

星图3.0模型训练分享

目录

1、背景

- 1.1 概念
 - 1.1.1、Shared-bottom 模型
 - 模型结构
 - 优点
 - 缺点
 - 1.1.2 ERNIE2.0
 - 1. 持续多任务学习框架
 - 1.1. 持续多任务学习
 - 2. 预训练任务分类
 - 1.1.3、星图3.0先验模型
 - 1. 模型结构
 - 2. 详细结构
 - 3. 垂类接入
 - 4. 训练样本
 - 5. 策略收益
 - 1.1.4 在线通路、星图平台接入

2、实现

- 2.1 框架
- 2.2 流程
 - 数据准备、特征处理
 - 模型训练、推理
 - 模型merge
 - 测试上线
- 2.2.1 特征获取流程
- 2.2.2 模型训练、推理
- 2.2.3 merge

3、现状

- 3.1 操作
- 3.2 收益

1、背景

1.1 概念

1.1.1、Shared-bottom 模型

Shared-bottom 模型是一种多任务学习 (Multi-task Learning, MTL) 的神经网络架构。它主要用于同时处理多个相关任务的场景，通过共享底层的特征表示来提高各个任务的学习效率和性能。

在电商搜索中，既要提高点击率又要增加转化率；视频推荐中，既要点击率又要完播率。

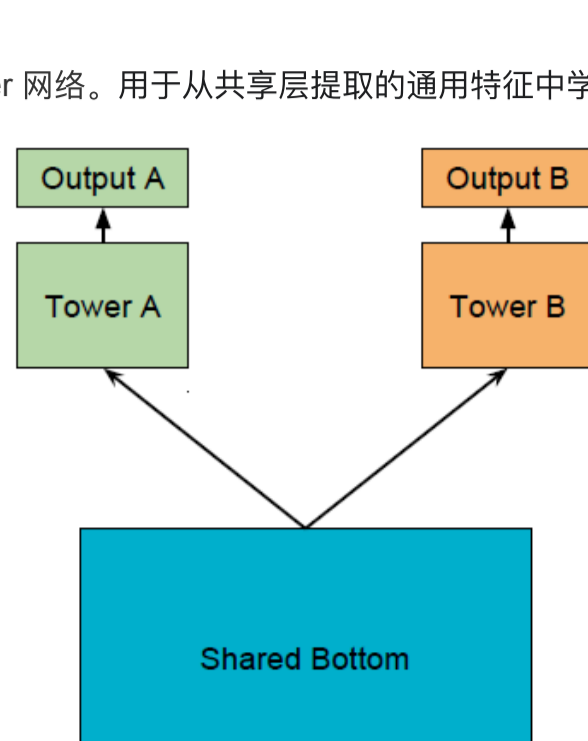
电商场景案例
当用户搜索“运动鞋”时：

1. Shared Bottom层：学习商品标题、用户画像、历史行为等通用特征
2. CTR Tower：专注用户点击偏好特征（如价格敏感度）
3. CVR Tower：侧重购买决策特征（如品牌忠诚度）

在垂类场景下，对多个垂类的模型进行训练，提升阿拉丁各垂类召回率。

模型结构

1. 底层共享部分 (Shared Bottom Layer)
 - 模型的底部，通常由一系列的神经网络层（如全连接层或卷积层）组成。提取输入数据的通用特征，所有任务共享。
2. 任务特定层 (Task-specific Layers)
 - 每个任务在共享层之上都有一个独立的 tower 网络。用于从共享层提取的通用特征中学习各自任务更具体、更相关的信息。



优点

1. 参数共享：减少参数数量，提高训练效率。
2. 信息共享：利用任务间的关联性，提升预测效果。
3. 易扩展：可以方便地增加新任务，只需为新任务添加特定分支和输出层。

缺点

1. 负迁移 (Negative Transfer)：如果任务之间的相关性较弱，共享底层可能会导致一个任务的噪声影响其他任务的表现。
2. 模型复杂性：随着任务数量增加，任务特定层和输出层会增多，增加模型的复杂性。

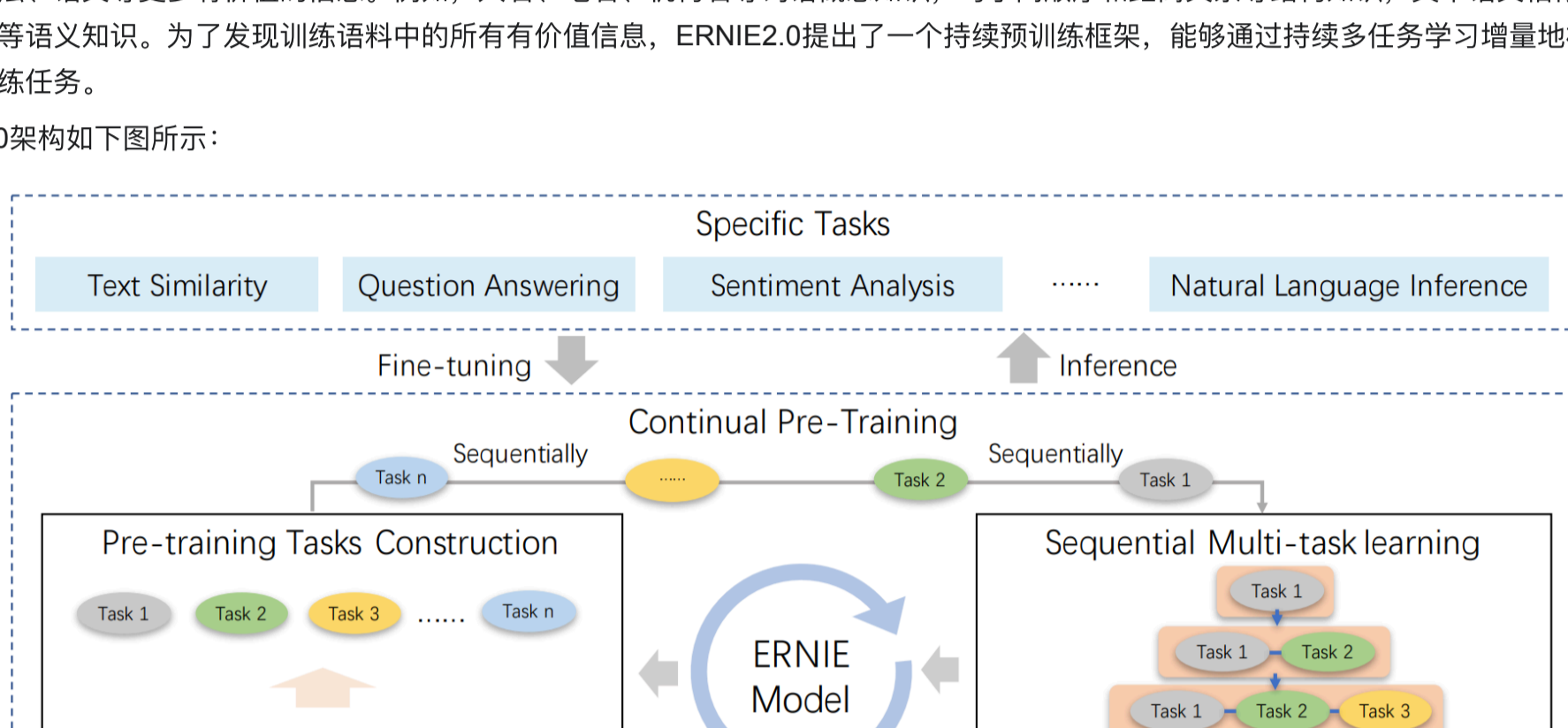
1.1.2 ERNIE2.0

ERNIE2.0于2019年7月被提出，该模型在共计 16 个中英文任务上超越了 BERT 和 XLNet，取得了 SOTA 效果。(State-Of-The-Art)

1. 持续多任务学习框架

ERNIE2.0之前的工作主要是通过词或句子的共现信号，构建语言模型任务进行模型预训练。然而，除了语言共现信息之外，语料中还包含词法、语法、语义等更多有价值的信息。例如，人名、地名、机构名等词语概念知识，句子间顺序和距离关系等结构知识，文本语义相似度和语言逻辑关系等语义知识。为了发现训练语料中的所有有价值信息，ERNIE2.0提出了一个持续预训练框架，能够通过持续多任务学习增量地构建并训练各种预训练任务。

ERNIE2.0架构如下图所示：



ERNIE 2.0 框架能通过多任务学习持续更新预训练模型，这也就是持续预训练的含义。在每一次微调中，ERNIE 会首先初始化已经预训练的权重，然后再使用具体任务的数据微调模型。持续预训练包含2个阶段，即预训练任务构建和持续预训练的多任务学习。

1.1. 持续多任务学习

对于持续的多任务学习，主要需要攻克两个难点：

1. 如何保证模型不忘记之前的任务？常规的持续学习框架采用的是一个任务接一个任务的训练，导致的后果就是模型在最新的任务上得到了好的效果但是在之前的任务上获得很惨的效果。
2. 模型如何能够有效地训练？为了解决上的问题，有人提出新的方案，我们每次有新的任务进来，我们都从头开始训练一个新的模型不就好了。虽然这种方案可以解决之前任务被忘记的问题，但是这也带来了效率的问题：每次都要从头训练一个模型，这样子导致效率很低。

针对第一个难点，ERNIE2.0的解决方案是当有新任务出现时，首先使用先前学习到的参数初始化模型，然后将新引入的任务与原有的任务同时进行训练。这可以保证已经学到的知识不被遗忘。为了提高训练的效率，ERNIE2.0为每一个任务分配了N个训练轮次 (iteration)，并且将每个任务的N个训练轮次自动分配到训练的不同阶段，这样不用每个轮次都训练所有任务，提高了效率。

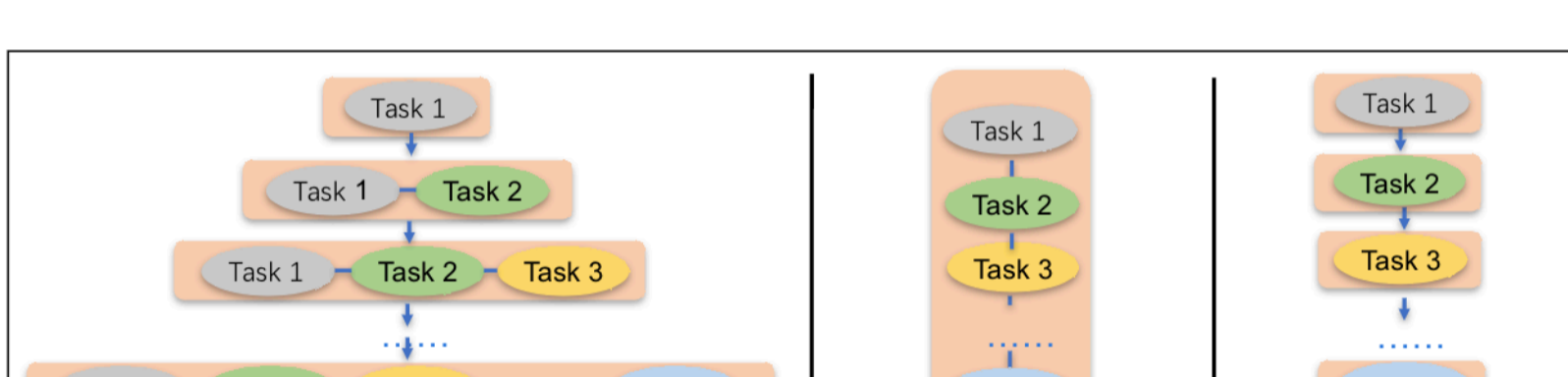
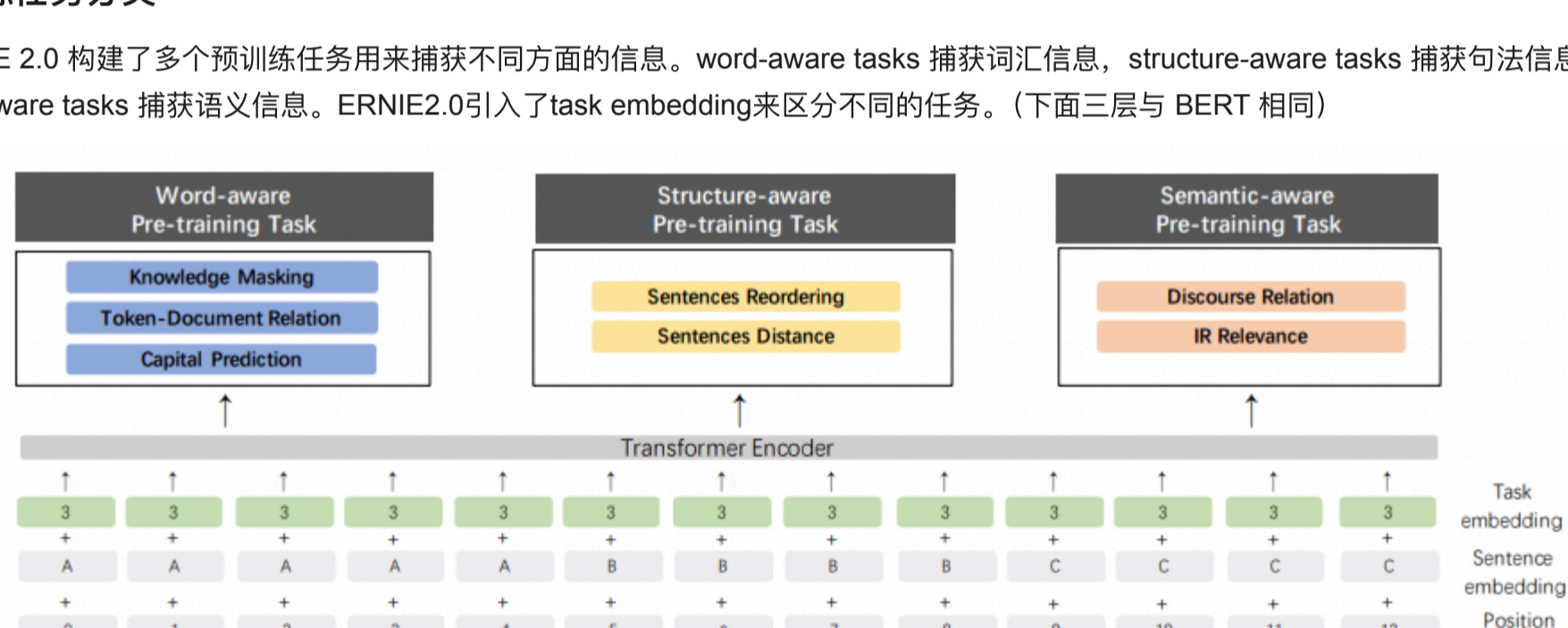


Figure 2: The different methods of continual pre-training.

2. 预训练任务分类

ERNIE 2.0 构建了多个预训练任务用来捕获不同方面的信息。word-aware tasks 捕获词汇信息，structure-aware tasks 捕获句法信息，semantic-aware tasks 捕获语义信息。ERNIE2.0引入了task embedding来区分不同的任务。(下面三层与 BERT 相同)



Word-aware Pre-training Tasks

- Knowledge Masking Task. 同 ERNIE 1.0。使用这个任务来训练模型的最初版本。
- Capitalization Prediction Task. 针对英文首字母大写的单词通常在文本中表示特殊的语义。
- Token-Document Relation Prediction Task (词频关系)。预测一个词在文中的A段落出现，是否会在文中的B段落出现。如果一个词在文章中的许多部分出现一般就说明这个词经常被用到或者和这篇文章的主题相关。通过识别这个文中关键的词，这个任务可以增强模型去获取文章的关键词语的能力。

Structure-aware Pre-training Tasks

- Sentence Reordering Task. 针对一个 paragraph，随机打乱 segments 的顺序，通过一个分类任务去预测打乱后的顺序类别。可以帮助模型学到文章的句子之间的关系。
- Sentence Distance Task. 构建一个三分类任务来判别句子的距离，0表示两个句子是同一篇文章中相邻的句子，1表示两个句子在同一篇文章中，但是不相邻，2表示两个句子属于不同的文章。通过构建这样一个三分类任务去判断句对(sentence pairs) 位置关系(包含邻近句子、文档内非邻近句子、非同文档内句子3 种类别)，更好的建模语义相关性。

Semantic-aware Pre-training Tasks

- Discourse Relation Task. 通过判断句对 (sentence pairs) 间的语义或修辞关系 (semantic & rhetorical relation)，更好的学习句间语义。
- IR Relevance Task. 从搜索引擎中拿到弱监督的数据，构建一个 query 和 title 的 3 分类任务来判断 query 和 title 的关系 (包括用户已点击、出现于结果中但用户未点击、未出现于结果中)，更好的建模短文本相关性。

相关资料：

- 论文：Ernie <https://arxiv.org/pdf/1904.09223> | Ernie 2.0 <https://arxiv.org/pdf/1907.12412> | Ernie3.0 <https://arxiv.org/pdf/2107.02137>
- github：<https://github.com/PaddlePaddle/ERNIE>
- 网络分享：机器之心 Ernie <https://www.jiqizhixin.com/articles/2019-03-16-3> | 机器之心 Ernie2.0 <https://www.jiqizhixin.com/articles/2019-07-31-10> | Ernie3.0 <https://www.jiqizhixin.com/articles/2021-07-06>

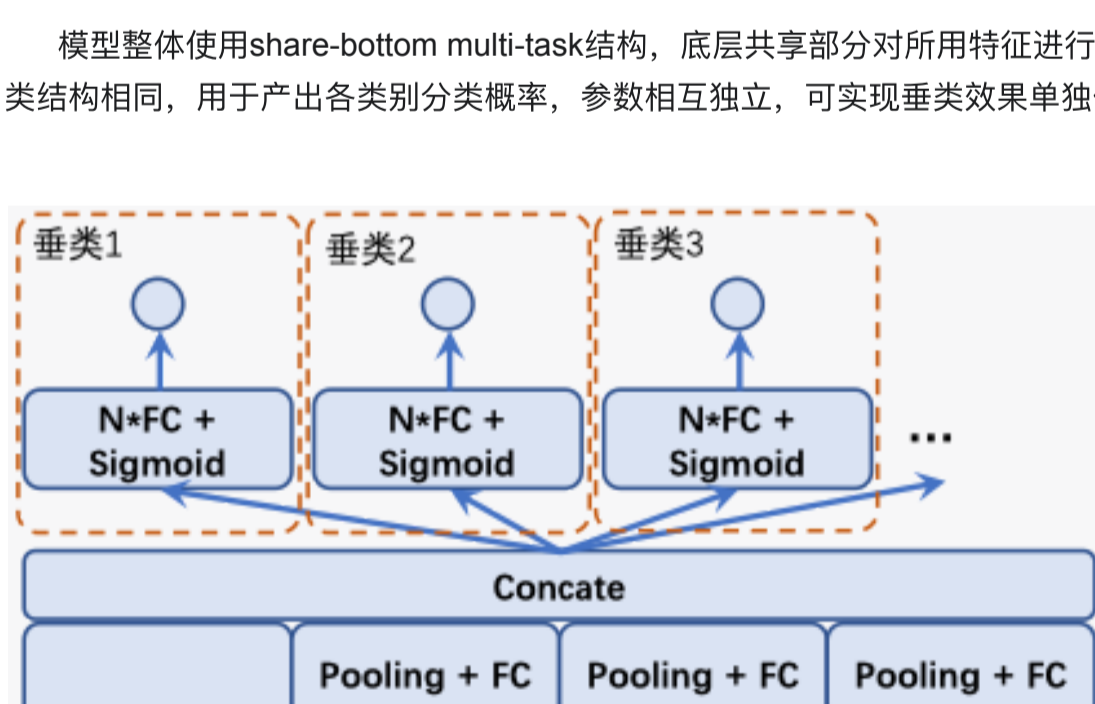
1.1.3、星图3.0先验模型

模型名：staratlas_zh_ntrigger_ernie_weixiaochi_20211115_cache

监控：http://sia.baidu.com/p/Search_ModelServer#/conf/result/dashboard/filter?name=_4143214&title=model_pv/lost_new&menuId=1622007

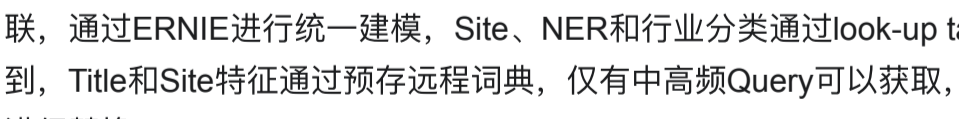
1. 模型结构

模型整体使用share-bottom multi-taskner结构，底层共享部分对所用特征进行通用embedding表示，产出query整体表示，上层multi-task部分各垂类结构相同，用于产出各类别分类概率，参数相互独立，可实现垂类效果单独优化。整体简化结构如下图所示：



底座是3层 ernie，上层接92个头（持续扩展）。使用的特征主要包括Query、Title、Site、NER和行业分类。Query和Title通过[CLS]和[SEP]级联，通过ERNIE进行统一建模，Site、NER和行业分类通过look-up table进行Embedding表示。Query、NER和行业分类特征在所有Query下都可拿到，Title和Site特征通过预存远程词典，仅有中高频Query可以获取，对于空缺Title特征，可直接省略，对于Site特征空缺，使用0向量对Embedding进行替换。

2. 详细结构



使用<https://netron.app/>打开，可以看到完整模型结构。

3. 垂类接入

星图3.0模型接入垂类：训练阶段用底座和单头组成一个单头模型，训练时把底座参数冻结，当效果达标之后，把单头的参数 merge 到线上模型上去；

merge：训练一个(n+1)个头的模型，其中n是线上模型的头数，训练的学习率是0，只起一个加载参数的作用。

4. 训练样本

星图统一触发框架以整页标准维护一套统一的样本，通过随机抽取100w Query作为统一的训练和测试集合，跑出各垂类先验触发Query，PM对触发Query进行标注。为了保证人工标注样本质量，我们对样本进行多轮Active Learning，每次人工标注样本训练完成后，通过对全部样本进行预测并降低阈值，逐步扩大样本标注范围，同时对预测值与Label不一致的Query也进行人工训练，之后再次训练，如此循环迭代。

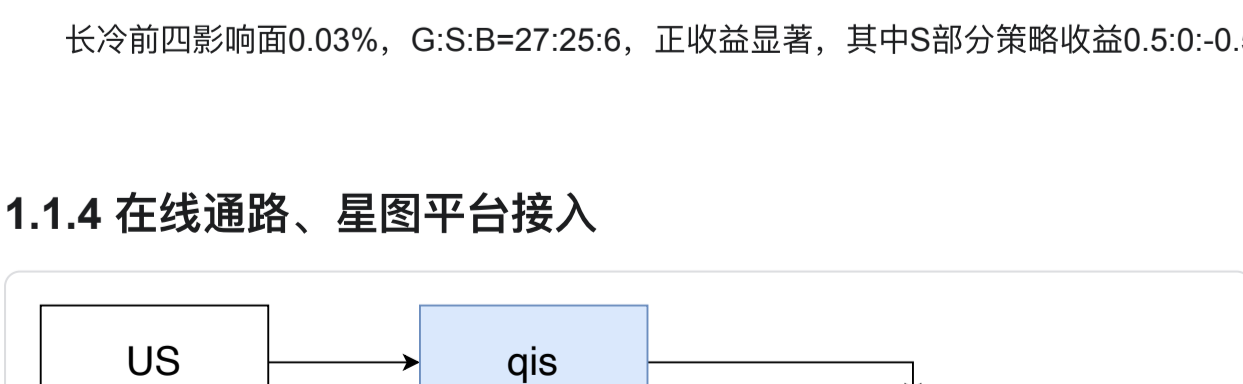
对于样本量较少的垂类，会抽取更多的线上触发Query对已有样本进行补充，保证所有垂类去重后的正样本量均大于500条，以保证模型训练效果。

5. 策略收益

随机前四影响面0.03%，G:S:B=27:20:8，正收益显著，其中S部分策略收益0.50:0.5=11:2.7

长冷前四影响面0.03%，G:S:B=27:25:6，正收益显著，其中S部分策略收益0.50:0.5=10:6.9

1.1.4 在线线路、星图平台接入

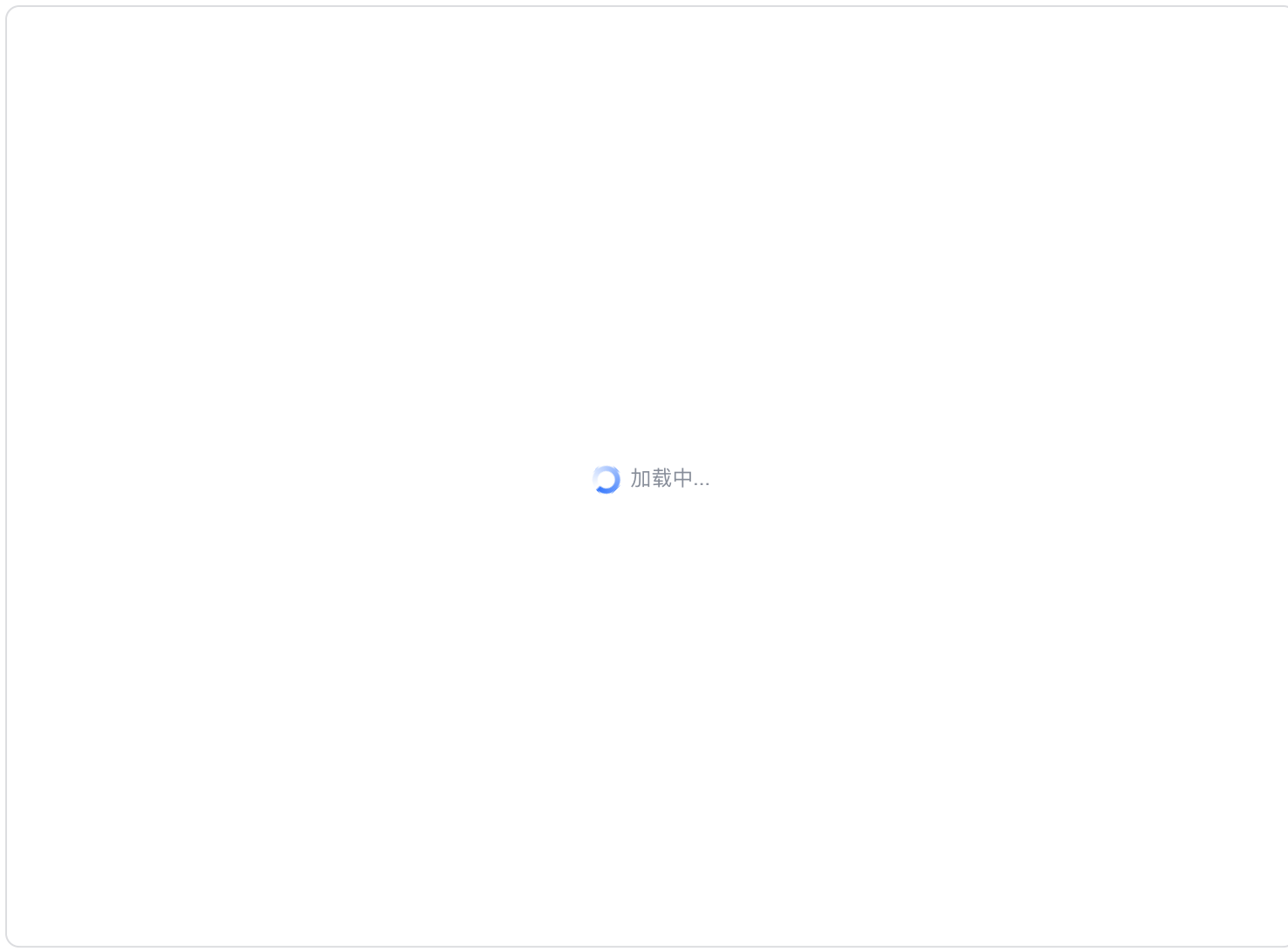


模型打分：

```
WARNING: 2025-04-09 11:00:17.17: 142358 [process@] baidu/ps-se-da-118n/strategy-request/model_star_atlas_trigge
r/processor_star_atlas_trigger_cpp:3193 [remote_data] flow_type=3, model=, srcId=, exact=model_server, 0.000000
111111:0.000000 757531:0.000000 757532:0.000000 757533:0.000000 757534:0.000000 757535:0.000000 757536:0.000000 757537:0.000000 757538:0.000000 757539:0.000000 757540:0.000000 757541:0.000000 757542:0.000000 757543:0.000000 757544:0.000000 757545:0.000000 757546:0.000000 757547:0.000000 757548:0.000000 757549:0.000000 757550:0.000000 757551:0.000000 757552:0.000000 757553:0.000000 757554:0.000000 757555:0.000000 757556:0.000000 757557:0.000000 757558:0.000000 757559:0.000000 757560:0.000000 757561:0.000000 757562:0.000000 757563:0.000000 757564:0.000000 757565:0.000000 757566:0.000000 757567:0.000000 757568:0.000000 757569:0.000000 757570:0.000000 757571:0.000000 757572:0.000000 757573:0.000000 757574:0.000000 757575:0.000000 757576:0.000000 757577:0.000000 757578:0.000000 757579:0.000000 757580:0.000000 757581:0.000000 757582:0.000000 757583:0.000000 757584:0.000000 757585:0.000000 757586:0.000000 757587:0.000000 757588:0.000000 757589:0.000000 757590:0.000000 757591:0.000000 757592:0.000000 757593:0.000000 757594:0.000000 757595:0.000000 757596:0.000000 757597:0.000000 757598:0.000000 757599:0.000000 757600:0.000000 757601:0.000000 757602:0.000000 757603:0.000000 757604:0.000000 757605:0.000000 757606:0.000000 757607:0.000000 757608:0.000000 757609:0.000000 757610:0.000000 757611:0.000000 757612:0.000000 757613:0.000000 757614:0.000000 757615:0.000000 757616:0.000000 757617:0.000000 757618:0.000000 757619:0.000000 757620:0.000000 757621:0.000000 757622:0.000000 757623:0.000000 757624:0.000000 757625:0.000000 757626:0.000000 757627:0.000000 757628:0.000000 757629:0.000000 757630:0.000000 757631:0.000000 757632:0.000000 757633:0.000000 757634:0.000000 757635:0.000000 757636:0.000000 757637:0.000000 757638:0.000000 757639:0.000000 757640:0.000000 757641:0.000000 757642:0.000000 757643:0.000000 757644:0.000000 757645:0.000000 757646:0.000000 757647:0.000000 757648:0.000000 757649:0.000000 757650:0.000000 757651:0.000000 757652:0.000000 757653:0.000000 757654:0.000000 757655:0.000000 757656:0.000000 757657:0.000000 757658:0.000000 757659:0.000000 757660:0.000000 757661:0.000000 757662:0.000000 757663:0.000000 757664:0.000000 757665:0.000000 757666:0.000000 757667:0.000000 757668:0.000000 757669:0.000000 757670:0.000000 757671:0.000000 757672:0.000000 757673:0.000000 757674:0.000000 757675:0.000000 757676:0.000000 757677:0.000000 757678:0.000000 757679:0.000000 757680:0.000000 757681:0.000000 757682:0.000000 757683:0.000000 757684:0.000000 757685:0.000000 757686:0.000000 757687:0.000000 757688:0.000000 757689:0.000000 757690:0.000000 757691:0.000000 757692:0.000000 757693:0.000000 757694:0.000000 757695:0.000000 757696:0.000000 757697:0.000000 757698:0.000000 757699:0.000000 757700:0.000000 757701:0.000000 757702:0.000000 757703:0.000000 757704:0.000000 757705:0.000000 757706:0.000000 757707:0.000000 757708:0.000000 757709:0.000000 757710:0.000000 757711:0.000000 757712:0.000000 757713:0.000000 757714:0.000000 757715:0.000000 757716:0.000000 757717:0.000000 757718:0.000000 757719:0.000000 757720:0.000000 757721:0.000000 757722:0.000000 757723:0.000000 757724:0.000000 757725:0.000000 757726:0.000000 757727:0.000000 757728:0.000000 757729:0.000000 757730:0.000000 757731:0.000000 757732:0.000000 757733:0.000000 757734:0.000000 757735:0.000000 757736:0.000000 757737:0.000000 757738:0.000000 757739:0.000000 757740:0.000000 757741:0.000000 757742:0.000000 757743:0.000000 757744:0.000000 757745:0.000000 757746:0.000000 757747:0.000000 757748:0.000000 757749:0.000000 757750:0.000000 757751:0.000000 757752:0.000000 757753:0.000000 757754:0.000000 757755:0.000000 757756:0.000000 757757:0.000000 757758:0.000000 757759:0.000000 757760:0.000000 757761:0.000000 757762:0.000000 757763:0.000000 757764:0.000000 757765:0.000000 757766:0.000000 757767:0.000000 757768:0.000000 757769:0.000000 757770:0.000000 757771:0.000000 757772:0.000000 757773:0.000000 757774:0.000000 757775:0.000000 757776:0.000000 757777:0.000000 757778:0.000000 757779:0.000000 757780:0.000000 757781:0.000000 757782:0.000000 757783:0.000000 757784:0.000000 757785:0.000000 757786:0.000000 757787:0.000000 757788:0.000000 757789:0.000000 757790:0.000000 757791:0.000000 757792:0.000000 757793:0.000000 757794:0.000000 757795:0.000000 757796:0.000000 757797:0.000000 757798:0.000000 757799:0.000000 757800:0.000000 757801:0.000000 757802:0.000000 757803:0.000000 757804:0.000000 757805:0.000000 757806:0.000000 757807:0.000000 757808:0.000000 757809:0.000000 757810:0.000000 757811:0.000000 757812:0.000000 757813:0.000000 757814:0.000000 757815:0.000000 757816:0.000000 757817:0.000000 757818:0.000000 757819:0.000000 757820:0.000000 757821:0.000000 757822:0.000000 757823:0.000000 757824:0.000000 757825:0.000000 757826:0.000000 757827:0.000000 757828:0.000000 757829:0.000000 757830:0.000000 757831:0.000000 757832:0.000000 757833:0.000000 757834:0.000000 757835:0.000000 757836:0.000000 757837:0.000000 757838:0.000000 757839:0.000000 757840:0.000000 757841:0.000000 757842:0.000000 757843:0.000000 757844:0.000000 757845:0.000000 757846:0.000000 757847:0.000000 757848:0.000000 757849:0.000000 757850:0.000000 757851:0.000000 757852:0.000000 757853:0.000000 757854:0.000000 757855:0.000000 757856:0.000000 757857:0.000000 757858:0.000000 757859:0.000000 757860:0.000000 757861:0.000000 757862:0.000000 757863:0.000000 757864:0.000000 757865:0.000000 757866:0.000000 757867:0.000000 757868:0.000000 757869:0.000000 757870:0.000000 757871:0.000000 757872:0.000000 757873:0.000000 757874:0.000000 757875:0.000000 757876:0.000000 757877:0.000000 757878:0.000000 757879:0.000000 757880:0.000000 757881:0.000000 757882:0.000000 757883:0.000000 757884:0.000000 757885:0.000000 757886:0.000000 757887:0.000000 757888:0.000000 757889:0.000000 757890:0.000000 757891:0.000000 757892:0.000000 757893:0.000000 757894:0.000000 757895:0.000000 757896:0.000000 757897:0.000000 757898:0.000000 757899:0.000000 757900:0.000000 757901:0.000000 757902:0.000000 757903:0.000000 757904:0.000000 757905:0.000000 757906:0.000000 757907:0.000000 757908:0.000000 757909:0.000000 757910:0.000000 757911:0.000000 757912:0.000000 757913:0.000000 757914:0.000000 757915:0.000000 757916:0.000000 757917:0.000000 757918:0.000000 757919:0.000000 757920:0.000000 757921:0.000000 757922:0.000000 757923:0.000000 757924:0.000000 757925:0.000000 757926:0.000000 757927:0.000000 757928:0.000000 757929:0.000000 757930:0.000000 757931:0.000000 757932:0.000000 757933:0.000000 757934:0.000000 757935:0.000000 757936:0.000000 757937:0.000000 757938:0.000000 757939:0.000000 757940:0.000000 757941:0.000000 757942:0.000000 757943:0.000000 757944:0.000000 757945:0.000000 757946:0.000000 757947:0.000000 757948:0.000000 757949:0.000000 757950:0.000000 757951:0.000000 757952:0.000000 757953:0.000000 757954:0.000000 757955:0.000000 757956:0.000000 757957:0.000000 757958:0.000000 757959:0.000000 757960:0.000000 757961:0.000000 757962:0.000000 757963:0.000000 757964:0.000000 757965:0.000000 757966:0.000000 757967:0.000000 757968:0.000000 757969:0.000000 757970:0.000000 757971:0.000000 757972:0.000000 757973:0.000000 757974:0.000000 757975:0.000000 757976:0.000000 757977:0.000000 757978:0.000000 757979:0
```

2、实现

2.1 框架



MXPS全称Model eXtensible Predictor Serving，是由PaddlePaddle深度学习框架训练产生的推理模型。即可扩展的模型预估服务。旨在通过合理的抽象与封装，能够快速满足业务方提供快速高效的模型推断需求。MXPS的目前应用范围包括大搜、多模搜索等，涵盖NLP、CV等多种业务场景。百川是离线GPU模型训练托管平台，旨在统一管理大搜、多模的GPU模型训练任务。

2.2 流程



数据准备、特征处理

数据集通常为随机 query + 业务定制数据，按比例划分为训练集、测试集、验证集。

特征处理通过查询 qae / sndb / udai 获取多个线上 dump 的特征，产出训练所需数据格式。

模型训练、推理

主要工作包括：提交模型训练任务到百川，关注训练任务详情，捞取训练日志；管理多轮训练任务中模型版本；对指定验证集提交推理任务，并展示结果模型效果。

<https://sdp.baidu-int.com/0/model-center/model-train>

模型merge

[星图3.0先验平台化-merge 阶段](#)

测试上线

[星图3.0测试、上线流程](#)

2.2.1 特征获取流程



- **sndb-client**

历史其他模块使用的sndbclient都需要搭建node服务通过proxy请求，这里采用了唯一一种无需使用proxy的方式

baidu/undb/client-go: 无需proxy, 可直连node;

[SNDB业务接入指南](#)

- **aeclient**

编译好的工具，请求线上qae，获取query对应的行业分类等等特征。通过conf配置请求bns、qps等。

- **task-queue**

由于特征获取时间较长，使用异步任务队列执行特征dump任务。整个流程由相应的worker完成。

[分布式任务队列分享](#)

- **bos**

所有数据集对应的特征文件都存储于bos，特征文件名与数据集名一一映射

特征文件路径：starmap.bj.bcebos.com/starmap/model_train/dataset_feature/{class_id}_{datasetname}

2.2.2 模型训练、推理

星图3.0模型主要基于平台（GDP）、训练服务（FastApi）以及百川完成整体模型训练流程。

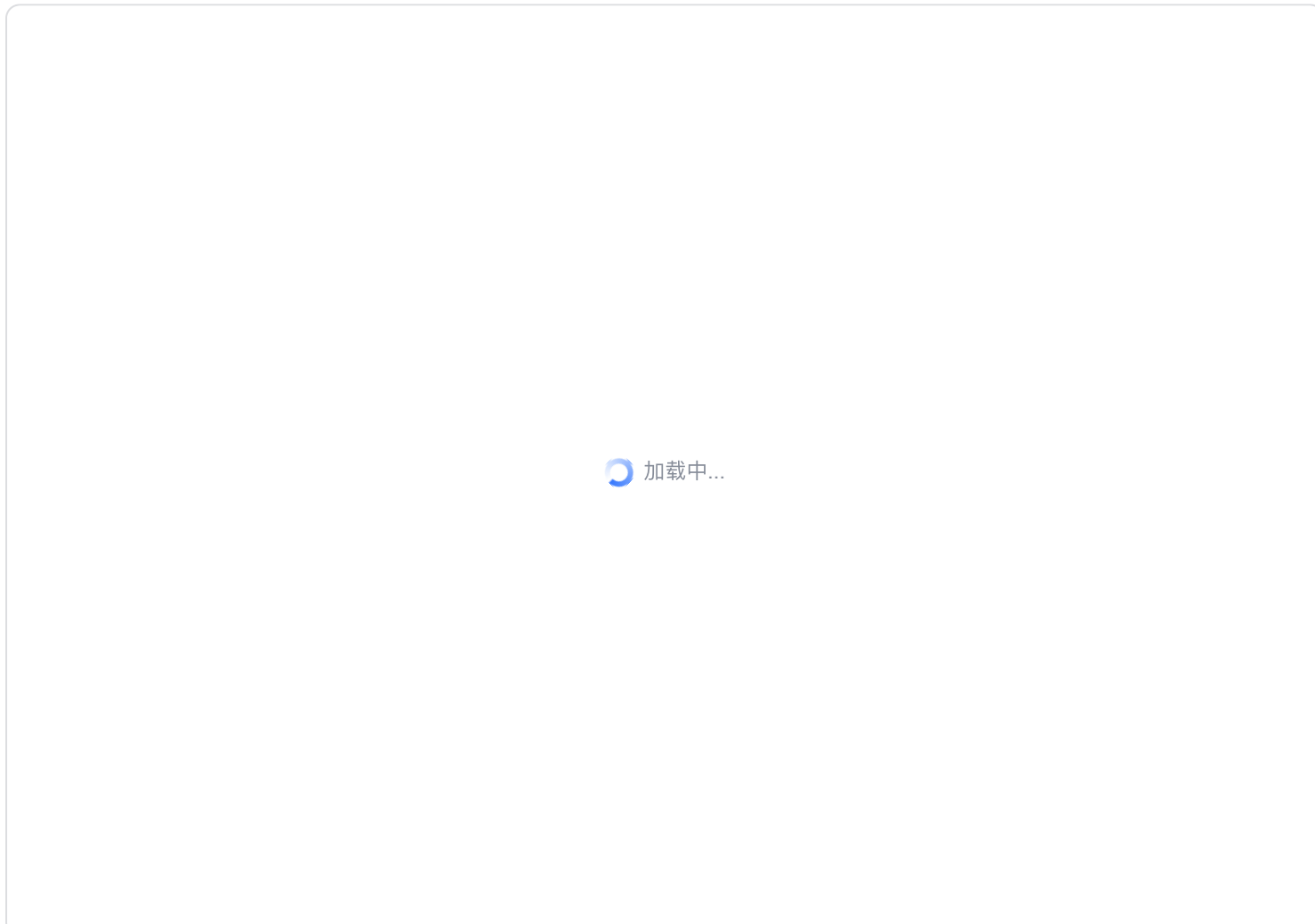
Mysql存储训练相关记录，AFS存储模型相关所有数据，BOS存储模型训练结果供平台接口使用

训练流程:



推理流程类似，都是平台提交任务后训练服务将数据推到baichuan最后返回结果。

2.2.3 merge



3、现状

3.1 操作

地址：<https://searchx.baidu-int.com/starMap/modelTrain>

- 操作页面

The screenshot shows the '3.0 Model Training' interface. It includes a '数据集' (Dataset) table with columns for ID, name, type, operator, and update time. Below it is the '训练模型' (Train Model) section with a version name input field and buttons for '配置数据集', '配置参数', '开始训练', '查看训练效果', and '保存模型'. A 'merge' section at the bottom shows a success message and a '开始merge' button.

3.2 收益

业务：
 效率：
 上线2个月，平台产出数据集150+，特征dump200+次，模型训练38次，产出16个模型，覆盖情感、q2c、贴吧、旅游、爱企查等垂类。
 效率统计如下表，可以看到，从数据集创建到产出模型完整流程，最快可在2小时内完成。
 除去异常数据（使用老数据集测试），整体平均耗时13.7538小时。
 以往产生一个3.0模型大概需要 2-3周，效率提升显著。且整体操作较为易用，业务方可自助训练，解放策略人力。

策略介入训一个符合预期的模型大概得2-3周，现在策略不用介入了，业务方自助，解放了策略的人力

	model_name	dataset_create_time	train_time	save_time	total	total/3600
1	travelknowledge_1_1740565830_step_25200	1740564435	1740565830	1740571190	6755	1.8764
2	travelknowledge_1_1740565830_step_25961	1740564435	1740565830	1740571190	6755	1.8764
3	entperpos_limeijia_v2_1742440276_1742455861_step_62500	1742453700	1742455862	1742463285	9585	2.6625
4	travelknowledge_0310_1741597547_1741602279_step_42000	1741596344	1741602279	1741614022	17678	4.9106
5	shortplay_hezhixingv3_1740587602_1740907732_step_318750	1740903905	1740907732	1740925999	22094	6.1372
6	shortplay_hezhixingv3_1740587602_1740907732_step_362131	1740903905	1740907732	1740925999	22094	6.1372
7	shortplay_hezhixingv3_1740587602_step_243750	1740583659	1740587603	1740644782	61123	16.9786
8	shortplay_hezhixingv3_1740587602_step_37500	1740583659	1740587603	1740644782	61123	16.9786
9	entperpos_limeijia_1741685715_1741694115_step_62501	1741693614	1741694115	1741768852	75238	20.8994
10	zxxtest_zhangxin14_1738841380_step_400	1738834804	1738841381	1738911282	76478	21.2439
11	travelknowledge_1_1740565830_1740642230_1740720560_step_23171	1740720281	1740720561	1740810498	90217	25.0603
12	shortplay_hezhixingv2_1740494562_step_559471	1740488968	1740494562	1740586238	97270	27.0194
13	shortplay_hezhixingv2_1740494562_step_525000	1740488968	1740494562	1740586238	97270	27.0194
14	zhangxin14_lvyou_retrain_v2_1742189081_step_11951	1741596344	1742189081	1742200828	604484	167.9122
15	zxxtest_zhangxin14v2_1739504566_step_500	1738834804	1739504566	1739880414	1045610	290.4472
16						