arXiv:2304.04057v1 [cs.IR] 8 Apr 2023

不完美的自动索引如何影响语义搜索性能?

Mengtian Guo,David Gotz,Yue Wang信息与图书馆科学学院北卡罗来纳大学教堂山分校教堂山，北卡罗来纳州，美国{mtguo, gotz,wangyue}@unc.edu

摘要——**健康领域的文档通常会使用受控词表中的语义概念（即术语）进行标注。随着这类文档的数量不断增加，标注工作越来越多地由算法完成。与人工相比，自动索引算法并不完美，可能会为文档分配错误的术语，从而影响后续包含这些术语的搜索任务。在这项工作中，我们旨在了解使用错误分配的术语对布尔语义搜索性能的影响。我们以 MeSH 术语和生物医学文献检索为案例研究。在由 MeSH 术语构成的真实世界布尔查询上，我们实现了多种自动索引算法，并发现：（1）概率逻辑比传统布尔逻辑更能处理错误分配的术语；（2）查询级性能主要受查询中性能最低的术语限制；（3）将少量人工索引与自动索引相结合，可以恢复出色的查询级性能。这些发现为自动索引的未来研究提供了重要启示。**

索引词—**自动索引、语义搜索、医学主题词、机器学习**

I. 引言

受控词表广泛用于搜索引擎，通过语义概念为非结构化内容建立索引。例如，美国国家医学图书馆（美国国家医学图书馆）开发医学主题词（医学主题词），用于在PubMed搜索引擎中索引和搜索生物医学信息。其他医学搜索引擎，如 SemEHR [29]和Thalia [24]，将统一医学语言系统（UMLS）中的术语分配给文档，并将这些术语用作搜索分面。放射学图像搜索引擎GoldMiner [9]将SNOMED-CT和医学主题词分配给图像文档，以便于图像搜索。

生物医学信息和临床文本的快速增长，使得为每篇文档手动分配受控词汇术语变得不可行。2022年，超过130万篇文章被添加到PubMed索引中 [15]。同年，美国国家医学图书馆宣布，所有PubMed文章均已通过名为Medical Text的自动方法完成索引

索引器 (MTI) [21], , 其中只有一小部分由人工专家审核和整理 [2]。

然而，最先进的自动索引方法的准确性仍无法与人工索引相匹敌 [16]。在专注于医学主题词索引的 BioASQ 2022 竞赛中，最佳系统的微平均 F1 分数低于 0.75 [17], , 与美国国家医学图书馆在同一竞赛中的系统持平或更优。

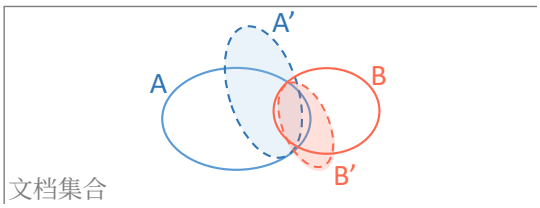


图 1：韦恩图展示了两个术语 A，B 对应的真实文章集合，以及相应的机器分类文章集合 A'，B'。这两个术语级分类器都不准确 ($A \neq A'$，$B \neq B'$)。如果查询为 A AND B ($A \cap B \approx A' \cap B'$)，查询级性能良好；但如果查询为 A OR B ($A \cup B \neq A' \cup B'$)，则表现较差。

事实上，美国国家医学图书馆知道自动索引可能存在错误 [2]。索引错误的文章可能导致令人不满意的搜索结果。例如，在系统综述中，研究人员会编写布尔查询，以便在搜索查询中精确定义纳入与排除标准；这些查询通常包含 MeSH 术语，用于指定感兴趣的主题。自动索引算法在为文章分配 MeSH 术语时可能出错，导致相关文章被漏掉，而无关文章被纳入搜索结果。

鉴于语义搜索引擎采用自动索引器已是不可避免的趋势，我们希望回答以下问题：**当查询是索引词的布尔组合时，不准确的自动索引结果将如何影响语义搜索性能?** 对于搜索引擎开发者而言，在将自动索引技术部署到生产环境之前，系统地衡量其对搜索质量的影响至关重要。对于搜索引擎用户而言，他们通常希望术语能够被准确地分配给文档。随着语义搜索引擎将索引过程自动化，开发者应了解这种期望在多大程度上仍能实现。

上述问题的答案可能并非显而易见。举例来说，请看图 1 中的示例。在这里，两个术语级分类器（自动索引器）单独使用时效果不佳。然而，将它们用于合取布尔查询时，得到的文章集合与真实集合非常吻合。与此同时，析取查询得到的文章集合与真实集合严重不符。这个例子

表明，不准确的术语级索引器可能会，也可能不会导致较差的查询级结果。其影响不仅取决于术语级分类器的准确性，还取决于查询中术语的逻辑组合方式。

在本研究中，我们采用数据驱动方法，探究自动索引对语义搜索质量的影响。具体而言，我们研究了医学主题词索引和生物医学文献检索。我们使用一组由医学主题词组成的真实搜索查询，将自动分配医学主题词的搜索结果与人工分配医学主题词的搜索结果进行了比较。我们考察了两种自动索引方法：将医学主题词作为二元标签（存在或不存在）分配，或作为概率标签分配（每个术语都关联一个预测概率）。相应地，搜索查询分别通过布尔逻辑或概率逻辑实现。我们对不准确索引如何影响搜索结果进行了分析，主要发现如下：

- 将术语作为概率标签分配，比将术语作为二元标签分配能够带来更好的检索性能。保留每个术语的预测概率，能够在后续搜索和排序流程中纳入术语级不确定性信息。
- 查询级检索性能并不与术语级分类性能呈线性关系。相反，我们观察到了“桶效应”：查询级性能受查询中表现较差的术语所限制，而这些术语往往是语料库中的稀有术语。
- 通过将索引工作分配给人和机器，可以实现出色的查询级性能（F1 值为 $> 95\%$）：仅对机器相对容易预测的术语应用自动索引（F1 值为 $> 80\%$），并对其余（罕见且机器难以预测的）术语进行人工索引。

我们的研究揭示了自动索引未来研究的若干启示。研究表明，预测性能最低的索引词是语义搜索性能的瓶颈。因此，自动索引研究亟须重点优化罕见术语。同时，在这些术语上投入人工审核工作，或可显著提升检索性能。

II. 相关工作

据我们所知，术语级索引性能与查询级检索性能之间的关系尚未得到充分研究。以往的工作通常将索引问题视为多标签分类或信息抽取问题，从术语或概念层面评估自动索引技术 [8, 19] 。很少有研究在搜索任务的背景下评估自动索引系统。Kim 等人早期的一项研究 [11]考察了自动分配的医学主题词与人工分配的医学主题词对检索结果的影响。他们使用一组非布尔查询来衡量这种影响。然而，医学主题词在实践中往往用于布尔查询 [13]。

人们已专门致力于从文献中提取生物医学概念，这通常被称为语义标注或语义索引。在一项专门的竞赛中，BioASQ 任务 A [26],提出了多种方法，用于自动为生物医学出版物标引医学主题词 [32, 31]。这类研究通常将语义索引系统视为极端多标签分类问题来评估。

自动医学主题词索引问题与信息抽取密切相关。信息抽取是一个活跃的研究领域，专注于从非结构化数据中提取有价值的信息。信息抽取服务对生物医学知识管理系统颇具吸引力，因为它们能够从非结构化文本中生成结构化且无歧义的数据，从而实现更有效、更高效的检索 [7]。因此，该技术已得到广泛研究，并应用于各种系统和场景，例如电子健康记录系统 [29], 图像检索 [8],和生物医学文献检索 [25]。传统上，机器抽取的术语不会直接呈现给用户，而是由搜索引擎在内部使用，以理解搜索查询并提升搜索质量 [23]。近年来，PubMed 等系统开始允许终端用户通过机器分配的索引词 [2]直接表达其检索和筛选条件。

III. 方法

A. 问题形式化

文档索引。设所有医学主题词的集合为 $H = \{h_1, \cdots, h_n\}$。索引过程将 H 中的术语分配给文档 d。对于 d，真实标签/人工术语分配向量为 $s = (s_1, \cdots, s_n)$。$s_i = 1$ 表示将 h_i 分配给了 d，而 $s_i = 0$ 则表示未分配。

自动索引算法（多标签分类器或索引词排序器）将术语分配给 d，其形式可以是二元标签或概率标签。二元索引算法为 d分配一个二元向量 $b = (b_1, \cdots, b_n)$。$b_i \in \{0, 1\}$ 表示该算法是否预测 h_i 属于 d。概率索引算法为 d分配一个概率向量$p = (p_1, \cdots, p_n)$。$p_i \in [0, 1]$ 表示 h_i 属于 d的预测概率。

使用索引词的布尔查询。布尔查询 q 是 MeSH 术语的组合，用于表达相关性标准。布尔组合是一个函数 $f_q(s)$，该函数接收术语分配向量 s 并输出二元相关性值。如果 d 的真实术语分配向量为 s，则 $f_q(s) \in \{0, 1\}$ 表示 d 是否与 q相关的真实标签。表 I 展示了一个检索主题及其对应的形式化表达。

当人工索引被自动索引取代时，我们可以根据自动分配给 d的术语，估计 d 对查询 q 的相关性。对于生成二元向量 b的二元索引算法，$f_q(b) \in \{0, 1\}$ 是用于判断 d 是否与 q 相关的估计二元相关性。对于生成概率向量p的概率索引算法，我们可以使用布尔函数的概率版本，来估计 d 与 q 相关的概率。我们将此概率逻辑函数称为f'_q。在本文中，我们使用乘积 t-范数 [5] 来实现概率逻辑。表 I 的最后一行展示了这样一个示例。

表 I：示例说明如何将检索主题表示为基于 MeSH 术语的布尔逻辑或概率逻辑。“$\oplus$”表示逻辑 OR，“$\odot$”表示逻辑 AND。

检索主题	“COVID-19 传播”
布尔检索查询	q =(“COVID-19” OR “SARS-CoV-2” OR “冠状病毒”) AND “传染病传播”
q中的 MeSH 术语	h_1 = “COVID-19” (D000086382), h_2 = “SARS-CoV-2” (D000086402), h_3 = “冠状病毒”(D017934), h_4 = “传染病传播” (D018562)
布尔逻辑函数 f_q	$f_q(b) = (b_1 \oplus b_2 \oplus b_3) \odot b_4$
概率逻辑函数 f'_q	$f'_q(p) = [(p_1 + p_2 - p_1 \cdot p_2) + p_3 - (p_1 + p_2 - p_1 \cdot p_2) \cdot p_3] \cdot p_4$

B. 索引算法

我们将自动索引实现为多标签分类任务，其中文档内容（标题和摘要）作为输入，一组相关的 MeSH 术语作为输出。我们训练了两种 MeSH 术语分类器：逻辑回归（LR）和双向变换器编码器表示（BERT）。较简单的逻辑回归分类器使用带有 TF-IDF权重的词袋特征。我们为每个 MeSH 术语训练了一个二元逻辑回归分类器。更复杂的 BERT 分类器通过微调 BioBERT [12]获得，其 Sigmoid单元的数量与 MeSH 术语的数量相同。两种分类器都会为每个术语生成一个预测概率向量。我们将预测概率以 0.5 为阈值进行二值化，从而获得每个术语的二元值向量。

这两个分类器为查询级别的文档搜索提供了四种算法。如果它们对每个术语的输出都是概率值（阈值处理前），我们就使用概率逻辑组合这些概率，得到 $[0,1]$ 中的查询级别相关性得分。我们将这两种算法称为**概率**$_{\text{LR}}$ 和**概率**$_{\text{BERT}}$。如果它们对每个术语的输出都是二元值（阈值处理后），我们就使用布尔逻辑组合这些二元值，得到$\{0,1\}$中的查询级别二元相关性得分。我们将这两种算法称为**二元**$_{\text{LR}}$ 和**二元**$_{\text{BERT}}$。

作为比较，我们还训练了以文档 d 作为输入并直接预测查询级真实标签 $f_q(s)$ 的分类模型，而不预测术语级标签向量 s。在这种方法中，模型可以采用端到端的方式进行优化，以直接提升查询级性能，因此可能优于上述流水线模型 [27]。因此，我们将此方法作为比较对象。对于每个查询，我们分别训练 LR 和 BERT 模型，以预测文档与查询相关的概率。我们将这些算法**查询**$_{\text{LR}}$ 和**查询**$_{\text{BERT}}$称为。

概率$_{\text{LR}}$、**概率**$_{\text{BERT}}$、**查询**$_{\text{LR}}$和**查询**$_{\text{BERT}}$ 会在 $[0,1]$ 中输出连续分数。因此，它们既可以对给定查询的文档进行分类，也可以对其排序。**二元**$_{\text{LR}}$ 和**二元**$_{\text{BERT}}$会输出二元值，只能对给定查询的文档进行分类，而不能排序。

我们没有使用美国国家医学图书馆的 MTI算法或来自 BioASQ 竞赛的其他算法，因为我们的首要目标是研究在布尔语义搜索中使用不完美分配的术语所产生的影响。这些算法在不完美索引性能方面，与我们在研究中训练的模型没有实质性差异。因此，我们转而训练了我们的自有模型以便开展实验和比较。

C. 数据集构建

文档索引语料库。我们使用来自 BioASQ 2021任务A [18]的一组规模庞大且具有代表性的生物医学文献及其 MeSH 术语标注，对上述逻辑回归和 BERT 模型进行了训练和评估。该语料库包含 15,559,157 篇 PubMed 文章，含有标题、摘要、专家标注的MeSH术语以及其他元数据。语料库涵盖 29,369 个独有的 MeSH 术语。每篇文章平均包含 12.68 个 MeSH 术语。

布尔查询集。为评估自动索引对搜索结果质量的影响，我们使用了一组为系统综述构建的布尔查询。这些查询来自三个来源：(1) CLEF-TAR 2019 [10],(2) 系统综述更新数据集 [1], (3) PubMed COVID- 19 文章过滤器 [22]。由于此处的目标是获取由 MeSH 术语组成的布尔查询，而非执行相同的系统综述任务本身，我们对查询进行了如下处理。首先，删除查询中的所有非 MeSH 术语，使每个查询都仅由 MeSH 术语进行逻辑组合。其次，如果某个术语前有 exp（爆炸）运算符，我们只保留该术语本身，不将其扩展为医学主题词层级中的下位词。这是因为此类扩展只会细化术语的范围，而不会增加查询的语义复杂度。最后，我们删除了在整个语料库中相关文档少于 50 篇的查询，以便训练效果较好的机器学习模型。最终的查询集包含 27 个查询，涉及 183 个不同的 MeSH 术语。每个查询都使用逻辑运算符（AND、OR、NOT）组合一组 MeSH 术语。所有查询及其来源参考信息均可在在线附录中找到。[1]

D. 实现细节

我们将数据集划分为训练集（89%；13,847,650 篇文章）、验证集（9%；1,400,324 篇文章）和测试集（10%；1,555,915 篇文章）。我们从训练集中随机抽取了一个包含 500K个文档的训练子集。无论采用哪种索引方法，类别分布不平衡都是分类器训练过程必须应对的问题。这是因为，在整个文档语料库中，只有极少数文档与任意给定的医学主题词或

[1]https://bit.ly/3mKmKIp

查询。这种类别分布不平衡通常给机器学习算法带来挑战。为缓解这一挑战，我们针对罕见的 MeSH术语和稀疏查询进行过采样，将分配有罕见术语或与稀疏查询相关的额外文档添加到训练子集中。我们为实验构建了这样一个包含 514,686 篇文章的训练子集。为进行公平比较，所有分类器均使用同一组文章进行训练。我们发现，使用更大的训练子集并不能显著改进医学主题词预测器。

请注意，在这个子集中，任何类别的正例仍然是少数。在正例与负例分布不平衡的数据集上训练机器学习分类器时，分类器往往会偏向多数类（负类），不愿预测少数类（正类）。为缓解这一问题，我们采用了代价敏感损失 [6] 和焦点损失[14]等特殊损失函数，以抵消数据不平衡的影响。有关详细的数据采样流程 A、超参数设置 B 以及术语级分类性能 (C)，请参阅附录。

E. 查询级评估指标

预测文章相对于查询的相关性，既可以视为分类任务（相关性的二元概念），也可以视为排序任务（相关性的相对概念）。为了评估分类性能，我们使用了精确率、召回率和 F1 分数，并对全部27 个查询取平均值。为了评估排序性能，我们使用了平均准确率均值（MAP）、P@10 精确率、P@50精确率和 R@1000 召回率，并对全部 27 个查询取平均值。对于**Prob** 和**Query** 方法，我们通过寻找在验证集上使F1分数最大化的阈值，将连续的相关性分数转换为二元分类决策；该方法由 You 等人提出 [32]。由于**Binary** 方法只能生成二元相关性标签，因此我们不评估其排序性能。

IV. 结果与分析

本节将深入分析第 III-B 节所述各种自动索引和排序算法的性能。除了比较不同算法外，我们还观察到不同查询之间存在较大的性能差异（IV-A）。这促使我们进一步分析术语级性能与查询级性能之间的相关性（IV-B），并考察具体案例（IV-C）。最后，我们分析一种混合索引场景：算法仅分配能够准确预测的术语，而对于算法难以预测的术语，则由人工分配（IV-D）。

A. 定量性能评估

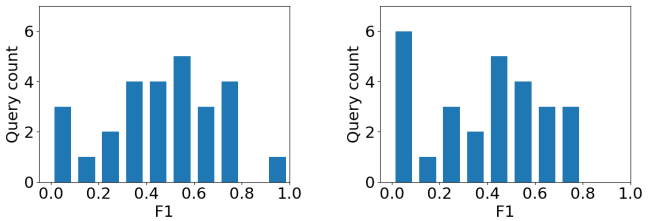
我们在表II 中报告了两种语义索引方法（**二值**$_{\text{LR/BERT}}$；**概率**$_{\text{LR/BERT}}$）和比较方法（**查询**$_{\text{LR/BERT}}$）的查询级性能。为了便于比较使用相同基础分类器的方法，我们将逻辑回归结果和 BERT 结果列为相邻行。全部指标在27个查询上取平均值。我们使用双侧随机化检验 [3]，并将第一类错误水平设为 $\alpha = 0.05$，对各项指标进行了相互比较。我们旨在了解使用自动索引方法时检索性能下降了多少。我们还希望比较不同索引方法、不同基础模型以及在不同搜索查询上评估时的性能差异。

由于我们将使用人工索引检索到的文档集视为基准真值，因此人工索引的精确率、召回率和 F1分数都应为 1。0。总体而言，所有实验中的索引方法都无法达到人工索引的效果。**Prob**$_{\text{BERT}}$的 F1分数最高，同时精确率最高，召回率也不错。**Binary**$_{\text{LR}}$ 以精确率较低为代价，获得了最高的召回率（尽管并未显著优于其他召回率结果）。**Prob**$_{\text{BERT}}$ 在全部四项指标中排名表现最佳（不过这些指标均未显著高于**Prob**$_{\text{LR}}$的对应指标）。

比较 二元和**概率**方法时，在使用逻辑回归作为基础分类器的情况下，除了召回率之外，**概率**方法在大多数指标上的表现至少与二元方法一样好。使用 BERT 作为基础分类器时，**概率**方法在精确率方面展现出优势。将 F1 作为衡量模型整体性能的指标时，使用逻辑回归时**概率**方法与二元方法表现相近，而使用 BERT 时则显著优于二元方法。**概率**方法还具有生成文档排名的额外优势，因此在自动索引方面比二元方法更具吸引力。

比较逻辑回归和 BERT，二元$_{\text{LR}}$ 与 二元$_{\text{BERT}}$之间没有显著差异。将其实现为 **概率型** 索引器时，**概率型**$_{\text{BERT}}$ 的查询级精确率和 F1 显著更高。**概率型**$_{\text{BERT}}$ 与 **概率型**$_{\text{LR}}$的召回率和排序指标相近。

在大多数指标上，词项级索引方法（二元$_{\text{LR/BERT}}$；**概率**$_{\text{LR/BERT}}$）的表现明显优于查询级分类器（**查询**$_{\text{LR/BERT}}$）。这意味着，当决策逻辑预先指定时，端到端训练未必优于组件级训练。因此，词项级索引不仅因其可扩展性和透明度而适用于语义搜索，也因其性能表现而成为一种良好策略。



(a) **概率**$_{\text{BERT}}$　　(b) **二进制**$_{\text{BERT}}$

图2：两种方法中的F1分数分布。

查看图2中的性能分布，性能下降幅度因查询而异。有些查询受自动索引的影响较小，F1 值达到

表 II：自动索引和排序算法的查询级平均性能。所有指标均为越高越好。使用相同基础分类器（逻辑回归或 BERT）的算法排列在相邻行中。在同一组行的任一列中：标注有‘*’的结果显著优于未标注‘*’的结果；如果两个结果均标注有‘*’，则它们之间没有显著差异。在同一列的所有行中：如果某个结果以**粗体**显示，则该结果显著优于该列中的任何其他结果；否则，该列中所有标注有‘*’的结果之间均无显著差异。（随机化检验，显著性水平 $\alpha = 0.05$）

方法	P	R	F1	MAP	P@10	P@50	R@1000
二元$_{\text{LR}}$	0.310*	0.625*	0.390*	-	-	-	-
概率$_{\text{LR}}$	0.325*	0.584	0.395*	0.415*	0.600*	0.459*	0.791*
查询$_{\text{LR}}$	0.241	0.543	0.305	0.314	0.507	0.401	0.745
二进制$_{\text{BERT}}$	0.358	0.553*	0.391	-	-	-	-
概率$_{\text{BERT}}$	**0.427***	0.554*	**0.457***	0.452*	0.611	0.481	0.810*
查询$_{\text{BERT}}$	0.389	0.434	0.389	0.397	0.530	0.471	0.711

0.9016。有些查询在使用自动索引时性能大幅下降，F1 值为 0.0。这促使我们分析影响查询性能的词项级因素，以及下面的案例研究。

B. 术语级性能与查询级性能之间的相关性

从图2中的查询性能分布可以看出，自动索引对不同查询的影响各不相同。这种差异可能有两个来源。首先，查询中使用的 MeSH 术语的术语级性能不同，从而导致查询级性能存在差异。其次，各查询的语义结构和复杂程度不同（例如，使用的 MeSH 术语数量不同，且 MeSH 术语的组合方式也不同）。为了解造成差异的因素，我们分析了查询的搜索结果质量与以下因素之间的相关性：（1）查询中所用 MeSH 术语的汇总术语级性能；（2）描述查询复杂程度的查询特征。分析中我们采用了**Prob** $_{\text{BERT}}$，因为在所有考察的方法中，它的整体性能最佳。

1）术语级性能与查询级性能之间的相关性：我们使用 F1指标表示医学主题词索引器的性能。我们计算了查询分类性能与该查询中所用 MeSH 术语的汇总性能之间的皮尔逊相关系数（即最差/平均/最佳术语级 F1、出现频率最高的 MeSH 术语的频率、出现频率最低的 MeSH 术语的频率，以及该查询中的 MeSH 术语数量）。结果如表III所示。

对于查询性能与 MeSH 术语性能之间的相关性，如果所有 MeSH 术语中 F1 最低的术语具有较高的 F1 值，则查询的召回率显著更高。此外，所有 MeSH 术语的平均 F1 值越高，查询的召回率和精确率也越高（具有中等显著性）。所有 MeSH 术语中 F1 最高的术语与查询的精确率和召回率之间不存在显著相关性。这表明，查询的召回率受到查询中性能最差的 MeSH 术语索引器的限制。为了提升查询检索性能，我们应重点关注性能最低的 MeSH 术语以及 MeSH 术语的平均性能。查询中有一个表现良好的 MeSH 术语，并不足以带来良好的查询级检索性能。

对于查询性能与医学主题词频率之间的相关性，我们观察到，查询的召回率与频率最低的医学主题词呈正相关。因此，如果查询中包含一个罕见的医学主题词，该查询的召回率更有可能较低。我们发现，医学主题词的F1分数与其在训练集中的频率呈正相关（皮尔逊相关系数：0.218，p值：0.003）。因此，罕见医学主题词的F1分数可能会因训练实例不足而较低，从而降低该查询的检索性能。因此，自动索引系统应更加努力地提升罕见医学主题词的性能。

2) 查询复杂度与查询级性能之间的相关性：查询的召回率与查询中使用的 MeSH 术语数量呈负相关。一种可能的解释是，包含更多 MeSH 术语的查询更可能包含性能较差的 MeSH 术语。我们发现，查询中 MeSH 术语的数量与其最差的 MeSH 术语级 F1 值之间存在较强的相关性（皮尔逊相关系数：-0.547，p-值：0.003）。因此，查询包含的 MeSH 术语越多，就越容易受到表现不佳的 MeSH 索引器的影响。

查询级性能与查询中 AND 的数量之间不存在显著相关性。我们发现，MeSH 术语的数量与查询中 OR 的数量高度相关（皮尔逊相关系数：0.977，p 值：0.000），这表明，当一个查询涉及更多 MeSH 术语时，这些 MeSH 术语大多通过 OR 连接。这也解释了为什么查询级性能与查询中 OR 的数量之间呈负相关。

C. 案例研究

为了了解检索性能下降如何影响用户搜索，我们下面使用**Prob**$_{\text{BERT}}$ 方法提供三个案例研究。我们选取了三个 **Prob**$_{\text{BERT}}$ 表现各异的查询：一个查询的$F1 = 0.7513$，一个查询的 $F1 = 0.4583$，以及一个

表III：查询级性能（精确率和召回率）与不同术语级特征之间的皮尔逊相关系数。“*”表示显著相关（$p < 0.05$），“†”表示中等显著相关（$0.05 \leq p < 0.1$）。

	最差 F1	平均 F1	最佳 F1	最高频率。	最低频率。	术语数量	AND 数量	OR 数量
查询级精确率	0.290	0.370†	0.321	0.196	0.140	-0.226	-0.137	-0.222
查询级召回率	0.387*	0.366†	-0.033	-0.139	0.403*	-0.378†	-0.144	-0.377†

的 $F1 = 0.0$。设想一位临床医生在包含 1,555,915 篇文章的测试集中发出这三个搜索查询。我们比较了使用**Prob**$_{\text{BERT}}$ 检索到的文章子集，以及使用人工提供的真实 MeSH 标签检索到的文章子集。使用人工标注的 MeSH 术语检索到的文章集被视为相关文章，未检索到的文章则被视为不相关文章。

1) 案例 *1*：对于案例研究 1，我们考察了与锁骨骨折相关的查询。详细查询如下。对于该搜索查询，**Prob** $_{\text{BERT}}$ 的精确率为 0.7184，召回率为 0.7872。如果临床医生在测试集中执行该搜索查询，**Prob**$_{\text{BERT}}$ 会检索到 103 篇文章，同时漏掉 20 篇相关文章（例如 PMID 32295588、PMID 28159682、PMID 27765500）。例如，PMID 28159682（“用于小骨骨折治疗的可注射拟骨热凝胶的开发”）介绍了一项关于研发药物的研究，该药物旨在治愈锁骨骨折等小骨骨折。在检索到的 103 篇文章中，有 29 篇不相关（例如 PMID 30948397、PMID 31677623、PMID 31415405）。例如，PMID 30948397（“成人桡骨远端骨折中的皮肤皱褶征”）提到，皮肤皱褶与锁骨骨折有关。然而，该文章介绍的是桡骨骨折病例，而非锁骨骨折。PMID 31677623 和 PMID 31415405 均标注了 锁骨，并且都被赋予了 内固定术，而非 骨折固定术。尽管它们与给定的搜索查询无关，但由于这两个术语存在关联，它们仍可能符合搜索者的兴趣。

1. 锁骨/2. 骨折，骨/或骨折固定/或骨折愈合/3.1 和2 4. 动物/ 5. 人类/ 6. 4 非 5 7. 3 非 6

2）案例 *2*：对于案例研究 2，我们考察了与婴儿败血症相关的查询。对于该搜索查询，**Prob**$_{\text{BERT}}$ 的精确率为 0.3738，召回率为 0.5923。如果临床医生在测试集中发出该搜索查询，**Prob**$_{\text{BERT}}$ 将检索到 618 篇文章，同时漏掉 159 篇相关论文（例如 PMID 375620、PMID 918581、PMID 879115）。在检索到的 618 篇文章中，有 387 篇与搜索查询无关（例如 PMID 6798907、PMID 6787924、PMID 6783064）。

1. 脓毒症/ 和 婴儿，新生儿/

3）案例 *3*：对于案例研究 3，搜索查询与肾功能不全的营养治疗有关。对于该搜索查询，**Prob**$_{\text{BERT}}$ 的精确率和召回率均为零 。因此，使用自动索引时，14 篇相关文章中没有一篇能够被检索到（例如 PMID 23123667、PMID 22177826、PMID 20708074）。**Prob**$_{\text{BERT}}$ 检索出 68 篇假阳性论文（例如 PMID87126、PMID 113851、PMID104103）。这三篇论文均标注了“Amino Acids”和其他可能与营养相关的 MeSH 术语，但没有一篇标注了“Renal Insufficiency”（该查询中表现最差的术语；F1 = 0.0057）。

1. 肾功能不全/ 2. 营养治疗/ 3. 营养需求/ 4. 能量摄入/ 5. 输注，肠外/ 6. 胃造口术/ 7. 膳食蛋白质/ 8. 氨基酸/ 9. 葡萄糖/ 10. 或/2-9 11. 1 和 10

D. 在人类与机器之间分配索引工作

我们上述的分析表明，自动索引中表现不佳的术语是查询级检索性能的主要瓶颈。它们在训练集中的低出现率使得机器学习模型难以对其进行预测。这一方面，尽管人工索引能够确保较高的检索性能，但无法扩展到大型文档集合。在本节中，我们旨在寻找完全自动索引与完全人工索引之间的平衡。我们设想一种混合索引场景：仅使用算法分配能够被准确预测的术语，而对于算法表现不佳的术语，仍然依靠人工进行分配。

为模拟这种混合索引情形，如果**Prob**$_{\text{BERT}}$ 未能达到高于阈值 t 的术语级 F1分数，我们就用人工标注的真实术语替换机器分配的术语。替换后，我们重新评估平均查询级性能。当 $t = 0$时，不替换任何机器分配的术语，查询级 F1 与表 II 中报告的全自动结果相同（P =0.427，R = 0.554）。另一方面，当 $t = 1$时，所有术语均为真实标注，查询级性能达到完美水平（P = R = 1.0）。随着我们将 t 从 0 增加到 1，183 个医学主题词中表现欠佳的子集将逐渐扩大，并被真实标注替换；我们

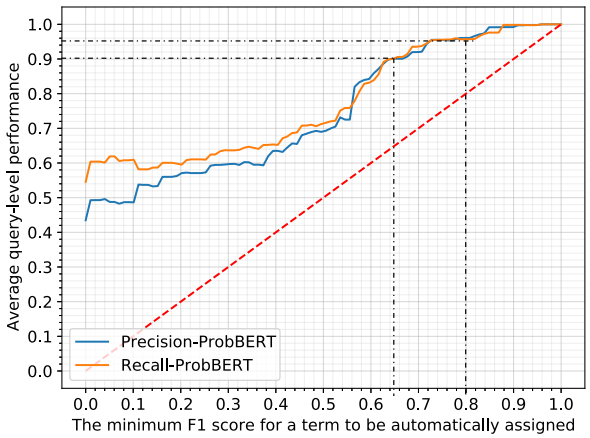


图3：将不准确的自动索引结果替换为真实值/人工索引结果时的平均查询级性能。曲线上的点（x, y）表示，当F1分数为 $\geq x$的术语由自动方式分配、F1分数为 $< x$ 的术语由人工方式分配时，平均查询级性能为 y 。两条虚线分别表示：当F1分数为 $\geq$ 0时，术语由自动方式分配，其余术语由人工方式分配，平均查询级精确率/召回率为0.9（0.95）；F1分数为.65（0.8）时，平均查询级精确率/召回率为0.9（0.95）。

预计查询级性能会呈现单调上升趋势。该趋势如图 3 所示。

图 3 显示，为了使查询级精确率/召回率达到 0.9（0.95）以上，可以使用自动索引为 F1 值分别高于 0.65（0.8）的术语赋值。对于 F1 值低于 0.65（0.8）的表现不佳术语，则由人工赋值；这些术语分别占医学主题词术语实例的 15.87%（21.33%）。这些比例相对较小的一个解释是，这些表现不佳的术语也属于罕见术语。这意味着，如果自动索引方法负责（1）将预测性能高的术语直接赋予文档，以及（2）推荐表现不佳但可能相关的术语，以协助人工做出索引决策，那么索引算法不必完美，也能有效提升语义搜索性能。

这项分析有两方面的用途。首先，对于 MeSH 术语索引算法的技术研究（例如 BioASQ 大规模索引挑战），当索引工作量由算法和人工专家分担时，这项分析能够回答“达到什么程度才算足够好”。从性能评估的角度来看，分析强调，与平等重视所有术语相比，提高对那些难以处理的术语的预测准确性更为重要。其次，对于语义搜索服务提供商（例如 PubMed），这项分析展示了一种方法，可用于确定哪些索引词能够自动分配而不会严重影响基于索引词的检索性能，并量化人工审核／整理成本与查询级检索性能之间的权衡。

E. 局限性

上述分析的一个主要局限是，我们的实验仅使用了一个较小的查询集。理想的查询集应是提交给 PubMed 等真实搜索引擎、且包含 MeSH 术语的搜索查询日志。我们找不到此类公开的查询日志，因此构建了一个较小的查询集。我们认为，本文提出的分析方法在此类查询日志可用时会很有用。

另一个局限是，在文献检索中，MeSH 术语通常与自由文本词一起使用。我们的研究考虑的是仅包含 MeSH 术语、不含自由文本词的布尔检索，而这类检索在文献检索中的使用较少。然而，临床术语的布尔组合被广泛用于计算表型识别 [4] 和值集 [28] 中，以定义复杂的高层概念。

V. 结论与未来工作

在本研究中，我们旨在了解在语义布尔检索中使用自动分配术语所带来的性能影响。我们发现，概率逻辑能够更好地处理文档检索中预测术语的不确定性；查询级性能与查询中表现不佳的术语高度相关；并且，在自动索引中混入少量人工索引即可实现出色的查询级性能。尽管我们采用的是医学主题词和文献检索这一具体案例，但这些发现对未来关于自动索引的一般性研究具有启示意义。

将预测词项的不确定性纳入语义布尔检索：自动索引算法通常会以一定程度的不确定性分配索引词，这种不确定性通过分类器预测的概率或排序器生成的分数来表示（这些分数可以转换为预测概率 [20]）。随后，通过选取分数高于阈值的词项或选取排名前k 的词项，将词项作为二元标签分配给文档。此过程丢弃了机器预测中固有的不确定性。当这些分配不准确的词项用于布尔检索时，它们会被视为仿佛是由人工准确分配的，从而导致后续检索错误。我们对比二元标签索引算法与概率索引算法的结果表明，在后续布尔检索中，以概率逻辑替代布尔逻辑，将预测词项的不确定性纳入其中是有益的。按照概率排序原则，这种方法能够自然地生成文档的排序列表。

在自动索引算法训练和评估中强调低性能术语：我们的分析表明，低性能术语是“木桶中最短的木板”——它们是限制查询级性能的因素。因此，未来关于自动索引算法训练和评估的研究应特别关注这样一组术语：它们较为罕见，和/或仍然存在相对较低的预测性能。关于长尾类别的一条新兴研究路线一直在探索类似的挑战 [30]。例如，如果目标是请人工索引人员审核潜在相关候选术语列表，那么在前-k个术语中正确术语的高召回率可能是一项有用的指标。

在机器和人之间分配索引工作：我们主张对自动索引效果不佳的术语采用人工索引。在 2022 [16] 之前，美国国家医学图书馆的医学主题词术语索引就采用了这一策略。然而，在美国国家医学图书馆最近的文档中，尚不清楚人工审核/整理工作如何与“MTI-Auto”系统中的自动术语分配相结合 [2]。我们的研究表明，用人工索引补充自动索引，可以防止表现不佳的术语影响查询级性能。由于表现不佳的术语也较为罕见，因此与之相关的人工索引工作量相对于通过术语分配实例数量衡量的整体索引工作量而言较小。从长远来看，罕见术语的标签数量可能会不断增加。这有望改善这些术语的自动索引性能，而人工索引工作量也将逐渐减少。

参考文献

[1] Amal Alharbi 和 Mark Stevenson。“系统综述更新数据集”。载于：*ACM SIGIR 2019*。法国巴黎：ACM，2019 年 7 月，第 1257–1260 页。[2] 自动索引常见问题。网址：https://support.nlm.nih.gov/knowledgebase/article/KA-05326/en-us。[3] D. Basu。“实验数据的随机化分析：费希尔随机化检验”。载于：美国统计学会会刊 75.371（1980 年 9 月）。出版商：Taylor & Francis，第 575–582 页。[4] Mike Conway 等。“分析面向电子健康记录的表型算法的异质性与复杂性”。载于：*AMIA*年度研讨会论文集。第 2011 卷。美国医学信息学会。2011 年，第 274 页。[5] Petr H´ajek。《模糊逻辑的元数学》。Springer Science & Business Media，1998 年。[6] Haibo He 和 Ed wardo A. Garcia。“从不平衡数据中学习”。载于：*IEEE*知识与数据工程汇刊 21.9（2009 年 9 月）。会议名称：IEEE知识与数据工程汇刊，第 1263–1284 页。[7] Jelena Jovanovic 等。“文本内容的自动语义标注”。载于：*IT Professional* 16.6（2014 年 11 月）。会议名称：IT Professional，第 38–46 页。[8] C. E. Kahn 和 D. L. Rubin。“对图注进行自动语义索引以改进放射学图像检索”。载于：美国医学信息学会杂志16.3（2009 年 5 月），第 380–386 页。[9] Charles E. Kahn 和 Cheng Thao。“GoldMiner：放射学图像搜索引擎”。载于：美国伦琴学杂志 188.6（2007 年 6 月）。出版商：美国伦琴射线学会，第 1475–1478 页。[10] Evangelos Kanoulas 等。“CLEF 2019 实证医学技术辅助评审概述”。载于：*CEUR*研讨会论文集。第 2380 卷。2019 年。

[11] W. Kim、A. R.Aronson和 W.J. Wilbur。“自动 MeSH 术语分配与质量评估。”载于：*AMIA*研讨会论文集（2001年），第319–323页。[12] Jinhyuk Lee 等。“BioBERT：用于生物医学文本挖掘的预训练生物医学语言表示模型。”载于：（2019年1月）。[13] Carol Lefebvre 等。“检索和选择研究。”载于：干预措施系统综述*Cochrane*手册（2019年），第67–107页。[14] Tsung-YiLin等。“用于密集目标检测的焦点损失。”载于：（2017年8月）。[15] *MEDLINE PubMed* 生产环境统计。网址：https: //www.nlm.nih.gov/bsd/medline pubmed production stats.html。[16] James Mork、Alan Aronson 和 Dina Demner-Fushman。“12年后——美国国家医学图书馆医学文本索引器是否仍然有用且相关？”载于：生物医学语义学杂志8.1（2017年），第1–10页。[17] Anastasios Nentidis 等。“BioASQ2022年：大规模生物医学语义索引与问答挑战赛第十届。”载于：*ECIR 2022*年，挪威斯塔万格，*2022*年*4*月*10–14*日，论文集，第二部分。Springer。2022年，第429–435页。[18] AnastasiosNentidis等。“BioASQ2021年概览：大规模生物医学语义索引与问答挑战赛第九届。”载于：（2021年6月）。[19] Aur´elieN´ev´eol等。“使用平行语料库评估法语和英语 MeSH索引系统。”载于：*AMIA*年度研讨会论文集2005（2005年），第565–569页。[20] Alexandru Niculescu-Mizil 和 Rich Caruana。“利用监督式机器学习预测良好的概率。”载于：第*22*届国际机器学习会议论文集。2005年，第625–632页。[21] 美国国家医学图书馆医学文本索引器（*MTI*）。网址：https://lhncbc.nlm.nih.gov/ii/tools/MTI.html。[22] *PubMed*帮助页面。网址：http://pubmed.ncbi.nlm.nih. gov/help/。[23] Ridho Reinanda、Edgar Meij 和Maarten deRijke。“知识图谱：信息检索视角。”载于：（），第156页。[24] Axel J Soto、Piotr Przybyła 和Sophia Ananiadou。“Thalia：生物医学摘要语义搜索引擎。”载于：生物信息学 35.10（2019年5月），第1799–1801页。[25] MariaTaboada 等。“罕见病病例的自动语义标注：案例研究。”载于：2014（2014年），第13页。[26] GeorgeTsatsaronis等。“BIOASQ大规模生物医学语义索引与问答竞赛概览。”载于：*BMC*生物信息学16.1（2015年4月），第138页。[27] Fei Wang、Lawrence PeterCasalino和 Dhruv Khullar。“医学中的深度学习——前景、进展和

挑战”。载于：《*JAMA*内科医学》179.3（2019年3月），第293–294页。[28] Duwayne L Willett 等人。“利用电子健康记录数据共享临床状况定义的 SNOMED CT 概念层级结构”。载于：《应用临床信息学》9.03（2018年），第667–682页。[29] Honghan Wu 等人。“SemEHR：一款通用语义搜索系统，可从临床笔记中提取语义数据，以支持个性化护理、临床试验招募和临床研究*”。载于：《美国医学信息学会杂志》25.5（2018年5月），第530–537页。[30] Lin Xiao 等人。“头部标签是否有助于长尾多标签文本分类”。载于：《*AAAI* 人工智能会议论文集》。第35卷。第16期。2021年，第14103–14111页。[31] Guangxu Xun 等人。“MeSHProbeNet：用于 MeSH 索引的自注意力探针网络”。载于：《生物信息学》35.19（2019年10月），第3794–3802页。[32] Ronghui You 等人。“BERTMeSH：利用全文进行大规模高性能 MeSH 索引的深度上下文表征学习”。载于：《生物信息学》37.5（2021年5月）。Wren J onathan 编，第684–692页。

附录

A. 考虑不平衡数据分布

1) 对稀有概念进行上采样：组合式方法中的医学主题词分类器和单体式方法中的查询分类器面临类别分布不平衡的问题，因为在训练集中的正例数量始终比负例少几个数量级。类别分布不平衡通常会给机器学习算法带来挑战。为了训练高质量的分类器，我们采用了以下采样方法，以纳入更多稀有类别的实例。

假设一个子样本中总共有 M 个可能的标签和 N 个文档。如果所有标签都相互排斥，那么每个标签应对应 N/M 个文档，这称为标签的“公平份额”。如果某个类别 c所对应的文档数量小于“公平份额”的10%（即 $N_c < 0.1 \times \frac{N}{M}$），我们将其定义为稀有类别，其中 N_c表示与该类别相关联的文档数量。随后，我们对稀有类别进行上采样，使其达到10%的“公平份额”。在本研究中，M 是医学主题词和查询的总数（即183+ 27 = 210），N 是两个训练子集的大小（即500,000和1,000,000）。采样过程完成后，两个训练数据集的大小分别为514,686（1.03 × 500,000）和1,025,855（1.03 × 1,000,000）。我们将这两个数据集分别称为“子集-500K”和“子集-1M”。我们发现，在子集-1M上训练的模型并未显著优于在子集-500K上训练的模型。因此，正文中使用的是在子集-500K上训练的模型。

2）类别分布不平衡的损失函数：为 MeSH 和查询预测训练分类器的一项挑战，是训练数据集中正实例与负实例的分布不平衡。对于 MeSH 术语，正实例与负实例的比例从 1:0.5 到 1:11968 不等；对于查询，比例则从 1:254 到 1:10090 不等。我们在损失函数中为正类和负类应用了加权因子，以应对这一问题。

对于逻辑回归，我们使用了以下公式：

$$CE(p_t) = -\alpha_t \log(p_t) \tag{1}$$

$$\alpha_t = \frac{n_{\text{samples}}}{n_{\text{classes}} \times n_{\text{samples in class } t}} \tag{2}$$

其中，CE 表示交叉熵损失；p_t 表示真实类别（$p_0 = 1 - p_1$）的估计概率；α_t 表示通过类别概率倒数计算得出的加权因子。

对于 BERT，我们采用了焦点损失 [14]，因为它提高了训练过程的稳定性。以下公式描述了焦点损失：

$$FL(p_t) = -\alpha_t(1 - p_t)^{\gamma} \log(p_t) \tag{3}$$

其中，FL 表示焦点损失；γ 和 α_t 是通过经验设定的超参数，我们采用了默认值（$\gamma = 2, \alpha_t$ =0.25），该值在提出焦点损失的论文中被证明效果最佳。

B. 模型超参数设置

逻辑回归。我们通过scikit-learn训练了逻辑回归模型，使用了L2正则化、LBFGS优化器，并将最大迭代次数设为1,000。我们将“类别权重”设置为“平衡”，以添加上一节中所述的加权因子。所有其他超参数均设为默认值。

BERT。我们使用 Hugging Face Transformers实现了 BioBERT。我们使用网格搜索进行超参数调优：学习率 {5e-6、1e-5、2e-5、3e-5}，训练轮数{1, 2, 3}。我们使用了8 的批量大小。我们发现，学习率为 1e-5 且训练三个轮次时，模型表现最佳。

*C. MeSH*术语分类器性能

表 IV：在子集-500K 和子集-1M上的宏平均分类性能（在实验查询集中的 183 个 MeSH 术语上取平均）。

Data	模型	精确率	召回率	F1
子集-500K	LR	0.393	0.503	0.422
	BERT	0.489	0.460	0.445
子集-1M	LR	0.393	0.538	0.434
	BERT	0.487	0.444	0.431