

一种用于电子商务查询分类的半监督可扩展统一框架

Chunyuan Yuan¹, Chong Zhang¹, Zheng Fang¹, Ming Pang^{1,*},
Xue Jiang¹, Changping Peng¹, Zhangang Lin¹, Ching Law¹

¹JD.COM,

{yuanchunyuan1,zhangchong78,fangzheng21,pangming8,jiangxue,pengchangping,linzhangang,lawching}@jd.com

摘要

查询分类,包括意图和类别预测等多个子任务,对于电子商务应用至关重要。电子商务中的查询通常很短且缺乏上下文,标签之间的信息无法使用,导致建模时先验信息不足。大多数现有的工业查询分类方法依赖用户的事后点击行为来构建训练样本,从而形成马太效应的恶性循环。此外,查询分类的子任务缺乏统一框架,导致算法优化效率低下。

本文提出了一种新颖的 **S** 半监督 **S** 可扩展 **U** 统一 **F** 框架 (SSUF), 包含多个增强模块以统一查询分类任务。知识增强模块利用世界知识来增强查询表示,并解决查询信息不足的问题。标签增强模块利用标签语义和半监督信号来减少对后验标签的依赖。结构增强模块基于复杂的标签关系来增强标签表示。每个模块都具有高度的可插拔性,可以根据每个子任务的需求添加或删除输入特征。我们进行了广泛的离线和在线 A/B 实验,结果表明 SSUF 显著优于最先进的模型。

1 介绍

像亚马逊、淘宝和京东这样的电子商务平台为用户提供数十亿种多样的产品,并已成为我们日常生活中的重要组成部分。由于用户需求和产品类别的多样性,捕捉用户的购买意图对于用户体验和平台效率都至关重要。查询分

类,包括意图、类别和品牌预测,在理解用户的购物需求方面发挥关键作用,并支持搜索系统的后续模块。

电子商务查询固有的特性,通常是短小且模糊的,给查询分类带来了重大挑战。为了解决由于短查询导致的信息不足问题,一些基于深度学习的模型,如 XML-CNN (Liu et al., 2017), KRF (Ma et al., 2020), HiAGM (Zhou et al., 2020) 和 LSAN (Xiao et al., 2019) 已被提出用于学习文档的上下文信息以增强查询的表示学习。一些最近的查询分类模型,如 HCL4QC (Zhu et al., 2023), SMGCN (Yuan et al., 2024) 和 HQC (He et al., 2024) 也探索了利用分层类别树结构或实例层次来帮助模型学习超出查询信息的外部信息。

工业查询分类方法通常依赖用户的点击行为来生成训练样本。虽然使用真实用户交互可以提高模型的准确性,但也引入了一个被称为“马太效应”的依赖循环。这个循环导致了有偏差的训练数据,其中流行的查询受到了过多的关注,扭曲了模型的理解并限制了其对长尾查询的一般化能力。此外,现有的模型通常分别处理子任务,忽略了可能增强模型优化和开发效率的潜在协同作用。缺乏统一框架进一步阻碍了在不同子任务之间共享见解和改进,从而限制了整体性能。

为了解决这些挑战,我们提出了一种半监督可扩展统一框架 (SSUF) 用于电子商务查询分类。SSUF 通过引入一组可扩展模块来克服

* 通讯作者。

上述问题：(1) 标签增强模块，(2) 知识增强模块，和 (3) 结构增强模块，以利用先验知识增强查询和标签表示，减少对后验标签的依赖，并提高模型从有限数据中泛化的能力。SSUF 内部的每个模块都设计为高度可插拔的，允许灵活适应不同子任务的具体需求。这种模块化确保了该框架可以定制以提升查询分类的各方面。

本文的贡献如下：

- 我们提出了一种新的统一框架来提升电子商务查询分类模型的优化效率。
- 我们设计了三个可扩展模块，以增强查询和标签表示，并打破“马太恶性循环”以提升查询分类的性能。
- 我们进行了广泛的离线和在线 A/B 实验，SSUF 显著优于现有的强基线。它已部署在电子商务平台上并带来了巨大的商业价值。

2 相关工作

2.1 多标签分类

多标签分类是机器学习中的一个重要领域，其中每个实例可以关联到多个标签。机器学习方法通过将多标签问题转化为几个单标签任务来解决这个问题 (Tsoumakas et al., 2007, 2009; Read et al., 2011)。最近，深度学习模型，如 XML-CNN (Liu et al., 2017)，LSAN (Xiao et al., 2019) 和 LEAM (Wang et al., 2018) 利用上下文信息或特定于标签的注意力来增强文档与标签之间的交互以进行分类。

2.2 查询分类

早期模型主要依赖深度学习模型，如 CNN (Hashemi et al., 2016)，LSTMs (Sreelakshmi et al., 2018) 和基于注意力的模型 (Zhang et al., 2021; Yuan et al., 2023) 来提取细粒度特征进行分类。近期的工作例如 PHC (Zhang et al., 2019) 探索多任务框架以联合优化查询分类和

文本相似性，而 DPHA (Zhao et al., 2019) 则利用基于标签图的神经网络来建模标签相关性。HCL4QC (Zhu et al., 2023)，SMGCN (Yuan et al., 2024) 和 HQC (He et al., 2024) 使用分层结构和实例层次来学习超出查询文本的信息。

3 模型

本节中，我们首先正式定义查询分类任务。然后，我们描述 SSUF 的不同模块，并分析模型在训练和推理过程中产生的影响。

3.1 标签增强模块

我们不直接使用标签索引作为标签嵌入，而是采用 BERT (Kenton and Toutanova, 2019) 作为标签的编码器来学习标签的语义表示。

文本编码器的输入是一个标签的字符序列，它由两部分组成：(1) 标签名称 $n = [n_1, n_2, \dots, n_L]$ ，和 (2) 增强的标签侧面信息 $m = [m_1, m_2, \dots, m_{L_m}]$ ，这些是从 (1) 标签描述中检索出来的，例如产品词、经常搜索的查询词等。(2) 由大语言模型生成的世界知识。

标签的字符序列被输入到 BERT 中以编码标签表示：

$$C_j = \text{BERT}_{CLS}([n_1, \dots, n_L, m_1, \dots, m_{L_m}]), \quad (1)$$

其中， $C_j \in \mathbb{R}^{1 \times d}$ 是 BERT 最后一层的“CLS”表示。同样地，我们可以得到查询 $Q_i \in \mathbb{R}^{1 \times d}$ 的表示。

3.2 知识增强模块

工业界的方法依赖于用户的后验点击行为来生成训练样本。然而，这导致了“马太恶性循环”，并造成训练数据的偏差，热门查询获得了更多关注，扭曲了模型的理解能力，并限制了其对较少出现查询的泛化能力。

我们提出了一种半监督模块来克服后验标签的局限性。然而，我们发现对于语义模糊的查询，直接计算查询和标签的半监督标签通常是不准确的。例如，查询“Black 16pro”指的是一个苹果手机型号，但由于语义信息不足，与

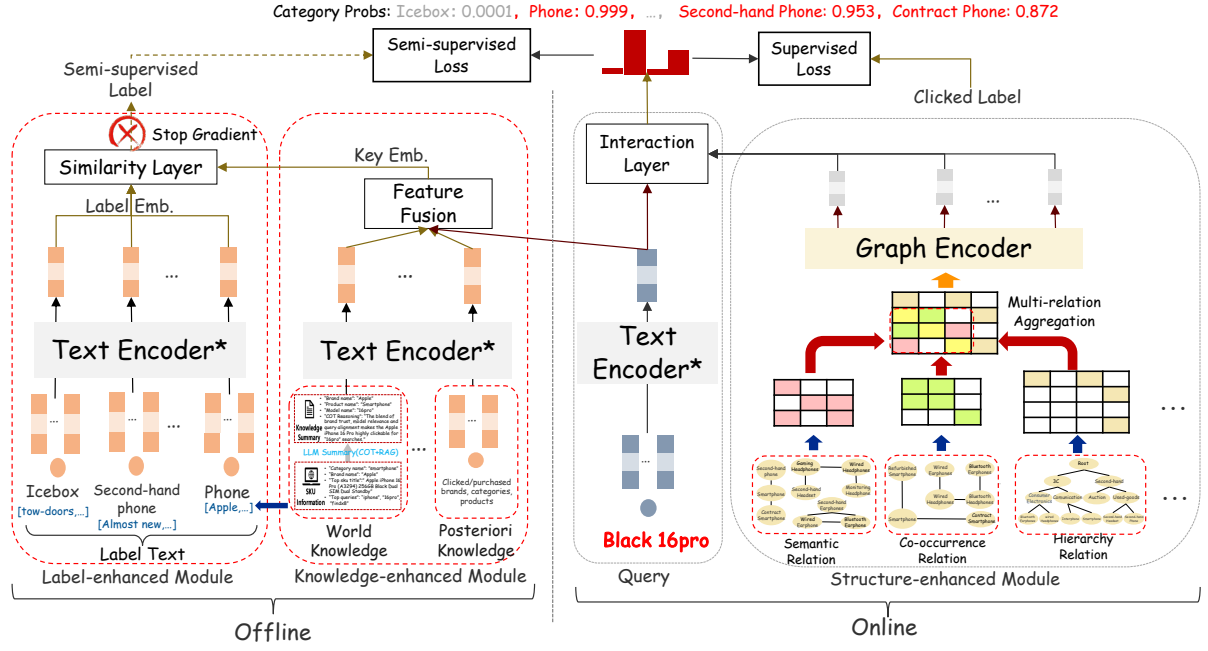


图 1: 半监督可扩展统一框架用于电子商务查询分类。离线部分参与模型训练，但不直接在线部署。带红色虚线的部分是一个可插拔模块。“文本编码器*”表示一个共享的文本编码器。

相关标签如“手机”和“二手手机”的相似度得分较低。这导致了半监督信号无法有效地召回相关的标签。为了解决这个问题，我们引入了一个知识增强模块来改善用于半监督标记的查询表示。

我们可以使用 (1) 后验知识，如用户经常点击或购买的产品标签，(2) 从 LLM 中提取的常识作为输入。为了获得查询的常识，我们将查询和相关产品提供给一个开源的 LLM 来总结简要描述，这可能包含相关的查询、类别、产品等。借助这些信息，模型可以全面编码查询的语义表示。

获得后验和世界知识后，我们将它们输入一个共享的文本编码器：

$$\mathbf{k}_i = \text{BERT}_{CLS}([k_1, \dots, k_n]), \quad (2)$$

以获取知识嵌入 $\mathbf{K} \in \mathbb{R}^{|K| \times d}$ 。

为了将这些知识嵌入与查询表示 $\mathbf{Q}_i$ 融合，我们使用一个注意力模块，其可以表述如下：

$$\begin{aligned} \alpha &= \text{softmax}(\mathbf{Q}_i \mathbf{K}^T), \\ \mathbf{q}'_i &= \mathbf{Q}_i + \sum_{j=1}^{|K|} \alpha_j \mathbf{K}_j, \end{aligned} \quad (3)$$

其中 α 是注意力得分，而 $\mathbf{q}'_i \in \mathbb{R}^{1 \times d}$ 是最终融合后的查询表示。

我们计算融合后的查询和标签表示之间的相似度分数，将其视为半监督标签。具体来说，

$$\mathbf{s}_i = \text{stop_grad} \left(\frac{\mathbf{q}'_i \mathbf{C}^T}{\|\mathbf{q}'_i\| \|\mathbf{C}\|} \right), \quad (4)$$

$$\mathbf{y}_{ij}^{\text{semi}} = \mathbf{s}_{ij} \cdot \mathbb{1}_{\mathbf{s}_{ij} \geq \tau},$$

其中 $\mathbf{s}_i \in \mathbb{R}^{1 \times |C|}$ 是查询 q_i 和所有类别之间的相关性分数。 τ 是用于过滤低分类别的阈值。 $\mathbf{y}_{ij}^{\text{semi}}$ 是半监督标签。

查询和标签都使用相同的文本编码器，但它们的词分布不同。将半监督信号的梯度反馈到半监督标签模块中可能会产生循环依赖，从而可能导致模型崩溃。为了防止这种情况，我们禁用了此分支的梯度反馈。

3.3 结构增强模块

3.3.1 图构建

首先，我们通过计算训练样本中类别共现的次数来获得类别间的共现关系。然后，我们计算两个类别的条件概率并得到邻接矩阵 $\mathbf{A}^{\text{coo}}$ ：

$$\mathbf{a}_{ij} = \frac{N(c_i, c_j)}{N(c_i)}, \mathbf{A}_{ij}^{\text{coo}} = \mathbf{a}_{ij} \cdot \mathbb{1}_{\mathbf{a}_{ij} \geq \alpha} \quad (5)$$

其中 $N(c_i, c_j)$ 是标签 c_i 和 c_j 的共现频率，而 $N(c_i)$ 表示标签 c_i 的频率。 α 是过滤低相关性得分边界的阈值。 $\mathbf{A}^{coo} \in \mathbb{R}^{|C| \times |C|}$ 是共现的邻接矩阵。

然后，我们可以通过计算每对类别的余弦相似度来获得类别间的语义相似关系：

$$\mathbf{a}_{ij} = \frac{\mathbf{C}_i \mathbf{C}_j^T}{\|\mathbf{C}_i\| \|\mathbf{C}_j\|}, \mathbf{A}_{ij}^{sim} = \mathbf{a}_{ij} \cdot \mathbf{1}_{\mathbf{a}_{ij} \geq \beta}, \quad (6)$$

其中 β 是用于过滤相关性得分较低的边的阈值。 $\mathbf{A}^{sim} \in \mathbb{R}^{|C| \times |C|}$ 是相似性邻接矩阵。

对于某些查询分类子任务，如意图或类别预测，各个级别的标签之间存在层次结构。这种结构有助于加强相关标签之间的关系并削弱不相关标签之间的紧密度。为了利用此结构，我们将它编码到层次邻接矩阵 $\mathbf{A}^{hier} \in \mathbb{R}^{|C'| \times |C'|}$ 中，并定义边为：

$$\mathbf{A}_{ki}^{hier} = \max \left(\frac{1}{|Child(k)|}, \frac{m_i}{\sum_{j \in Child(k)} m_j} \right), \quad (7)$$

其中 $Child(k)$ 是节点 k 的子节点集，且 $i, j \in Child(k)$ 。 m_j 是数据集中用户点击节点 j 的频率。 $|C'|$ 表示所有标签的数量，包括一级、二级和叶标签。 $|C|$ 表示叶标签的数量。

除了上述三个标签关系图之外，每个子任务还可以根据其现有的输入数据和业务特征增加或减少标签图的数量。

3.3.2 图融合与学习

获得标签相关性矩阵后，我们融合这些相关性矩阵，并使用归一化方法 (Kipf and Welling, 2017) 对融合后的矩阵进行归一化：

$$\mathbf{A} = \frac{1}{2}(\mathbf{A}^{coo} + \mathbf{A}^{sim}) \rightarrow \mathbf{A}^{hier}, \quad (8)$$

$$\hat{\mathbf{A}} = \mathbf{D}^{-\frac{1}{2}} (\mathbf{A} + \mathbf{I}) \mathbf{D}^{-\frac{1}{2}},$$

其中 $\rightarrow$ 表示赋值符号。赋值过程如图 1 所示。 $\mathbf{A} \in \mathbb{R}^{|C'| \times |C'|}$ 是最终的邻接矩阵。 $\mathbf{I}$ 是一个单位矩阵。 $\mathbf{D}$ 是一个对角度矩阵，具有 $D_{ii} = \sum_j \mathbf{A}_{ij}$ 。最后，我们使用 GCN (Kipf and Welling, 2017) 从最终的邻接矩阵 $\mathbf{A}$ 中学习节点的表示 $\mathbf{H} \in \mathbb{R}^{|C'| \times d}$ 。

尽管尾标签的训练样本有限，但可以通过复杂的标签关系轻松将这些标签与其相关的热门标签关联起来。这样的关系使得能够从具有热门标签的样本向具有尾标签的样本转移梯度，从而更有效地对尾标签进行表征训练，并缓解后验标签的限制。

3.4 训练和推理

在我们的应用场景中，我们只需要将用户的输入查询 $\mathbf{q}_i \in \mathbb{R}^{1 \times d}$ 分类到叶标签空间而不是所有标签。因此，我们从 $\mathbf{H}$ 中提取得到叶标签嵌入 $\mathbf{H}_l \in \mathbb{R}^{|C| \times d}$ 。最后，我们使用一个交互层将查询投影到标签空间：

$$\hat{\mathbf{y}}_i = \text{sigmoid}(\mathbf{q}_i \mathbf{H}_l^T + \mathbf{b}), \quad (9)$$

其中 $\mathbf{b} \in \mathbb{R}^{1 \times |C|}$ 是偏置项， $\hat{\mathbf{y}}_i \in \mathbb{R}^{1 \times |C|}$ 是查询 q_i 的预测标签。

为了用后验标签和先验标签优化模型，我们将它们融合如下：

$$\mathbf{y}_i = \min(\mathbf{y}_i^{click} + \mathbf{y}_i^{semi}, 1.0), \quad (10)$$

其中 $\mathbf{y}_i^{click}$ 是查询 q_i 的点击标签的多热编码，且 $\mathbf{y}_i$ 的值域为 $\mathbf{y}_i \in [0, 1]$ 。我们使用二元交叉熵损失作为训练模型的目标。

4 实验

4.1 数据集

为了评估 SSUF 的有效性，我们在一个电子商务应用程序的用户点击日志中提取了两个大规模的真实世界数据集上进行了一系列实验。数据集的统计信息列在表 1 和 2 中。实验集中在以下两个任务上：

- **意图任务**：该任务基于用户的查询预测多种购买意图。电子商务平台由专家精心定义了一个分层的意图架构，涵盖超过 1000 种不同的用户意图。训练数据和测试数据均从历史用户点击日志中提取。
- **类别任务**：该任务旨在预测用户需求的产品类别。此前点击率较高的类别（前 95% 的点击通过率）被认为是查询的类别。

表 1: 意图分类任务的数据统计。

统计	意图任务		
	训练	值	测试
Queries	67,450,702	20,0000	31,792
Avg. chars	7.63	5.00	8.36
Total Labels	1,605	1,605	1,605
Avg. # of labels	1.04	1.67	1.91
Min. # of labels	1	1	1
Max. # of labels	7	3	16

表 2: 类别任务的数据统计。

统计	分类任务		
	训练	值	测试
Queries	113,686,150	20,0000	33,960
Avg. chars	8.50	6.53	6.02
Total Labels	6,634	6,634	6,634
Avg. # of labels	1.52	2.05	5.33
Min. # of labels	1	1	1
Max. # of labels	16	13	20

4.2 基线模型

我们将 SSUF 与多个强大的基线进行了比较，包括多标签分类方法和查询分类模型。详细介绍如下：

(1) 多标签分类基线：

- **XML-卷积神经网络 (Liu et al., 2017)**: 这是一个基于 CNN 的模型，它结合了 CNN 模型的优势，并超越了多标签共现模式。
- **学习自动机 (Wang et al., 2018)**: 这是一个标签嵌入注意力模型，它将单词和标签嵌入到同一个空间，并衡量词-标签对的兼容性。
- **LSAN (Xiao et al., 2019)**: 这是一个基于标签的注意力网络，它使用文档和标签文本通过自我和标签注意力机制来学习特定于标签的文档表示。

(2) 查询分类基线：

- **DPHA (Zhao et al., 2019)**: 它包含一个基于标签图的神经网络，并使用基于相关性的标签表示进行软训练。
- **MMAN (Yuan et al., 2023)**: 这是一个基于 BERT 的模型，它从查询-类别交互矩阵中抽取字符和语义层面的特征，以减轻非正式查询与类别之间的表达差距。
- **HCL4QC (Zhu et al., 2023)** 使用层级结构和实例层次来学习超出查询文本的信息。
- **SMGCN (Yuan et al., 2024)**: 它扩展了类别信息，并利用类别的共现和语义相似性图来增强标签之间的关系。
- **HQC (He et al., 2024)**: 它通过利用对比损失来辨别层次结构中细粒度实例关系的增强表示学习，以及关注内在标签分类学的微妙层次分类损失来使用分层信息。

4.3 实验设置

查询分类本质上是一个文本分类任务。与之前的研究所一致 (Zhang et al., 2021; Yuan et al., 2023)，我们使用微平均和宏平均的精度、召回率和 F1 分数指标来评估模型性能。

我们的模型是使用 PyTorch 框架实现的，我们使用的 Adam 算法 (Kingma and Ba, 2014) 的学习率为 $1e^{-4}$ 。BERT 嵌入具有 768 维。我们使用一个两层 GCN 从图中学习标签嵌入，嵌入维度为 768。最大查询长度设置为 20。边阈值 (α) 和 (β) 均设为 0.5，由网格搜索确定。模型训练使用预热策略，半监督阈值 (τ) 初始设置为 1.0，并在训练过程中逐渐降低到 0.8。训练进行 20 个周期，批量大小为 1024。

4.4 离线评估

4.4.1 离线性能

实验结果如表 3 所示。具体来说，我们有以下观察：(1) SSUF 在两项任务中都显示出显著的性能优势，超过了多标签基线模型。尽管

表 3: 实验结果与多标签文本分类和查询分类模型进行了比较。

模型	意图任务						类别任务					
	微小			宏			微小的			宏		
	精度	回忆	F1	精度	回忆	F1	精度	回忆	F1	精度	回忆	F1
XML-CNN	78.66	32.09	45.58	50.33	20.76	27.24	86.95	24.60	38.34	40.50	15.44	20.16
LEAM	76.22	37.21	50.01	55.11	25.72	32.40	76.79	26.68	39.60	39.40	17.19	21.31
LSAN	76.46	34.96	47.98	54.47	25.12	31.71	86.39	23.66	37.15	44.69	17.79	22.84
DPHA	77.22	36.91	49.94	55.09	25.74	32.53	87.29	22.49	35.76	36.08	13.11	17.26
MMAN	79.26	38.96	52.24	56.27	26.32	33.36	82.05	32.57	46.63	57.41	28.26	34.68
HCL4QC	74.28	40.25	52.21	54.13	31.33	37.94	79.39	33.02	46.64	54.03	30.17	36.11
SMGCN	75.83	49.91	59.72	63.18	43.90	48.54	82.51	40.05	53.92	55.83	35.62	40.15
HQC	75.02	37.03	49.58	50.28	30.87	36.77	80.87	31.03	44.85	54.73	28.74	33.98
苏夫	74.89	52.62	61.81	62.74	45.91	49.46	80.74	43.40	56.45	54.98	36.02	41.22
w/o. SE-S	73.49	50.92	60.16	59.49	41.32	45.21	79.92	41.31	54.47	54.36	34.34	39.72
w/o. SE-C	74.03	51.19	60.53	59.92	40.21	44.92	79.17	40.91	53.94	54.12	34.92	39.24
w/o. SE-H	74.32	52.02	61.20	60.33	44.02	47.29	79.32	41.88	54.82	54.43	35.13	39.95
w/o. SE	76.88	48.28	59.31	56.88	37.58	43.30	81.44	38.92	52.67	55.42	34.39	38.52
w/o. KE	74.91	49.12	59.33	56.91	42.12	45.82	81.83	39.12	52.93	55.88	35.43	39.24
w/o. LE&KE	77.03	45.05	56.85	55.49	32.21	42.36	82.02	35.35	49.41	56.02	30.51	36.47
BERT	81.28	37.59	51.41	51.63	29.97	36.84	82.83	31.99	46.15	56.72	27.80	33.80

改进查询和标签表示可以缓解由于短查询而导致的上下文信息不足问题，但它们忽略了工业应用中的复杂性。工业数据集存在类别不平衡的问题，数据分布严重偏向于热门标签，导致“马太恶性循环”。因此，如果直接将这些模型应用于在线系统中，其有效性可能会降低。

(2) 与查询分类方法相比，SSUF 在两项任务上也实现了更好的性能。如表中所示的结果，相关类别的召回率在两项任务上都获得了近 3% 的 F1 提升。虽然 HCL4QC 和 HQC 也使用层次结构来增强标签表示，但它们无法建模完整的标签关系和先验知识以打破恶性循环。此外，当查询缺乏足够的语义信息时，模型的泛化能力不足，并退化为记忆模型。SSUF 通过融合后验信号和先验知识的三个可扩展模块解决了这些问题，从而表现出更优的性能。

4.4.2 消融研究

为了发现每个模块在 SSUF 中的相对重要性，我们对其变体进行了消融研究：

- 无 KE：移除知识增强模块。
- 无 KE+LE：移除标签增强模块和知识增强模块。
- 无 KE：移除结构增强模块。
- 无 KE-S：移除结构增强模块的语义关系。
- 无 KE-C：移除结构增强模块的共现关系。
- 无 KE-H：删除结构增强模块的层级关系。
- BERT：仅保留 BERT 作为查询分类的文本编码器。

实验结果见表 3。实验结果表明：

(1) 移除 SE 时，与 SSUF 相比，在两个数据集上的性能都有轻微下降。移除共现图时也可以看到类似的现象，表明相似性或共现图包含在后验数据中被忽略的额外信息。

(2) 当我们同时消除相似性图和共现图时，性能相比完整的 SSUF 下降了超过 5%。结果

表明这两个图在类别表示学习中扮演着不同的角色。

(3) 在移除这三个模块后，我们可以看到微 F1 和宏 F1 与完整的 SSUF 相比下降了 8%。这一结果进一步证明了 SSUF 中的所有这些组件彼此提供了互补信息，并且对于查询分类是必需的。

4.5 在线评估

4.5.1 在线部署

为了降低部署延迟，SSUF 的文本编码器使用了 4 层 BERT，这与在线模型一致。此外，我们只需要缓存由 GCN 生成的类别嵌入 $H \in \mathbb{R}^{|C| \times d}$ ，而不是直接部署 GCN。这样，我们就可以在不增加任何额外计算和延迟的情况下部署 SSUF。

4.6 在线架构

图 2 展示了 SSUF 在搜索系统中的作用。当用户输入查询时，SSUF 首先预测用户的意图并识别相关的类别，并将这些信息传递给下游模块。基于向量的检索模块然后找到与这些类别相关联的项目。检索到的项目与其他检索源的项目结合，并通过一个子模块过滤掉不符合用户所需类别的项目。经过过滤的项目随后被发送到排名模块。

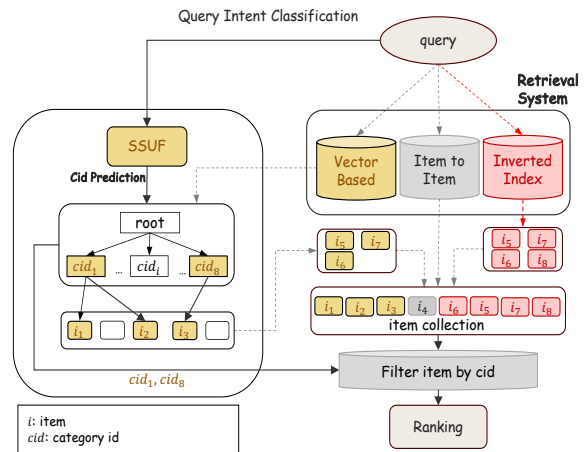


图 2: SSUF 的部署以及类别在电子商务系统中的作用。

4.6.1 在线性能

我们在广告引擎中部署了 SSUF 和基础模型进行 A/B 测试。每个模型分配了 5% 的流量。A/B 测试至少持续了一周时间。对于在线评估，我们使用了几项业务指标：展示次数 (Imp.)，点击次数，CPM（千次展示成本）以及广告收入。

表 4: 在线改进 SSUF。改进在配对 t 检验中具有统计显著性，值为 $p < 0.05$ 。（%）

Models	影响。 点击 CPM 广告收入			
Online	-	-	-	-
SSUF	+3.14	+2.72	+1.35	+4.49
w/o. SE-S	+3.07	+2.38	+0.90	+3.97
w/o. SE-C	+2.43	+2.27	+1.51	+3.94
w/o. SE-H	+2.72	+2.34	+1.13	+3.86
w/o. SE	+2.51	+2.38	+1.02	+3.53
w/o. KE	+2.67	+2.34	+0.95	+3.61
w/o. LE	+2.93	+2.47	+1.24	+4.17

如表 4 所示，SSUF 在商业指标上相较于在线模型取得了显著提升。广告展示次数和点击次数的增加表明广告系统检索到了更多相关的产品，并且这些产品与用户的偏好和搜索意图有效对齐。移除 SSUF 的任何子模块都会导致性能下降，这进一步验证了每个模块及其在 SSUF 中协同集成的有效性。

总之，离线和在线实验结果一致展示了 SSUF 的高效性、通用性和可扩展性。

5 结论

本文提出了一种用于电子商务查询分类的半监督可扩展统一框架，解决了诸如简短且模糊的查询上下文以及依赖后验点击行为等关键挑战。SSUF 集成了三个创新模块：标签增强模块、知识增强模块和结构增强模块，这些模块共同提升了查询和标签表示，打破了“马太恶性循环”，并允许在不同子任务之间灵活适应。广泛的离线和在线 A/B 测试显示 SSUF 显著超越了基线模型，验证了其有效性和实用性。SSUF 在商业电子商务平台的成功部署突显了

其巨大的商业价值。

在未来的工作中，我们计划通过整合用户特定信息和历史搜索行为来增强 SSUF，以实现个性化查询分类，旨在提高分类准确性和用户满意度。

伦理考虑

我们从以下几个方面讨论伦理问题：

- **预期用途。**如果技术按预期运行，电子商务平台的卖家和用户都能从 SSUF 模型中获益。SSUF 可以帮助客户快速识别他们想要的产品。它还通过减少在发布新产品时选择更准确产品类别所需的努力来帮助卖家。
- **失效模式。**在发生故障的情况下，SSUF 可能会输出不准确的产品信息。这些非事实的信息可能会潜在地影响用户的购物体验。系统可能会预测错误的产品类别，从而向客户推荐不需要的产品，并对其购物体验产生不利影响。

References

- Homa B Hashemi, Amir Asiaee, and Reiner Kraft. 2016. Query intent detection using convolutional neural networks. In *International conference on web search and data mining, workshop on query understanding*, pages 134–157.
- Bing He, Sreyashi Nag, Limeng Cui, Suhang Wang, Zheng Li, Rahul Goutam, Zhen Li, and Haiyang Zhang. 2024. Hierarchical query classification in e-commerce search. In *Companion Proceedings of the ACM Web Conference 2024*, pages 338–345.
- Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.
- Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Thomas N Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*.
- Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu, and Yiming Yang. 2017. Deep learning for extreme multi-label text classification. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, pages 115–124.
- Qianwen Ma, Chunyuan Yuan, Wei Zhou, Jizhong Han, and Songlin Hu. 2020. Beyond statistical relations: Integrating knowledge relations into style correlations for multi-label music style classification. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 411–419.
- Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. 2011. Classifier chains for multi-label classification. *Machine learning*, 85(3):333.
- K Sreelakshmi, PC Rafeeqe, S Sreetha, and ES Gayathri. 2018. Deep bi-directional lstm network for query intent detection. *Procedia computer science*, 143:939–946.
- Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. 2009. Mining multi-label data. In *Data mining and knowledge discovery handbook*, pages 667–685. Springer.
- Grigorios Tsoumakas, Ioannis Vlahavas, and Ioannis Vlahavas. 2007. Random k-labelsets: An ensemble method for multilabel classification. In *European conference on machine learning*, pages 406–417. Springer.
- Guoyin Wang, Chunyuan Li, Wenlin Wang, Yizhe Zhang, Dinghan Shen, Xinyuan Zhang, Ricardo Henao, and Lawrence Carin. 2018. Joint embedding of words and labels for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2321–2331.
- Lin Xiao, Xin Huang, Boli Chen, and Liping Jing. 2019. Label-specific document representation for multi-label text classification. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 466–475.
- Chunyuan Yuan, Ming Pang, Zheng Fang, Xue Jiang, Changping Peng, and Zhangang Lin. 2024. A semi-supervised multi-channel graph convolutional network for query classification in e-commerce. In *Companion Proceedings of the ACM Web Conference 2024*, pages 56–64.
- Chunyuan Yuan, Yiming Qiu, Mingming Li, Haiqing Hu, Songlin Wang, and Sulong Xu. 2023. A multi-granularity matching attention network for query intent classification in e-commerce retrieval. In *Companion Proceedings of the ACM Web Conference 2023*, pages 416–420.
- Hongchun Zhang, Tianyi Wang, Xiaonan Meng, Yi Hu, and Hao Wang. 2019. Improving semantic matching via multi-task learning in e-commerce. In *eCOM@ SIGIR*.
- Junhao Zhang, Weidi Xu, Jianhui Ji, Xi Chen, Hongbo Deng, and Keping Yang. 2021. Modeling across-context attention for long-tail query classification in e-commerce. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 58–66.

- Jiashu Zhao, Hongshen Chen, and Dawei Yin. 2019. A dynamic product-aware learning model for e-commerce query intent understanding. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1843–1852.
- Jie Zhou, Chunping Ma, Dingkun Long, Guangwei Xu, Ning Ding, Haoyu Zhang, Pengjun Xie, and Gongshen Liu. 2020. Hierarchy-aware global model for hierarchical text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1106–1117.
- Lyxing Zhu, Kexin Zhang, Hao Chen, Chao Wei, Weiru Zhang, Haihong Tang, and Xiu Li. 2023. Hcl4qc: Incorporating hierarchical category structures into contrastive learning for e-commerce query classification. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 3647–3656.