一种用于电子商务分类体系中语义查询分类的思维链方法

Jetlir Duraj[1], IshitaKhan[1], Kilian Merkelbach[2] 和 MehranElyasi[1]

[1]*eBay Inc*，美国[2]
eBay Inc，德国

摘要

电子商务搜索的核心依托于对库存的结构化表示，通常以类别分类体系的形式构建。在采用层级分类体系的电子商务中，一项重要能力是选出一组与给定用户查询在语义上相符的相关叶类别。在此背景下，我们探讨现实电子商务分类体系中搜索查询分类这一基本问题。正确地对查询进行分类，不仅能让人们深入到正确的库存范围，还能为查询的多种意图理解能力打开大门。该问题的实用且准确的解决方案在电子商务领域有诸多应用，包括限制检索商品的范围以及提升搜索结果的相关性。

针对这项任务，我们探索了一种新颖的思维链（CoT）范式，将简单的树搜索与大语言模型语义评分相结合。通过在人工作出的查询-类别对评判、相关性测试以及基于大语言模型的参考方法上评估其分类性能，我们发现，思维链方法优于采用基于嵌入的查询类别预测的基准方法。我们展示了思维链方法如何检测层级分类体系中的问题。最后，我们还提出了同类的基于大语言模型的查询分类方法，但这些方法在数百万条查询的规模下具有更好的扩展性。

关键词

Q查询分类、电子商务搜索、分类体系、思维链推理、大语言模型 els

1. 引言

将用户查询映射到相关类别，对于电子商务搜索和导航至关重要，因为这能提升搜索相关性、用户导航体验和库存定位。传统上，业界和学术文献中已有研究利用点击率等用户行为数据的基于需求的方法。然而，这些方法面临展示偏差和信号稀疏等问题，尤其是在处理长尾查询和新库存时（例如，Joachims 等人 [1], Xv 等人 [2]）。

近期，基于语义的方法通过利用语言和上下文理解来推断相关性，为解决这些问题提供了有前景的方案；这类方法能够应对稀疏性和偏差，并泛化至新场景。这些方法将查询语义与分类体系相结合，以实现更精准的映射，但往往缺乏针对特定任务的适配，并侧重于静态表示（例如，Dehghani 等人 [3], Gao 等人 [4]）。

我们提出了一种新颖的语义投影系统，作为基于需求的方法的补充。我们将大语言模型（LLM）的思维链（CoT）推理范式（参见 Wei 等人 [5]）应用于层级分类体系中的分类这一特定问题。给定一个查询，我们的系统会沿分类体系从根节点逐步导航至叶类别，整合查询语义和分类体系细节，从而生成精准且可解释的映射。我们的方法与基于需求的方法相互正交，旨在丰富这些方法并减少其偏差。

我们的系统与思维链（CoT）推理理念相似：大语言模型通过将复杂任务拆解为较小的步骤和任务来解决问题，而我们构建的系统则通过结构化的多步流程逐步求解。系统会在分类体系的每个层级迭代预测并排列类别，从根节点逐步深入到叶类别。模型会根据当前类别节点、其子节点以及查询语义的信息，动态调整预测阈值。

SIGIR eCom'25：电子商务搜索与信息检索研讨会，*2025*年7月，意大利帕多瓦
jduraj@ebay.com(J. Duraj);ishikhan@ebay.com(I. Khan); kmerkelbach@ebay.com (K. Merkelbach);melyasi@ebay.com (M. Elyasi)
https://ceur-ws.org/Vol-4123/paper_24.pdf (J. Duraj)
© 2025 年本文作者版权所有。允许在知识共享署名 4.0 国际许可协议（CC BY 4.0）下使用



我们方法的一项关键特性是能够为查询指定上下文，例如购买、浏览或配件／互补意图等用户意图。指定上下文有助于进行更有针对性且符合情境的类别映射。模型会为每项预测提供置信度分数，并据此排列类别，从而提供可用于采取行动的洞见和可解释性。此外，它还可作为诊断工具，用于完善和改进分类体系，解决结构噪声与买家信号经常不匹配的问题。

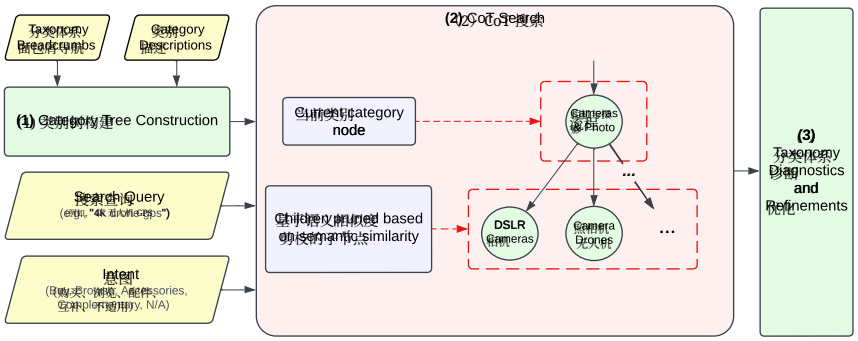

图 1： 我们系统架构的概览。首先，我们构建类别树，并为其补充类别描述。然后，我们在树中搜索最符合搜索查询的节点。最后，我们可以提取分类体系的诊断结果和优化建议。

1.1. 相关工作

电子商务中的推荐系统和个性化搜索在很大程度上依赖于对用户交互数据进行语义理解。文献中主要有两种方法：一种是利用行为数据的基于需求的方法，另一种是通过查询理解和内容理解来增强推荐效果的语义方法。

基于需求的类别预测方法这些方法利用隐式反馈（如点击数据）来个性化推荐。Ai 等人 [6] 提出了一种利用局部上下文提升排序效果的模型，而 Joachims 等人 [1] 则指出，尽管点击数据存在偏差，但仍具有实用价值。Xv 等人 [2] 利用图神经网络更好地处理长尾查询和冷启动产品，从而应对数据稀疏性这一持续存在的挑战。

基于语义的方法与查询理解这类方法侧重于理解查询意图，并将其与相关内容相匹配。Guo 等人 [7] 提出了深度相关性匹配模型，Mitra 和 Craswell [8] 则引入了用于对齐查询与文档的语义嵌入。这些方法有助于缓解偏差和数据稀疏问题。Dehghani 等人 [3] 展示了弱监督在稀疏数据集中的有效性，而 Gao 等人 [4] 则探索了分类体系中的语义泛化。

大语言模型 大语言模型善于泛化到未见过的场景，这是处理电子商务中长尾查询的一项关键能力。Brown 等人 [9] 展示了大语言模型在少样本学习方面的优势。思维链推理（Wei 等人 [10], Kojima 等人 [11]）能够增强层次推理和预测能力。

分类体系集成 分类体系为改进类别预测和推荐提供了结构化基础。Huang 等人 [12] 展示了如何在电子商务搜索中集成语义信号和分类体系。

2. 方法论

2.1. 高层概述

在给定输入数据的情况下，识别语义相关的叶类别是一项多标签分类问题。标签对应于树结构分类体系中的叶类别。因此，树路径搜索算法是一个自然的研究选择。此外，对于应用而言，评估查询与叶类别之间语义关系的强度至关重要。因此，我们将直接的树搜索方法与用于衡量语义相关性的大语言模型评分相结合。在大语言模型评分中，我们要求大语言模型提供一个分数，用以衡量查询与某个电子商务类别及其描述之间语义关系的强度。正如表1.所示，在我们的任务中，大语言模型评分对语义相关性的评估与人工判断十分接近。

表1大语言模型评分的分类性能（以人工判断作为真实值）。

LLM	F1	精确率	召回率
Mixtral-8x7B	0.727	0.825	0.649
Llama3-70B	**0.805**	0.797	**0.813**
OpenAI-GPT-4oMini	0.743	**0.862**	0.654

此表展示了 4,897 个查询-叶类别对在语义相关性方面的分类指标，真实值由人工评审员确定。我们评估的大语言模型在 F1值、精确率和召回率方面表现良好。在推理速度方面，Mixtral-8x7B 是一种可在本地托管的混合专家模型，尽管其 F1值排名第三，但推理速度显著高于另外两个受评模型。[1] 本文的方法论结果采用 Mixtral-8x7B。

在语义相关性任务中，倾向于采用大语言模型评分而非更具生成性的 大语言模型 方法，另一个原因是指令遵循问题。由于当前大语言模型架构具有生成性，这类问题无法彻底消除。若提示大语言模型直接从类别节点的可用子节点中进行选择，它们有时会修改类别名称，而不是准确复现输入中的名称。这个问题在 OpenAI 或 Gemini 等闭源大语言模型中较少出现（我们的实验中失败率为 4%），但在 Hugging Face 的开源模型中则更为明显；这些模型可在本地托管，并支持大规模推理。

基于这些原因，我们专注于为方法中的大语言模型组件采用评分方法：我们要求大语言模型给出语义相关性的置信度评分，范围为 1（最低）到 10（最高）。是否继续搜索每个类别节点，取决于其子节点的语义评分，以及其他上下文信息或算法运行时信息。

2.2. 广度优先搜索设计（CoTBFS）

将分类任务视为分类之后的回归（赋分）任务后，我们的方法会先完整解决相关叶类别的分类问题，再处理回归问题。具体而言，对于给定的查询，在分类体系的第 1 层对所有第 1 层类别的语义相关性进行评分后，我们仅保留语义相关性相对最高的子节点，并剪除其余子节点。我们使用两个依赖于查询的选择阈值：选择阈值和最小阈值，二者范围均为 1 到 10。类别子节点的相关性得分（1 到 10）会映射到标准正态分布。将选择阈值除以 10 后，应用于标准化得分，以剪除相关性较低的子节点。例如，当选择阈值为 9（满分 10）时，语义得分低于均值加上 0.9 倍标准差的子节点将被剪除。为了应对得分分布在低分端可能存在的高度偏斜，子节点的

[1]实验在 eBay 的基础设施中运行耗时不到两小时。OpenAI-GPT-4oMini 是闭源模型，而 Llama3-70B 则部署在内部。

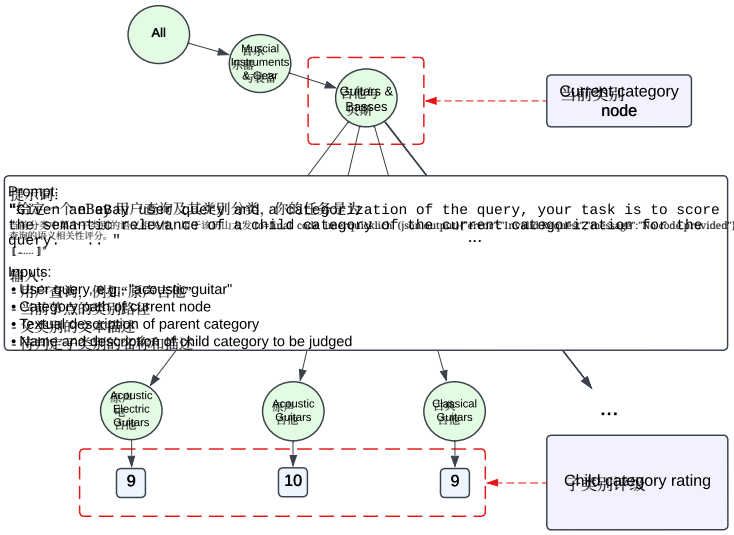

图 2：针对查询原声吉他的 CoT BFS 输出。使用 Mixtral-8x7B 作为大语言模型。提示词包含从树的根节点到当前节点的路径信息、直接父节点，以及待评级的子节点。

原始语义分数必须超过最小阈值（第二个阈值），才能保留并继续探索。

接下来，算法从通过初始剪枝的第 1 层子节点开始，逐一检查每棵子树。我们按照上述相同的相对阈值处理流程，对这些子树中的子节点重复剪枝，以识别未被剪枝的第 2 层节点。该迭代过程持续进行，直到到达没有子节点的节点，即叶类别，并将其添加到候选叶类别集合中。

最后，由于搜索采用的是相对而非绝对语义阈值法，我们仅使用叶类别信息（分类路径和描述）为最终叶类别集合评分。语义相关性较高且高于最小阈值的叶类别，即为最终预测。

对于实证应用，我们按如下方式选择选择阈值和最小阈值：鉴于语义评分范围为 1（最低）至 10（最高），我们不会考虑低于 6 的阈值。在 7、8、9 这几个阈值中，对于选择阈值和最小阈值，我们只考察选择阈值高于最小阈值的组合。在这些组合中，我们选取针对由约 1000 个代表性查询构成的人工判断数据集，F1 分数表现最佳的一组。[2]

我们将这种方法称为思维链广度优先搜索（CoT BFS）。

图 2 展示了查询原声吉他的 CoT BFS 分类结果，选择阈值为 9，最小阈值为 8。在第一步中，CoT BFS 将范围缩小到一个唯一的一级类别： *AllCats* > 乐器与设备，其获得的中间语义分数为 10。在 35 个一级类别中，其他所有一级类别的分数都较低，众数分数为 1，其他类别的最高分数为 4（类别*AllCats* > 音乐）。存留的一级类别 *AllCats* > 乐器与设备 有 16 个子节点。下一步分类会从这 16 个子节点中剪枝掉 15 个，因为它们的语义分数都低于 8。子节点 吉他与贝斯 的语义分数为 10，并且

[2]请注意，最高分 10，也就是 1..10 的取值范围跨度，同样是一个超参数。随着时间推移，需要定期验证这些超参数，以应对查询的分布漂移以及分类体系的变化。受篇幅限制，我们无法详细分析超参数的影响。这里定性报告如下：降低选择阈值和最小阈值通常会提高召回率，但会降低精确率。我们还发现，使用 1..5 这一替代范围而非 1..10，会同时降低精确率和召回率；而使用 1..20 则会略微提高召回率、降低精确率，并降低 F1值。

得以保留。此外，节点 *AllCats* >乐器与设备 > 吉他与贝斯 有 13 个子节点。在下一步对子节点进行评分时，除三个子节点外，其余全部被剪枝。存留的子节点为以下节点：古典吉他 （最终分数为 9）、电箱吉他 （最终分数为 9）以及原声吉他 （最终分数为 10）。此时，搜索停止，因为到达的节点已经是类别树的叶类别。

2.3. 思维链和大语言模型评分的可扩展方法

对于给定查询，CoT BFS 中大语言模型调用的总次数与分类体系中访问的类别节点数量处于同一数量级。针对大型查询数据集的实验表明，CoT BFS 能够访问 eBay 分类体系中 1.7% 至 24.8% 的类别节点。这体现了我们方法的高效性，尤其是考虑到 eBay 分类体系中类别数量极其庞大。尽管如此，要将该方法扩展至数百万个查询并实现低延迟，仍需进行调整。我们提出了两种方法，其中第二种比第一种更具可扩展性。

2.3.1. 思维链-k近邻混合 BFS

基于类别名称或描述嵌入的 k近邻检索，可在树搜索流程的每一步中用作过滤器。无需对每个子节点进行穷举评分，只需由大语言模型对通过嵌入距离过滤器（比较用户查询与类别文本表示）的子集进行评分。这样可以减少分类体系中每个节点的大语言模型调用次数，并将搜索范围限制在最有希望的方向上。

2.3.2. 叶类别上的 k近邻搜索 + 大语言模型评分

用叶类别嵌入上的 k近邻搜索取代树搜索，将其作为预筛选步骤，随后对通过 k近邻识别出的候选项进行大语言模型评分。在流程开始时使用较多的近邻（例如 20 个）运行 k近邻，可以提高召回率。我们在第 3.2 节中使用此方法的一个变体，为评估 CoT BFS 构建合成基准真值。

3. 实验

3.1. 基线模型：k近邻分类

我们用于评估 CoT BFS 的基准，是使用 sentence-BERT 的嵌入（未经微调）对叶类别进行 k 近邻搜索。余弦相似度用作 k近邻的度量指标。[3]我们详细报告分类表现，并从微观聚合、宏观聚合和样本聚合三个方面，对比我们方法的 F1值、精确率和召回率等分类指标。微观聚合考察所有查询和叶类别上的整体表现。样本聚合考察每个查询的表现，然后进行汇总。宏观聚合考察每个叶类别的表现，然后进行汇总。

3.2. 与基线进行比较

3.2.1. 人工判断

人工判断通过运用人类直觉和专业知识，提供定性和定量评估。评估人员会审查预测类别的语义相关性，不过这一流程具有主观性，而且在大型数据集中成本高昂。尽管成本不菲，人工判断能够捕捉其他方法常常遗漏的细微差别，因此不可或缺。我们的人工判断数据集包含 1018 个查询

[3]我们可以使用基于 eBay 专用数据训练的语言模型。这类模型在 eBay 相关任务中的表现通常优于通用语言模型，但我们并未展示使用 eBay 专用语言模型得到的结果。这是因为我们关注的是 CoT BFS 方法在通用、未经微调的大语言模型上的表现。此外，使用可公开访问的语言模型有助于确保结果的可复现性。

和 4897 个查询-类别对，并根据语义相关性进行判断（是/否）。我们注意到，用于判断的叶类别是根据用户行为信号选定的，因此会受到呈现偏差的影响，而这种偏差源于 eBay 当前在生产环境中使用的模型。标注员是三位由 eBay 资助、专攻 eBay 分类体系的独立领域专家。

表 2 展示了在真实值由人工判断给出的情况下，CoT BFS 相对于基准的性能。CoT BFS 的表现优于基线模型，尤其是在 F1值和精确率方面。

表 2基于人工判断数据的分类结果。CoT BFS 选择阈值为 9，最小阈值为 8。相对于基准的相对提升。

指标	F1	精确率	召回率
微观聚合	+89.8%	+86.1%	+3.4%
样本聚合	+109.7%	+190.6%	-13.5%
宏观聚合	+44.7%	+60%	+31.4%

3.2.2. AI伪参考方法

我们采用 AI 伪参考方法创建一个数据集，在不受需求信号数据集中呈现偏差影响的情况下近似真实值。首先，我们从 3000 个用户查询开始，由高质量大语言模型模拟人工判断，评估查询-类别对的语义相关性。为避免在大语言模型评分时遗漏潜在相关的叶类别，我们使用包含大量邻居的基于嵌入的 k 近邻搜索。具体而言，我们为每个查询选出 100 个最相关的类别，并从中排除余弦相似度低于 0.01 的类别。

随后，我们使用一个比用于 CoTBFS 的本地部署 Mixtral-8x7B 更强的大语言模型（OpenAI-GPT-4o-Mini），对每一对进行 1 到 10 分的评分。这种采用大型 k的混合方法，另请参见小节 2.3.2，提供了真实值的代理指标。表3展示了结果。CoT BFS 再次优于基线，尤其是在 F1值和精确率方面。

表3大语言模型判定的数据（OpenAI-GPT-4oMini）上的分类结果。CoTBFS选择阈值9，最小阈值 8。相对于基准的相对提升。

指标	F1	精确率	召回率
微观聚合	+96.7%	+137.5%	-5.4%
样本聚合	+132.1%	+89.8%	+3.7%
宏观聚合	+112.1%	+47.8%	+111.4%

3.2.3. 检索测试

检索测试通过比较我们的模型与基线在检索商品层面的召回率和相关性，评估预测的叶类别。根据叶类别从库存中检索商品，估计召回规模显示找到的相关商品所占比例。相关性采用 eBay 内部的 PEGFB 模型进行衡量；该模型基于人工判断数据训练，并将结果划分为五个等级：完美、优秀、良好、一般和差。检索测试评估凸显了该模型在提升用户满意度和搜索效率方面的实际效用。我们的模型在召回率和相关性两方面均显著优于 k近邻基准，曼-惠特尼U检验结果表明，CoT BFS 的表现显著更优。

表43000 条查询的分类结果。CoT BFS 选择阈值为 9，最小阈值为 8。相对于基准的提升。

	估计召回规模	相关性得分
Mean	+72%	+34%
Mann-Whitney U 统计量	5401539.0	5363320.0
Mann-Whitney U p 值	8.95e-88	1.30e-82

4. 应用

4.1. 上下文学习

通过从大量真实世界数据集中学习，大语言模型能够识别揭示用户意图和偏好的模式，从而实现个性化搜索并提升语义相关性。上下文学习可以根据给定的查询上下文优化模型输出，确保预测结果具有相关性。这一能力对于 eBay 这样的平台至关重要，因为准确识别买家意图有助于提升搜索体验。我们考虑了上下文学习的两种应用，一种针对用户意图，另一种针对品牌来源。

更具体地说，eBay 分类体系在电子产品、汽车和时尚等不同商品细分领域中都包含配件相关类别。CoT BFS 方法可以通过修改大语言模型提示词，将购买、寻找配件、寻找互补商品等意图纳入其中，从而轻松整合买家意图。图3 展示了查询佳能相机。

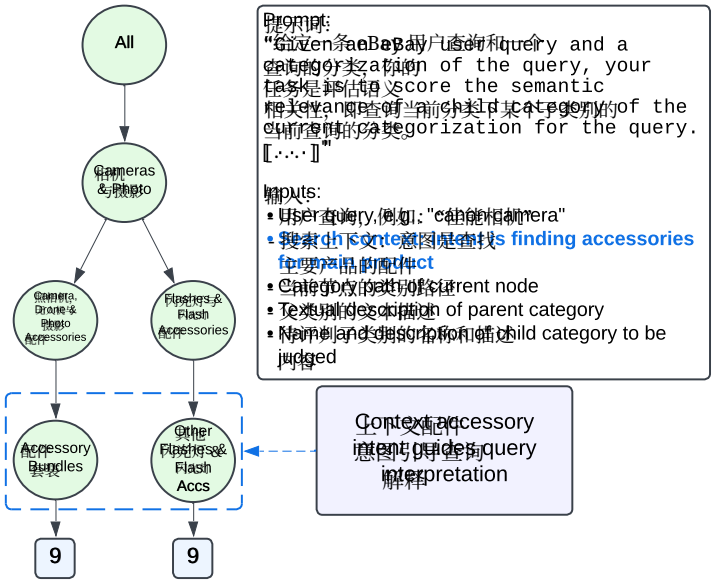

图 3： 针对具有配件意图的查询 佳能相机 的思维链 BFS 输出。将查询意图作为上下文提供，可引导类别搜索至与查询中实体相关的配件类别。

若不提供意图，该查询识别出的首要类别为 *AllCats* > 相机与摄影 >数码相机，得分为 10。将 配件 意图作为搜索上下文注入提示词后，识别出的首要类别为 *AllCats* > 相机与摄影 > 相机、无人机与摄影配件 > 配件套装，以及 *AllCats* > 相机与摄影 >闪光灯及闪光灯配件 > 其他闪光灯及闪光灯配件，各自得分均为 9。更广泛地说，表5 展示了指定 配件 意图时，15 个选定查询的配件类与非配件类预测类别的平均语义得分如何变化。

在思维链 BFS 中纳入买家意图，有助于更精准地定位相关类别。

类似地 接下来，我们考察注入有关品牌来源的上下文所产生的影响。举例来说，对于 e

表 5上下文学习：配件意图可提升配件类别的预测效果。

配件类别？（是/否）	无意图时的平均分	有配件意图时的平均分
No	**9.143**	1.000
Yes	2.643	**7.214**

查询 跑车，我们考虑两种不同的上下文：品牌来源是德国和品牌来源是意大利。对于此查询，第一种情况下，CoT BFS 仅预测与德国品牌奥迪、宝马、梅赛德斯-奔驰、保时捷相关的跑车叶类别；第二种情况下，则仅预测与意大利品牌阿尔法·罗密欧、德托马索、法拉利、菲亚特、玛莎拉蒂、兰博基尼相关的叶类别。

4.2. 检测问题并改进分类体系

通过分析查询模式并识别类别表示方面的缺口，CoT BFS 可以为改进电子商务分类体系结构提供可操作的洞见。

在这方面，我们使用一个具有代表性的 25,000 条查询样本开展了一项实验，并通过高阈值的 CoT BFS 方法处理这些查询：选择阈值为 10，最小阈值为 9。共有 3,110 条查询的模型预测结果为空，这表明当前 eBay 分类体系在前几个层级缺少与这些查询在语义上高度相关的类别节点。我们通过对嵌入进行 k近邻搜索，对这些查询进行聚类，从而发现了一些规律。例如，其中两个“失败”查询簇分别对应电子商务类别设计师太阳镜和光学仪器及配件。对于识别出的第一个查询簇（设计师太阳镜），当前 eBay 分类体系中最接近的叶类别是*AllCats* > 服装、鞋履及配饰 > 女装 >女士配饰 > 太阳镜及太阳镜配件 > 太阳镜 和*AllCats* > 服装、鞋履及配饰 > 男装 > 男士配饰 > 太阳镜及太阳镜配件 > 太阳镜，二者在分类体系中的深度均为 5。第二个查询簇也存在类似问题。从大量用户查询中得出的这类洞见，有助于产品管理团队改进分类体系。例如，新增一个名为光学产品及眼镜 的第 2 层类别，并设置设计师太阳镜 和光学仪器及配件 等子类别，或许有助于提升搜索体验。

归根结底，要维护一个能为用户提供高价值的电子商务分类体系，涉及复杂的业务和产品管理决策。我们的方法论提供了探索分类体系问题的工具，旨在提升这类复杂业务决策的决策水平。

5. 结论与未来工作

本研究提出了一种用于层级分类体系内查询分类的新方法论。该方法结合了大语言模型的世界知识和简单的树搜索算法，实现高质量分类，并深入揭示分类体系的特点。

第 2.3 节介绍的可扩展方法计划进行 A/B 测试。2.3这些测试包括直接测试，即缓存模型预测结果以供生产环境使用；也包括间接测试，即对利用经思维链广度优先搜索标注的数据训练的低延迟分类模型进行 A/B 测试。

此外，我们还开发了一个思维链算法版本，在每个分类体系节点上采用绝对阈值处理，而非论文中讨论的相对阈值处理；由于篇幅所限，论文中未予介绍。该方法称为思维链深度优先搜索（CoT DFS），它以深度优先的方式搜索叶类别；当遇到语义相关性绝对值较低的中间节点时，便停止沿该路径继续搜索，这与本文所述的相对阈值处理不同。由于要求更为严格，CoT DFS 方法会产生更多预测结果为空的查询。在进行分类体系诊断时，CoT DFS 能比 CoT BFS 更有效地利用用户查询活动和大语言模型语义知识。

生成式人工智能声明

作者未使用任何生成式人工智能工具。

参考文献

[1] T. Joachims、L. Granka、B. Pan、H. Hembrooke、G. Gay，将点击数据准确地解释为隐式反馈，载于：第28届ACM SIGIR国际信息检索研究与发展年会论文集，SIGIR '05，计算机协会，美国纽约州纽约，2005年，第154–161页。URL：https://doi.org/10.1145/1076034.1076063。doi:10.1145/1076034.1076063.[2] G. Xv、C. Lin、W. Guan、J. Gou、X. Li、H. Deng、J. Xu、B. Zheng，基于内容协同图神经网络的电子商务搜索，载于：第29届ACM SIGKDD知识发现与数据挖掘会议论文集，KDD '23，计算机协会，美国纽约州纽约，2023年，第2885–2897页。URL：https://doi.org/10.1145/3580305.3599320。doi:10.1145/3580305.3599320.[3] M. Dehghani、H. Zamani、A. Severyn、J. Kamps、W. B. Croft，基于弱监督的神经排序模型，载于：第40届ACM SIGIR国际信息检索研究与发展会议论文集，SIGIR '17，计算机协会，美国纽约州纽约，2017年，第65–74页。URL：https://doi.org/10.1145/3077136.3080832。doi:10.1145/3077136.3080832.[4] J. Gao、P. Pantel、M. Gamon、X. He、L. Deng，利用深度神经网络建模趣味性，载于：自然语言处理实证方法会议，2014年。URL：https://api.semanticscholar.org/CorpusID:2141094。[5] J. Wei、X. Wang、D. Schuurmans、M. Bosma、B. Ichter、F. Xia、E. H. Chi、Q. V. Le、D. Zhou，思维链提示能激发大语言模型的推理能力，载于：神经信息处理系统进展，第35卷，2022年，第24824–24837页。[6] Q. Ai、K. Bi、J. Guo、W. B. Croft，学习用于排序优化的深度列表式上下文模型，载于：第41届ACM SIGIR国际信息检索研究与发展会议，SIGIR '18，计算机协会，美国纽约州纽约，2018年，第135–144页。URL：https://doi.org/10.1145/3209978.3209985。doi:10.1145/3209978.3209985.[7] J. Guo、Y. Fan、Q. Ai、W. B. Croft，用于即席检索的深度相关性匹配模型，载于：第25届ACM信息与知识管理国际会议论文集，CIKM '16，计算机协会，美国纽约州纽约，2016年，第55–64页。URL：https://doi.org/10.1145/2983323.2983769。doi:10.1145/2983323.2983769.[8] B. Mitra、N. Craswell，神经信息检索导论，《信息检索基础与趋势》13（2018），第1–126页。URL：http://dx.doi.org/10.1561/1500000061。doi:10.1561/1500000061。[9]T. Brown、B. Mann、N. Ryder、M. Subbiah、J. D. Kaplan、P. Dhariwal、A. Neelakantan、P. Shyam、G. Sastry、A. Askell、S. Agarwal、A. Herbert-Voss、G. Krueger、T. Henighan、R. Child、A. Ramesh、D. Ziegler、J. Wu、C. Winter、C. Hesse、M. Chen、E. Sigler、M. Litwin、S. Gray、B. Chess、J. Clark、C. Berner、S. McCandlish、A. Radford、I. Sutskever、D. Amodei，语言模型是少样本学习器，载于：H. Larochelle、M. Ranzato、R. Hadsell、M. Balcan、H. Lin（编），神经信息处理系统进展，第33卷，Curran Associates, Inc.，2020年，第1877–1901页。URL：https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf。[10] J. Wei、X. Wang、D. Schuurmans、M. Bosma、B. Ichter、F. Xia、E. Chi、Q. Le、D. Zhou，思维链提示能激发大语言模型的推理能力，2023年。URL：https://arxiv.org/abs/2201.11903。arXiv:2201.11903.[11] T. Kojima、S. S. Gu、M. Reid、Y. Matsuo、Y. Iwasawa，大语言模型是零样本推理器，ArXiv abs/2205.11916（2022）。URL：https://api.semanticscholar.org/CorpusID:249017743。

[12] P.-S. Huang、X. He、J. Gao、L. Deng、A. Acero、L. Heck，利用点击数据学习用于网络搜索的深度结构化语义模型，载于：第22届ACM信息与知识管理国际会议论文集，CIKM ’13，计算机协会，美国纽约州纽约，2013年，第2333–2338页。URL：https://doi.org/10.1145/2505515.2505665。doi：10.1145/2505515.2505665。