 | 1.000 |
| LMJM | 归一化折损累积增益@10 | | | | | 0.649 |
| BM25 | 归一化折损累积增益@10 | | | | | 1.000 |
| LMDir | 归一化折损累积增益@10 | | | | | 1.000 |

(b) 与表 4a 类似，但 QPP 性能采用逐点方法 APAE 进行评估。与表 4a 的比较表明，对于信息检索指标的变化，QPP系统相对排名的一致性更好。

| 指标 | 模型 | LMJM (0.6) | BM25 (0.7,0.3) | BM25 (0.3,0.7) | LMDir (500) | LMDir (1000) |
|---|---|---|---|---|---|---|
| AP@100 | LMJM (0.6) | 0.826 | 0.904 | 0.819 | 0.714 | 0.895 |
| nDCG@100 | LMJM (0.6) | 0.780 | 0.694 | 0.695 | 0.759 | 0.759 |
| 前100项召回率 | LMJM (0.6) | 0.824 | 0.769 | 0.782 | 0.904 | 0.904 |
| AP@100 | BM25 (0.7,0.3) | | 0.703 | 0.712 | 0.904 | 0.823 |
| nDCG@100 | BM25 (0.7,0.3) | | 0.781 | 0.827 | 0.811 | 0.811 |
| 前100项召回率 | BM25 (0.7,0.3) | | 0.813 | 0.725 | 0.731 | |
| AP@100 | BM25 (0.3,0.7) | | | 0.903 | 0.785 | 0.785 |
| nDCG@100 | BM25 (0.3,0.7) | | | 0.897 | 0.786 | 0.786 |
| 前100项召回率 | BM25 (0.3,0.7) | | | 0.812 | 0.752 | 0.779 |
| AP@100 | LMDir (500) | | | | 0.887 | |
| nDCG@100 | LMDir (500) | | | | 0.901 | 0.895 |
| 前100项召回率 | LMDir (500) | | | | 0.889 | 0.901 |
| AP@100 | LMDir (1000) | | | | 0.901 | |
| nDCG@100 | LMDir (1000) | | | | 0.893 | |
| 前100项召回率 | LMDir (1000) | | | | 0.903 | |

(c) 此处衡量了不同信息检索模型对之间 QPP 系统相对排名的秩相关性。与表 4a一样，使用 $\tau$评估 QPP 系统。信息检索模型旁边的数字表示各自的参数。

| 指标 | 模型 | LMJM (0.6) | BM25 (0.7,0.3) | BM25 (0.3,0.7) | LMDir (500) | LMDir (1000) |
|---|---|---|---|---|---|---|
| AP@100 | LMJM (0.6) | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| nDCG@100 | LMJM (0.6) | 1.000 | 0.864 | 1.000 | 0.843 | 0.864 |
| 前100项召回率 | LMJM (0.6) | 1.000 | 0.864 | 1.000 | 1.000 | 1.000 |
| AP@100 | LMJM (0.6) | | 1.000 | 1.000 | 1.000 | 1.000 |
| nDCG@100 | LMJM (0.6) | | 0.914 | 1.000 | **0.813** | 0.914 |
| 前100项召回率 | LMJM (0.6) | | 1.000 | 1.000 | 1.000 | 1.000 |
| AP@100 | BM25 (0.7,0.3) | | | 1.000 | 1.000 | 1.000 |
| nDCG@100 | BM25 (0.7,0.3) | | | 1.000 | 1.000 | 1.000 |
| 前100项召回率 | BM25 (0.7,0.3) | | | 0.812 | 0.905 | 1.000 |
| AP@100 | BM25 (0.3,0.7) | | | | 1.000 | 1.000 |
| nDCG@100 | BM25 (0.3,0.7) | | | | 1.000 | 1.000 |
| 前100项召回率 | BM25 (0.3,0.7) | | | | 1.000 | 1.000 |
| AP@100 | LMDir (500) | | | | | 1.000 |
| nDCG@100 | LMDir (500) | | | | | 1.000 |
| 前100项召回率 | LMDir (500) | | | | | 1.000 |

（d）与表4c不同，此处的查询性能预测结果是通过 APAE 进行评估的（而非 $\tau$）。

**查询性能预测方法相对有效性的方差**为了研究**研究问题2**，我们考察查询性能预测情境变化（即不同的信息检索模型和目标指标）下查询性能预测系统排名的相对稳定性，同时比较列表式方法和逐点式方法（见表3）。举例来说，假设使用三种查询性能预测方法，即 AvgIDF、NQC 和 WIG，我们观察到，在以 AP@100 为相对测量指标时，LMDir 的排序为 $\tau$（NQC）$>\tau$（WIG）$>\tau$（AvgIDF）。我们预计

当选择不同的信息检索模型和目标信息检索指标时，也能观察到类似的排序，例如使用 BM25 和 nDCG@100。与之前的实验一样，我们通过 Kendall 的 $\tau$ 来测量总共七个 QPP 系统之间的排名相关性（见表1）。

## 4. 结论

不同于标准的列表式查询性能预测评估机制（该机制以查询的参考排名为依据，按照检索效果衡量整体排名相关性），我们提出了一种逐点评估方法，用于计算归一化查询性能预测分数与真实信息检索评估指标（例如 AP@100 或 nDCG@20）之间的相对差异。实验表明，所提出的指标与标准列表式方法具有较高的相关性，并且相比列表式评估指标，对查询性能预测实验设置的变化更稳健。因此，利用这一指标，可以按查询评估不同查询性能预测方法在下游任务中的有效性。

**致谢**。。第一作者和第三作者的研究获得了爱尔兰科学基金会（SFI）资助，资助编号为 SFI/12/RC/2289_P2。

## 参考文献

[1] Y. Zhou, W. B. Croft，排名稳健性：一种预测查询性能的新框架，载于：CIKM '06 论文集，2006年，第567–574页。[2] A. Shtok, O. Kurland, D. Carmel，利用统计决策理论和相关性模型进行查询性能预测，载于：SIGIR '10 论文集，2010年，第259–266页。[3] Y. Lv, C. Zhai，信息检索中的自适应相关性反馈，载于：CIKM '09 论文集，2009年，第255–264页。[4] R. Cummins，用于查询性能推断和预测的文档得分分布模型，《ACM 信息系统汇刊》32（2014年），2:1–2:28。[5] H. Zamani, W. B. Croft, J. S. Culpepper，利用来自多个信号的弱监督进行神经网络查询性能预测，载于：SIGIR '18 论文集，2018年，第105–114页。[6] C. Buckley，当前信息检索引擎为何失效，载于：SIGIR'04 论文集，2004年，第584–585页。[7] D. Ganguly, S. Datta, M. Mitra, D. Greene，查询性能预测有效性变化分析，载于：ECIR'22 论文集，2022年，第215–229页。[8] C. Hauff, D. Hiemstra, F. de Jong，检索前查询性能预测器综述，载于：CIKM '08 论文集，2008年，第1419–1420页。[9] S. Cronen-Townsend, Y. Zhou, W. B. Croft，预测查询性能，载于：SIGIR '02 论文集，2002年，第299–306页。[10] A. Shtok, O. Kurland, D. Carmel, F. Raiber, G. Markovits，通过查询漂移估计预测查询性能，《ACM 信息系统汇刊》30（2012年）。[11] Y. Zhou, W. B. Croft，网络搜索环境中的查询性能预测，载于：SIGIR '07 论文集，2007年，第543–550页。[12] O. Zendel, A. Shtok, F. Raiber, O. Kurland, J. S. Culpepper，信息需求、查询与查询性能预测，载于：SIGIR '19 论文集，2019年，第395–404页。[13] G. Faggioli, O. Zendel, J. S. Culpepper, N. Ferro, F. Scholer，改进的查询性能预测评估框架，载于：《信息检索进展》，2021年，第115–129页。