**面向内容文本相互预测的多面向密集检索预训练**

# 孙晓杰

中国科学院大学网络数据科学与技术重点实验室

中国北京sunxiaojie21s@ict.ac.cn

# 马新宇

中国科学院大学网络数据科学与技术重点实验室

中国北京xinyuma2016@gmail.com

# 毕片

中国科学院大学网络数据科学与技术重点实验室

中国北京bikeping@ict.ac.cn

# 宜兴风机

中国科学院大学网络数据科学与技术重点实验室

中国北京fanyixing@ict.ac.cn

# 郭嘉峰

中国科学院大学网络数据科学与技术重点实验室

中国北京guojiafeng@ict.ac.cn

张忠义刘蚂蚁集团

{xinzong、qishen.zqs、zhongyi.lzy}@阿里巴巴-

inc.com

## 摘要

arXiv：2308.11474 v1[cs.IR]2023年8月22日

基于预训练语言模型（PLMs），密集检索已经在纯文本上得到了广泛的研究。相比之下，使用密集模型检索具有多个方面的数据的研究很少。在产品搜索等场景中，方面信息在相关性匹配中起着重要作用，例如，类别：电子产品、计算机和宠物用品。利用方面信息进行多方面检索的一种常见方法是引入辅助分类目标，即使用项目内容来预测项目方面的注释值ID。然而，通过从头开始学习值嵌入，该方法可能无法充分捕获值之间的各种语义相似性。为了解决这一限制，我们在预训练期间利用方面信息作为文本字符串而不是类ID，以便可以在PLM中自然地捕获它们的语义相似性。为了促进方面字符串的有效检索，我们提出了项目方面文本和内容之间的相互预测目标。这样，我们的模型比对方面和内容的串联文本进行无差别屏蔽语言建模（MLM）更充分地利用了方面信息。对两个真实世界数据集（产品和小程序搜索）的广泛实验表明，我们的方法可以优于竞争基线，既将方面值视为类，又对方面和内容字符串进行相同的MLM。代码和相关数据集将在URL 1中提供。

1 https://github.com/sunxiaojie99/ATTEMPT

允许免费制作本作品的部分或全部的数字或硬拷贝供个人或课堂使用，前提是制作或分发副本不是为了盈利或商业利益，并且副本的第一页带有本通知和完整的引用。必须尊重本作品第三方组件的版权。对于所有其他用途，请联系所有者/作者。

*CIKM'23，2023年10月21-25日，英国伯明翰*

©2023版权归所有者/作者所有。ACM ISBN 979-8-4007-0124-5/23/10。

https://doi.org/10.1145/3583780.3615157

## CCS概念

* 信息系统→信息检索。

## 关键词

密集检索，多方面，预训练

**ACM参考格式：**

孙晓杰、毕克平、郭家峰、马新宇、范宜兴、山宏宇、张启申和刘忠义。2023.用于多方面密集检索的方面-内容文本相互预测预训练。第32届ACM信息与知识管理国际会议(CIKM'23)论文集，2023年10月21日至25日，英国伯明翰。美国纽约州纽约市ACM，5页。https://doi.org/10.1145/3583780.3615157

## 导言

密集检索模型【9-11，26，28，29】以预训练语言模型（PLM）【5，24】为骨干，取得了令人信服的性能。关于密集检索的研究大多集中在由纯文本组成的非结构化数据上，而对产品和人员搜索等结构化项目检索关注较少。在这些场景中，查询或项目内容之外的附加方面信息对于相关性匹配至关重要，例如brand-nike、affiliation-Stanford。然而，很少有工作探索如何在密集检索模型中有效地使用它们。

利用方面信息进行多方面检索的一种典型方法是用辅助方面预测目标来细化项目表示[12]。具体地，对于项目的每个方面，项目内容用于在训练期间预测其注释值ID。这种方法有两个主要缺点：1)它将方面的值视为孤立的类，并从头开始学习值id的嵌入，忽略它们的语义关系。例如，在类别值中，“狩猎和钓鱼”与“运动和户外”更相关，而与“宠物用品”无关。然而，如果我们将这些语义关系视为独立的类，则可能无法充分捕获它们。2）它不使用查询/项目方面如类别、品牌、颜色，

等等。这限制了潜在的检索增益。尽管在在线服务期间获取查询方面可能代价高昂，但是可以离线提取项目方面，并且如果它们已经在训练中使用，则也很容易在推理期间使用它们。

在本文中，我们提出了一种用方面-内容文本相互预测（尝试）进行预训练的方法来解决上述限制。具体来说，尝试利用方面值作为文本字符串，并使用中间的前导指示符将它们与内容连接起来。为了更有效的检索，我们专门设计了一个方面-内容相互预测目标，而不是简单地对连接的方面和内容文本进行无差别的MLM。它保持整个方面/内容，并预测内容/方面中被屏蔽的方面。此外，为了适应在线获取查询方面的开销很高的场景，我们在推理期间将查询方面文本设置为空。与普通方法相比，我们的方法有几个优点：1）在尝试中，方面值的文本重用了来自强大PLM的to-ken嵌入，因此可以自然地捕获值之间的语义关系。2)与内容连接，项目方面也可以在测试时间内用于相关性匹配。3）方面-内容相互预测目标促进了方面和内容在令牌级别的充分交互，为检索产生更好的项目表示，这一点得到了广泛实验结果的证实。据我们所知，还没有合适的大规模公共数据集用于多方面检索。我们通过从页面中抓取项目类别来构建这样一个数据集，以补充亚马逊ESCI数据集[19]中的方面。我们在这个改进的数据集和一个真实世界的商业小程序数据集上的实验表明，尝试在预测方面值的类别和

对方面和内容字符串进行相同的MLM。

## 相关工作

有三条工作线索与我们的研究相关。（1）多方位检索。一些工作已经利用多方面的信息在PLM出现之前对产品或实体进行排名【1,2，21】。在PLM时代[6]，对多方面检索的研究有限，直到Kong等人[12]首次尝试这样做。它们通过用项内容预测它们的值id来学习方面嵌入，并将它们融合以产生项嵌入。后来，Shan等人。[23]提出了一种使用局部方面级匹配信号来增强全局查询项嵌入匹配的微调方法。（2）多领域检索。如何有效地利用多个字段（例如，标题、正文等。）一直是一个长期的研究课题。最著名的方法是BM25F[22]。在PLM出现之前和之后，也提出了利用多场的方法[3,18,27,30]。多字段本质上是非结构化文本，基本问题是如何在匹配过程中对它们进行不同的加权。与字段不同，方面通常具有比字段文本的空间小得多的固定值集。



内容到方面

[CLS][]

$ %

#! #!

…**[]**

$

#"

…[C]

$

&

% '

& &

…

[九月]



[CLS][]

< 多 级 类别 >

[]<品牌> [九月][C]

<标题

+说明>

[九月]



服装，女人，鞋子， 阿迪达斯运动，跑步

adidas女式Cloud foam Pure跑步鞋白色6.5美国…

[CLS][]

$ %

#! #!

…**[]**

$

#"

…[C]

$

&

% '

& &

…

[九月]

方面到内容

**图1：相互预测MLM的尝试。外观和内容文本为绿色和紫色。**

## 方法论

### 初步

对于查询q或候选项，我们将内容文本（例如，查询字符串、标题、描述）表示为，将方面文本（例如，品牌、颜色和类别的值）表示为。假设或有

方面，进一步表示为1，..，。对于每个方面

（1），它有一个有限的方面值词汇表，表示为

≤≤

。先前的工作[12]通过预测对应于每个方面的注释值的IDs来合并方面信息

在空间内。相反，我们建议通过在文本和文本之间进行相互预测来预训练编码器。

## 尝试

为了自然地建模方面的各种值之间的语义关系，我们将方面值视为文本字符串，并将它们与内容文本连接起来。为了充分捕捉项目方面和内容之间的交互，我们引入了相互预测目标，如图1所示。

编码器输入。为了指示不同类型的文本段，我们在预期文本和原始内容（例如，编码器输入是

[ ] ≤ ≤ []

112233。当查询/项没有特定方面信息时，对应的方面文本

[ ] [ ] [ ] [ ][ ] [ ]

将是空的。在这种情况下，指示符令牌仍然可以学习查询/项目内容的一些隐式表示。请注意，在相关性匹配期间，我们总是将查询方面文本保持为空，以适应获取查询方面的开销较高的实际检索场景，同时也避免了潜在的语义漂移。表3将显示查询端指示符令牌（

[ ]

1，）在预训练期间单独学习有利于检索。因为尝试的其他部分正是

(≤≤)[]

和之间相同，我们以举例说明。内容屏蔽语言建模（MLM）。为了在没有任何辅助信息的情况下捕获内容令牌之间的交互，尝试对项目内容进行MLM。它随机屏蔽内容文本中的标记，并用Transformer层编码的上下文相关表示来预测屏蔽的标记[5]。对应的损失函数为：

Σ<UNK>

因此，它们的核心挑战是不同的。（3）预训练模型

用于密集检索。许多研究探索了提高PLM的密集检索能力，包括引入额外的

L(ˆ)=−

∈(ˆ)

(ˆ

\(ˆ)),

(1)

训练目标[4,13,15-17]、特殊屏蔽方案[25]、以及模型架构变化[7]等。我们的方法基于基本的双BERT编码器[5]。

其中，它表示通过随机屏蔽一些

文本中的令牌，(ˆˆ)表示屏蔽令牌，并且

\ ( )

表示中的剩余令牌。

方面到内容MLM预测。在预测内容文本中的屏蔽标记时，我们将整个方面文本作为上下文。在这种情况下，屏蔽内容令牌的预测具有额外的考虑证据，并且可以与单独的内容MLM不同。aspe∑ct-含量（a2c）损失L 2为：

L2(⊕ˆ)=−

(\(ˆ)),

(2)

**表1：总体性能。最好的结果以粗体显示。指示第一/第二/第三组中尝试和最佳基线之间的显著差异。**

方法 MA-亚马逊 支付宝

†

r@100 r@500 ndcg@50 r@100 r@500 ndcg@50

比伯特 0.6075 0.7795 0.3929 0.4464 0.6284 0.2033冷凝器 0.6091†0.7801†0.3960†0.4520†0.6423†0.2072†

其中表示串联。特别是，输入中的前导令牌[]（1）和[]将不会被屏蔽。

≤≤

⊕

∈(ˆ)

内容到方面MLM预测。内容到-

方面预测与文献[12]中的方面分类相似，都是利用原始内容来预测方面。然而，尝试预测方面文本中的屏蔽词而不是值类（ID），这以更软的方式编码方面信息。具体地，内容到方面(c2a)损失是：

Σ<UNK>

MTBERT 0.6137†0.7852†0.3969† 0.4498 0.6280 0.2064

马德拉尔 0.6088 0.7815 0.3950 0.4506†0.6383†0.2057†

比伯特-C 0.6137 0.7814 0.4005 0.4517 0.6291 0.2103

BIBERT-C(A)0.6137 0.7841 0.4019 0.4611 0.6432† 0.2091

MTBERT-C 0.6142 0.7839 0.3997 0.4391 0.6189 0.2026

MADRAL-C 0.6169†0.7850†0.4041† 0.4376 0.6141 0.2044

尝试 0.6233 0.79240.4097 0.4667 0.65920.2113

纯文本的方法。它采用MLM【5】使用查询/项目的内容文本来预训练编码器。（2）冷凝器【7】：它增加了一个

L2(ˆ⊕)=−

∈(ˆ)

(ˆ\(ˆ)⊕)。 (3)

BERT[5]下层和上层以外的令牌之间的短路，以增强最终的CLS表示。

总体学习目标。通过引入2和2，尝试可以通过双向交互将方面信息充分地合并到项目表示中。在

L L

总而言之，我们的总体预培训目标是：

L=L(ˆ)+(L2(⊕ˆ)+L2(ˆ⊕)),(4)

其中是超参数。

## 实验装置

### 数据集

我们在两个真实世界的数据集上进行模型比较：多方面亚马逊ESCI数据集（MA-Amazon）。亚马逊ESCI产品搜索[19]最初具有多语言现实世界查询，产品信息如品牌、颜色、标题、描述等。，以及4级相关性标签：精确、替代、补充和不相关。我们只使用英文部分，并通过从项目页面收集多级产品类别来丰富数据集。我们合并所有项目，得到482K唯一项目的语料库，用于预训练。为了微调，我们将原始训练集按查询分为训练集和验证集，并保留测试集，分别产生17K、3.5 K和8.9 K的查询。如[19]所示，在训练和召回计算中，我们将Exact视为相关，将其他标签视为不相关。MA-Amazon只有item aspect信息，1-2-3-4级的品牌、颜色、品类覆盖率分别为94%、67%和87%-87%-85%-71%。

支付宝搜索数据集。支付宝是一个小程序（类似应用程序的服务）搜索数据集，带有二进制手动相关性注释。预训练查询/项目语料库具有130万/1.8 0万个具有方面信息的不同查询/项目，即品牌（查询/项目的44%/0.6%覆盖率）和三级类别（查询/项目的类别1-2-3的91%-90%-56%/90%-90%-62%覆盖率）。微调数据集由训练/验证/测试集中的60K/3.3K/3.3K唯一查询组成。请注意，用于验证和测试的查询不会出现在预训练查询语料库中。

### 基线

我们将尝试与以下预训练方法进行比较（-C表示输入对方面和内容文本采用与尝试相同的串联策略）：（1）BIBERT[14，20]：BIBERT是尝试的主干，是一种普遍的密集检索

（3）BIBERT-C：仅在编码器输入上与BIBERT不同。在预训练和微调期间，它以与尝试相同的方式使用方面文本。（4）BIBERT-C（a）：它通过专门为纵横文本分配更高的掩码比率来细化BIBERT-C，这与尝试一致。（5）MTBERT[-C][12]：它在预训练期间对CLS执行额外的方面分类任务。

（6）MADRAL[-C][12]：它启动额外的多个方面嵌入，并通过预测每个方面的值类来学习它们，并将它们融合以产生最终的项目表示。

### 实施和评估细节

我们自己实现了尝试和所有基线。对于所有方法，编码器对于查询和项都是共享的。预培训。最大令牌长度为156，MA-Amazon/支付宝数据集的学习率和纪元分别设置为1e-4/5e-5和20/10。我们用Google的公共检查点初始化BERT参数，并使用Adam优化器进行线性预热。对于所有C基线和尝试，项目/查询内容的掩码比率设置为0.15/0.3，以考虑较短的查询长度。除了BIBERT-C（a）和尝试之外，它们在纵横比和内容文本之间都具有相同的遮罩比，其中纵横比文本为0.6。在等式4中设置为1.0。我们每两个时期微调预训练的模型检查点，并在验证数据集上选择最佳检查点。

微调。在这两个数据集上，所有模型都用Tevatron工具包训练了20个时期[8]。我们使用5e-6的学习率和64的批量大小。所有方法都用softmax交叉熵损失训练，具有批内负和一个硬负。请注意，我们没有对MTBERT和MADRAL使用辅助分类目标，因为没有实现显著的改进。

度量。我们报告召回@100、召回@500和ndcg@50。在MA-Amazon上计算ndcg时，遵循【19】，我们将E、S、C和I的增益分别设置为1.0、0.1、0.01和0.0。我们进行双尾t检验（p值≤0.05）以查看显著差异。

## 实验结果

### 主要结果

总体性能如表1所示。我们有以下观察结果：（1）一般来说，使用方面信息的方法优于不使用方面信息的方法，证实了方面在

**表2：MA-Amazon上各种组件选择的研究。**

**†表明比BIBERT有显著改进。**

r@100 r@500 ndcg@50

比伯特 0.6075 0.7795 0.3929

尝试 0.6233† 0.7924† 0.4097†

唯一品牌 0.5977 0.7710 0.3859

只有颜色 0.5867 0.7626 0.3773

仅cate1-4 0.6212† 0.7893† 0.4050†品牌+颜色+类别1 0.6192† 0.7863† 0.4040†品牌+颜色+类别1-2 0.6199† 0.7898† 0.4073†品牌+颜色+类别1-3 0.6223† 0.7910† 0.4092†

尝试−L2 0.6211† 0.7882† 0.4068†尝试−L2 0.6127† 0.7846† 0.3997†尝试−L 0.6145† 0.7851† 0.4013†

比伯特+ 0.6246† 0.7913† 0.4112†尝试+ 0.6393† 0.8019† 0.4257†

相关性匹配。值得注意的是，MADRAL在MA-Amazon上的表现比MT-BERT差，这可能是由于没有足够的预训练数据来充分地从头学习方面嵌入。（2）将方面信息视为文本字符串的模型（BIBERT-C/-C（A））超过了将方面值视为离散类的模型（MTBERT和MADRAL）。当方面文本的掩码比内容大时（在BIBERT-C（a）中），检索性能将得到提高。这表明，应该特别注意方面文本，以鼓励充分的学习。（3）当方面文本连接与使用方面值进行分类的方法（MTBERT和MADRAL）相结合时，检索性能并不总是变得更好。这可能是因为输入方面文本成为模型预测其相应类ID的快捷方式。当预训练数据很大时（例如，在支付宝上），这种关系更有可能被模型抓住，阻碍了有益交互的学习。（4）更强大的预训练方法（Condenser）有时比使用方面的方法（支付宝上的MTBERT和MADRAL）表现更好。注意，高级预训练技术的益处与方面信息正交，并且它们可以被组合以获得更好的性能。我们把这方面的研究留到以后的工作中。（5）总的来说，我们的尝试在两个数据集上都取得了最佳性能，显示了其专门为方面和内容的级联文本提出的预训练目标的有效性。

### 进一步分析

我们还从不同的角度进行了探讨，以验证其有效性。为了再现性，我们的分析基于MA-Amazon。唯一的例外是查询/项目方面的消融研究，因为只有支付宝两者都有。

方面的消融研究。我们在表2中研究了尝试中各种方面（从1级到4级的品牌、颜色和类别）的影响。我们发现：（1）当单独使用每个方面时，只有类别信息提高了模型的性能。这可能是因为品牌和颜色通常已经包含在项目内容中，而类别是额外的元信息。类别最重要的观察结果与[12]一致。（2）结合所有方面优于仅使用类别，表明品牌和颜色在与类别互动时可能会产生更好的效果。（3）除三级和四级结果相似外，类别信息层次越多，性能越好。虽然增加更多的类别级别可以提供更丰富的信息，但减少的覆盖范围（参见第4.1节）可能会限制好处。

**表3：支付宝上查询/项目方面的消融研究。表明比BIBERT有显著改进。**

r@100 r@500 ndcg@50

†

比伯特 0.4464 0.6284 0.2033

尝试 0.4667† 0.6592† 0.2113†

尝试 0.4563† 0.6437† 0.2105†尝试 0.4526 0.6366† 0.2059

损耗函数的消融研究。我们将三项损失中的每一项从总损失中剔除，看看它有多重要。在表2中，我们发现：（1）双向预测损失有利于尝试，排除任何一个都会导致性能下降。

（2）方面到内容（a2c）预测是最有帮助的，表明使用方面作为内容MLM预测的上下文是将方面信息注入项目的可行方法。（3）当香草MLM损失被消除时，性能也下降了很多，表明不受外部信息影响的原始内容语义也很重要。结合先进的微调技术。同意

[23]是最近提出的一种微调方法，它将局部方面查询匹配损失与原始全局查询项匹配损失结合起来。AGREE没有研究如何利用查询方面，这很适合MA-Amazon，因为它没有查询方面。由于AGREE将项目方面与内容连接起来，因此在预训练和尝试之后的微调期间很容易集成AGREE。表2中的最后一个块显示了AGREE单独和两者结合的性能。表明基于更好的微调技术，尝试可以获得更好的性能。值得注意的是，将AGREE与进行方面分类的方法相结合并不一定会带来更好的性能（检查表1中的MTBERT-C和MADRAL-C）。

查询/项目方面的消融研究。我们在表3中检查了查询和项目方面的影响。结果表明，查询方面和项目方面都对检索性能有贡献，项目方面更为重要。由于我们只在相关性匹配期间使用项目方面，查询方面只在预训练期间生效，并且可能有更少的贡献。

## 结论

在本文中，我们提出了一种有效的预训练方法，该方法使用方面作为文本字符串，并在方面和内容文本之间进行相互预测，用于多方面检索。与以前将方面值视为分类ID的方法相反，尝试可以通过它们的文本字符串捕获方面之间的语义关系，并通过相互预测在项目方面和内容之间执行更细粒度的交互。我们在两个真实世界数据集上的实验表明，尝试可以显著优于多个竞争基线。此外，我们发布了丰富的多方面亚马逊产品搜索数据集，以鼓励对多方面密集检索的研究。

## 致谢

这项工作得到了国家自然科学基金（NSFC）批准号61902381，CAS青年创新促进会批准号2021100，JCKY2022130C039和2021QY1701，联想-CAS联合实验室青年科学家项目。这项工作也得到了蚂蚁集团通过Ant创新研究计划的支持。

## 参考文献

1. 艾庆耀、瓦希德·阿齐兹、陈旭和张永峰。2018.学习异构知识库嵌入以进行可解释推荐。算法11，9（2018），137。https://doi.org/10.3390/a11090137
2. 艾庆耀、张永峰、毕和W.布鲁斯·克罗夫特。2020.具有动态关系嵌入模型的可解释产品搜索。ACM Trans。信息。系统。38,1(2020),4:1–4:29.https://doi.org/10.1145/3361738
3. 赛义德·巴拉内辛科尔丹、亚历山大·科托夫和费多尔·尼古拉耶夫。2018.用于即席结构化文档检索的注意力神经架构。第27届ACM信息与知识管理国际会议论文集，CIKM 2018，意大利都灵，2018年10月22-26日，Alfredo Cuzzocrea、James Allan、Norman W.Paton、Divesh Srivastava、Rakesh Agrawal、Andrei Z.Broder、Mohammed J.Zaki、K.Sel ç uk Candan、Alexandros Labrinidis、Assaf Schuster和Haixun Wang（编辑）。美国计算机学会，1173-1182。https://doi.org/10.1145/3269206.3271801
4. 张伟成、于、尹文文、杨和桑吉夫·库马尔。2020.基于嵌入的大规模检索的预训练任务。第八届学习表征国际会议，ICLR 2020，2020年4月26-30日。OpenReview.net.https://openreview.net/forum?id=rkg-mA4FDr
5. 雅各布·德夫林、张明伟、肯顿·李和克里斯蒂娜·图塔诺娃。2018.Bert：用于语言理解的深度双向Transformers的预训练。

CoRR abs/1810.04805(2018).arXiv：1810.04805 http://arxiv.org/abs/1810.04805

1. 范宜兴、谢晓辉、蔡银琼、陈嘉、马新宇、李祥生、张如清和郭家峰。2022.信息检索中的预训练方法。

找到了。趋势信息。后退。16,3(2022),178–317.https://doi.org/10.1561/1500000100

1. 高鲁豫和杰米·卡兰。2021.Condenser：用于密集检索的预训练架构。2021年自然语言处理经验方法会议论文集，EMNLP 2021，虚拟事件/多米尼加共和国蓬塔卡纳，2021年11月7-11日，Marie-Francine Moens，Huang，Lucia Specia和Scott Wen-tau Yih（编辑。）.计算语言学协会，981-993。https://doi.org/10.18653/v1/2021.emnlp-main.75
2. 高鲁豫、马学光、林志颖和杰米·卡兰。2022.Tevatron：高效灵活的密集检索工具包。更正abs/2203.05765(2022)。https://doi.org/10.48550/arXiv.2203.05765 arXiv：2203.05765
3. 塞巴斯蒂安·霍夫施塔特、林胜杰、杨正宏、林志颖和艾伦·汉伯里。2021.用平衡的主题感知采样有效地教授有效的密集检索器。在SIGIR’21：第44届国际ACM SIGIR信息检索研究与发展会议，虚拟活动，加拿大，2021年7月11-15日，Fernando Diaz，Chirag Shah，Torsten Suel，Pablo Castells，Rosie Jones和小川哲也酒井（编辑。）.美国计算机学会，113–122。https://doi.org/10.1145/3404835.3462891
4. 塞缪尔·胡莫、库尔特·舒斯特、玛丽·安妮·拉肖和杰森·韦斯顿。2019.使用深度预训练变压器在多句子任务中的实时推理。

CoRR abs/1905.01969(2019).arXiv：1905.01969 http://arxiv.org/abs/1905.01969

1. 奥马尔·哈塔布和马泰·扎哈里亚。2020.科尔伯特：通过BERT上的情境化后期交互进行高效和有效的段落搜索。第43届国际ACM SIGIR信息检索研究与发展会议论文集，SIGIR 2020，虚拟活动，中国，2020年7月25-30日，Jimmy X.Huang，Yi Chang，Xueqi Cheng，Jaap Kamps，Vanessa Murdock，Ji-Rong Wen和Yiqun Liu（编辑。）.ACM，39–48。https://doi.org/10.1145/3397271.3401075
2. 孔维泽、斯瓦拉吉·卡丹加、李成、莎琳·库马尔·古普塔、张明阳、徐文松和迈克尔·本德斯基。2022.多方位密集检索。在KDD’22：第28届ACM SIGKDD知识发现和数据挖掘会议，美国DC华盛顿州，2022年8月14-18日，Aidong Zhang和Huzefa Rangwala（编辑。）.美国计算机学会，3178-3186。https://doi.org/10.1145/3534678.3539137
3. Kenton Lee，Ming-Wei Chang和Kristina Toutanova。2019.弱监督开放域问答的潜在检索。《计算语言学协会第57届会议论文集》，ACL 2019，意大利佛罗伦萨，2019年7月28日至8月2日，第1卷：长篇论文，Anna Korhonen、David R.Traum和Lluís Màrquez（编辑）。计算语言学协会，6086-6096。https://doi.org/10.18653/v1/p19-1612
4. 林志颖、罗德里戈·诺盖拉和安德鲁·耶茨。2021.用于文本排名的预训练变形金刚：BERT和Beyond。摩根&克莱普尔出版社。https：

//doi.org/10.2200/S01123ED1V01Y202108HLT053

1. 陆舒琪、何迪、熊、柯、瓦利德·马利克、窦志成、保罗·贝内特、刘铁彦和阿诺德·奥弗维克。2021.少即是多：使用弱解码器预训练用于密集文本检索的强连体编码器。2021年自然语言处理经验方法会议论文集，EMNLP 2021，虚拟事件/多米尼加共和国蓬塔卡纳，2021年11月7-11日，Marie-Francine Moens，Huang，Lucia Specia和Scott Wen-tau Yih（编辑。）.计算语言学协会，2780-2791。https://doi.org/10.18653/v1/2021.emnlp-main.220
2. 马新宇、郭家峰、张如清、范、程。2022.通过对比跨度预测预训练用于密集检索的判别文本编码器。SIGIR’22：第45届国际ACM SIGIR信息检索研究与开发会议，西班牙马德里，2022年7月11日至15日，Enrique Amigó、Pablo Castells、Julio Gonzalo、Ben Carterette、J.Shane Culpepper和Gabriella Kazai（编辑）。美国计算机学会，848–858。https://doi.org/10.1145/3477495.3531772
3. 马新宇、张如清、郭家峰、范、程。2022.用于密集检索的判别自动编码器的对比预训练方法。第31届ACM信息与知识管理国际会议论文集，美国佐治亚州亚特兰大，2022年10月17-21日，Mohammad Al Hasan和李雄（编辑。）.ACM，4314-4318。https://doi.org/10.1145/3511808.3557527
4. 秦振、李忠良、迈克尔·本德斯基和唐纳德·梅茨勒。2020.为学习在个人搜索中排名匹配交叉网络。在WWW’20：2020年网络会议，台湾台北，2020年4月20-24日，Yennun Huang，Irwin King，Tie-Yan Liu和Maarten van Steen（编辑。）.ACM/IW3C2，2835-2841。https://doi.org/10.1145/3366423.3380046
5. Chandan K.Reddy、Lluís Màrquez、Fran Valero、Nikhil Rao、Hugo Zaragoza、Sambaran Bandyopadhyay、Arnab Biswas、Anlu Xing和Karthik Subbian。2022.购物查询数据集：用于改进产品搜索的大规模ESCI基准。更正abs/2206.06588(2022)。https://doi.org/10.48550/arXiv.2206.06588 arXiv：2206.06588
6. 尼尔斯·赖默斯和伊琳娜·古列维奇。2019.句子-BERT：使用暹罗BERT网络的句子嵌入。2019年自然语言处理经验方法会议暨第九届自然语言处理国际联席会议论文集，EMNLP-IJCNLP 2019，中国香港，2019年11月3-7日，Kentaro Inui，Jing Jiang，Vincent Ng和Xiaojun Wan（编辑。）.计算语言学协会，3980-3990。https://doi.org/10.18653/v1/D19-1410
7. 里多·雷南达、埃德加·梅伊和马丁·德·里克。2015.挖掘、排序和推荐实体方面。第38届国际ACM SIGIR信息检索研究与发展会议论文集，智利圣地亚哥，2015年8月9-13日，Ricardo Baeza-Yates，Mounia Lalmas，Alistair Moffat和Berthier A.Ribeiro-Neto（编辑。）.美国计算机学会，263–272。https://doi.org/10.1145/2766462.2767724
8. 斯蒂芬·E·罗伯逊、雨果·萨拉戈萨和迈克尔·J·泰勒。2004.对多个加权字段的简单BM25扩展。在2004年11月8-13日在美国DC华盛顿州举行的2004年ACM CIKM信息和知识管理国际会议论文集中，David A.Grossman、Luis Gravano、ChengXiang Zhai、Otthein Herzog和David A.Evans（编辑。）.美国计算机学会，42–49。https://doi.org/10.1145/1031171.1031181
9. 单宏宇、张启申、刘忠义、张冠南和李。2023.超越双塔：候选检索的属性引导表征学习。ACM网络会议论文集2023.3173-3181。
10. 孙宇、王硕焕、李宇坤、冯世坤、陈旭一、和。2019.ERNIE：通过知识整合增强表征。CoRR abs/1904.09223(2019).arXiv：1904.09223 http://arxiv.org/abs/1904.09223
11. 肖石涛、刘正、邵和曹赵。2022.RetroMae：通过屏蔽自动编码器预训练面向检索的语言模型。2022年自然语言处理经验方法会议论文集，EMNLP 2022，阿拉伯联合酋长国阿布扎比，2022年12月7-11日，Yoav Goldberg、Zornitsa Kozareva和张悦（编辑。）.计算语言学协会，538–548。https://aclanthology.org/2022.emnlp-main.35
12. Lee Xiong、Chenyan Xiong、Ye Li、Kwok-Fung Tang、Jialin Liu、Paul N.Bennett、Junaid Ahmed和Arnold Overwijk。2021.密集文本检索的近似最近邻负对比学习。第九届学习表征国际会议，ICLR 2021，虚拟活动，2021年5月3日至7日。OpenReview.net.https://openreview.net/forum?id=zeFrfgyZln
13. 哈米德·扎马尼、巴斯卡尔·米特拉、宋夏、尼克·克拉斯威尔和索拉布·蒂瓦里。2018.具有多个文档字段的神经排序模型。第11届ACM网络搜索和数据挖掘国际会议论文集，WSDM 2018，美国加利福尼亚州玛丽娜德尔雷，2018年2月5-9日，易昌，翟成祥，刘燕和Yoelle Maarek（编辑。）.美国计算机学会，700–708。https://doi.org/10.1145/3159652.3159730
14. 詹景涛、毛嘉新、刘益群、郭家峰、张敏和马少平。2021.用硬否定优化密集检索模型训练。在SIGIR’21：第44届国际ACM SIGIR信息检索研究与发展会议，虚拟活动，加拿大，2021年7月11-15日，Fernando Diaz，Chirag Shah，Torsten Suel，Pablo Castells，Rosie Jones和小川哲也酒井（编辑。）.

ACM，1503-1512。https://doi.org/10.1145/3404835.3462880

1. 詹景涛、毛嘉新、刘益群、张敏和马少平。2020.Repbert：用于第一阶段检索的上下文文本嵌入。CoRR abs/2006.15498(2020).arXiv：2006.15498 https://arxiv.org/abs/2006.15498
2. 张宏春、王天一、孟晓楠、胡一。2019.通过电子商务中的多任务学习改善语义匹配。在SI-GIR 2019电子商务研讨会会议录中，与第42届国际ACM SIGIR信息检索研究与发展会议同期举行，eCom@SIGIR 2019，法国巴黎，2019年7月25日（CEUR研讨会会议录，第2410卷），Jon Degenhardt，Surya Kallumadi，Utkarsh Porwal和Andrew Trotman（编辑。）.CEUR-WS.org。http://ceur-ws.org/Vol-2410/paper2.pdf