

工业领域下的分布式 索引算法

测试方案

2023 年 12 月

目 录

1	背景介绍	1
2	性能测试	1
2.1	测试内容	2
2.1.1	工业领域百万级数据分布式存储	2
2.1.2	工业领域索引算法性能检测	3
2.1.3	工业场景下准确的数据推荐模型	4
2.1.4	工业领域意图识别模型	5
2.1.5	工业领域推荐算法	6
2.1.6	工业领域知识图谱推理算法	7
2.2	测试方法	4
2.2.1	工业领域百万级数据分布式存储	4
2.2.2	工业领域索引算法性能检测	5
2.2.3	工业场景下准确的数据推荐模型	4
2.2.4	工业领域意图识别模型	5
2.2.5	工业领域推荐算法	6
2.2.6	工业领域知识图谱推理算法	7
3	测试环境	25
4	测试时间地点	28
5	参考文献	28

1 背景介绍

随着技术的不断进步，工业场景下检索也越来越受到关注了。人们不仅仅满足于模型在标准数据集上的表现，更希望它们在真实的工业应用环境中也能够稳定可靠地工作。然而，实现这一目标面临着一些重要的挑战。工业环境与干净的标准训练数据集存在着差异，数据的分布情况可能会发生变化，来源也更为丰富，这意味着我们思考如何更好地设计系统和算法。

随着深度学习技术的发展，工业分布式检索在实践中取得了一些重要的突破。然而，由于缺乏对先验知识和数据本身的整理，工业分布式检索的过程仍然比较复杂，很难在不同的工业场景中进行迁移和推广。这导致了算法在做出决策和检索时的性能不佳，限制了工业分布式检索在各种应用场景中的适用性。

2 性能测试

本课题以工业多场景下的数据源作为来源，分别研究了工业数据分布式存储和索引算法、意图识别、推荐推理算法。本测试大纲规定了测试内容、测试方法、技术指标、测试环境和条件。

2.1 测试内容

2.1.1 工业领域百万级数据分布式存储

项目的数据来源于多个工业场景，涉及装备制造、行业制药等领域，同时涵盖了生产、检测、归因等多个环节。在实际生产过程中，各个企业会将不同流程的数据置放于不同的库中，这要求我们在对多个企业的数据进行处理时，客观上必须进行分布式的存储，从而保证必要的安全。

项目的数据量达到了百万量级，最直接的诉求在于数据本身的表示以及处理上。对于企业而言，往往会将结构化的数据直接存储在数据库中，而其中又夹杂了大量自动化数据。因此，在我们处理数据时，要先对数据的质量进行筛选，剔除一部分“脏”数据，同时，整理数据属性之间的关系图，构建图谱，从而为后续实现多源的检索打下了基础。

为了让数据本身是否达到数量和分布式的要求，项目将数据分为了结构化数据和图谱数据进行分别存储，并加以实现检索，同时加入了决策库等工业知识作为补充。

将工业数据处理整理到库中后进行测试，拟达到的性能指标：实现数据的分布式存储和数据总量达到百万级。

指标名称：

工业领域百万级数据分布式存储。

完成时指标值：

- 1、工业数据量达到百万级；
- 2、工业数据根据自身结构存储在多个库中。

具体测试内容：

1、测试结构化数据（江中、保隆、商飞）的数据规模；测试数据特征主要包括数据的存储信息、属性特征等特点。

2、测试图谱化数据（上汽）的数据规模；测试数据特征主要包

括数据的存储信息、属性特征等特点。

2.1.2 工业领域索引算法性能检测

索引算法的设计和实现是让用户按照自然语言的输入，能够高效地从相关的库中查询到较高匹配度的答案的关键技术。而索引是一种常见的数据库优化手段，设计优良的索引能够大幅提升数据库的查询效率，提升并发能力。本项目根据不同场景、不同类型下的数据，构建不同的索引，针对结构化数据，选取实际场景中具有唯一性标识的属性（工单、物品等）建立索引，针对图谱化数据，选取实体节点建立索引。

将算法集成到应用示范平台后进行测试，拟达到的性能指标：在三个任务中系统根据索引进行随机检索测试，结果准确率高于80%。

指标名称：

工业领域索引算法性能检测。

完成时指标值：

- 1、结构化数据索引检测性能大于65%；
- 2、图谱化数据索引检测性能大于65%。

具体测试内容：

基于工业领域数据集（江中、保隆、上汽、商飞），分别从不同库中随机对索引进行抽样检索，记录下随机抽样的索引值以及返回的结果。

2.1.3 工业场景下准确的数据推荐模型

在设计工业领域数据推荐模型的测试方案时，关键在于验证模型是否能够准确识别和处理工业专业术语和概念，并据此提供相关且有价值的数据推荐。首先，需要准备包含丰富工业术语的文本数据集，如技术文档和研究报告，用于训练和测试模型。模型训练应重点在于使其理解行业特有的语言和概念。在性能测试阶段，通过句子相似度以及权重系数分析检测模型对专业术语的处理能力，并评估其推荐的准确性和相关性。同时，通过分析推荐成功案例的归因，验证模型在理解行业需求方面的效果。

为了全面评估模型的准确性和实用性，收集行业专家和用户

反馈至关重要，这有助于了解模型推荐的准确性和实用性。基于这些反馈，对模型进行调整和优化，确保其能够持续提供高质量的推荐。结果评估将侧重于三个主要方面：推荐准确性、相关性以及用户满意度。通过这种综合评估方法，可以确保模型不仅能处理复杂的工业数据，还能根据用户的具体需求提供高效、精准的数据推荐。

为测试工业领域数据以及相关权重系数对于提高推荐结果准确性的影响，项目将对两类推荐方法的结果进行归因分析并加以比较：一类为无相关权重支撑的推荐，另一类为有相关权重支撑的推荐。两个方法都在江中、保隆和商飞提供的数据集上进行测试。

指标名称：

数据推荐模型的推荐准确性。

完成时指标值：

相比无相关权重支撑的数据推荐的准确率提高5%。

具体测试内容：

（1）测试无相关权重支撑的推荐算法对江中提供的数据集推断的准确率；测试有相关权重支撑的推荐算法对江中提供的数据集推断的准确率；计算在江中提供的数据集数据推荐任务中，有相关权重支撑的推荐算法相比无相关权重支撑的推荐算法准确率的提升效果。

（2）测试无相关权重支撑的推荐算法对保隆提供的数据集推断的准确率；测试有相关权重支撑的推荐算法对保隆提供的数据集推断的准确率；计算在对保隆提供的数据集推荐任务中，有相关权重支撑的推荐算法相比无相关权重支撑的推荐算法准确率的提升效果。

（3）测试无相关权重支撑的推荐算法对商飞提供的数据集推断的准确率；测试有相关权重支撑的推荐算法对商飞提供的数据集推断的准确率；计算在对商飞提供的数据集推荐任务中，有相关权重支撑的推荐算法相比无相关权重支撑的推荐算法准确率的提升效果。

2.1.4 工业领域意图识别模型

意图识别模型的设计和实现是让系统能够高效地根据自然语言输入，准确地识别用户意图的关键技术。

针对工业领域文本数据，标注出问答语句的意图，基于深度学习模型，本项目构建了工业领域意图识别模型。模型最终能够准确识别用户意图的比例高于65%。

指标名称：

工业领域意图识别算法性能检测。

完成时指标值：

意图识别准确率达到65%以上。

具体测试内容：

基于工业领域数据集（保隆和上汽），从中选择问句进行随机抽样测试，对验证集中的数据进行意图识别并返回识别结果。

2.1.5 工业领域推荐算法

在工业应用中，推荐系统通常可分为两部分：召回和排序。推荐技术主要处理的是问答或搜索系统的搜索结果候选集。利用推荐模型将搜索得到的结果集经过召回阶段和排序阶段，得到最终的结果展示给用户。

项目数据涉及工业领域的多个场景，与其他领域不同，工业领域下的推荐场景复杂且专业性较强，推荐结果需要包含多方面的内容，而不是单值对应，因此工业领域的推荐算法不仅要满足快速响应以及降低信息过载的要求，还要加深对用户的了解，为其提供定制化服务。因此合适的推荐算法可以帮助优化设备维护计划，减少人工投入，提高生产效率，并确保产品质量。本项目利用决策库数据，该数据来自已完成的工单，包含工单信息、产品的配方信息、物料信息以及工艺状态信息等，通过每条信息的工单号或产品名称等属性，将相关的完整方案推荐给用户。

将算法集成到应用示范平台后进行测试，拟达到的性能指标：在已完成工单所构成的决策库数据中根据具有唯一标识性的词条进行随机测试，将推荐结果输出。

指标名称：

工业领域推荐算法性能检测。。

完成时指标值：

1. 推荐结果的数量；
2. 输出推荐方案并评价推荐结果的准确率达到65%以上。

具体测试内容：

基于工业领域数据集（江中），从决策库中进行随机抽样测试，对输入的信息进行关键词匹配并返回推荐结果。

2.1.6 工业领域知识图谱推理算法

工业领域知识图谱推理是一种基于知识图谱的智能推理技术，旨在通过整合和分析工业领域的复杂知识结构，进行逻辑推断和决策支持。这涉及建立知识建模，其中节点表示实体（如设备、工艺、质量等），边表示它们之间的关系，而推理引擎则通过这些关系进行推断，为知识补全，推荐，智能问答等下游任务提供支撑。

具体而言，先设计 Drools 推理引擎与专家先验知识规则库。其次编写业务逻辑功能对输入数据进行解析，得到规范输入结果后进行推理。最后通过接口返回。

算法的输入为图谱三元组数据，经过推理模块得到知识补全后的增强三元组，测试指标为推理出三元组的规模，规模越大说明推理补全的效果越好。

指标名称：

工业领域知识图谱推理算法。

测试指标：

推理出新三元组的规模。

具体测试内容：

基于工业故障溯源数据集（保隆），通过意图识别得到的用户意图，在推荐模块中获取相应的推荐结果，若无推荐结果则执行推理模块得到推理结果。

2.2 测试方法

2.2.1 工业领域百万级数据分布式存储

针对工业数据的分布式存储，要求合理地将数据中的信息提取出来并存储到合适的存储介质中。为了评价分布式存储的性能，采用数据量和数据源作为定量指标。

测试中对拟达到的性能指标采取下述测试步骤完成测试：

(1) 针对图谱结构的数据：进入智能制造过程知识获取平台，选择资产管理，点击图谱资产，用户即可在我的资产中查看相关的图谱信息，涉及所属资源组、使用说明书、资产描述、实体以及关系等等信息。

(2) 针对结构化的数据：输入数据来源和点击“查看库信息”按钮，用户即可查看库的信息、数据的属性、数量以及示例等等。

表 1 工业领域百万级测试数据记录表

测试项目	技术要求	检验结果	判定
百万级工业数据	1. 数据规模达到百万级； 2. 数据分布式存储。		

2.2.2 工业领域索引算法性能检测

对结构化数据和图谱化数据索引性能进行检测，对随机获得的索引进行检索、获得的结果中出现索引时则是未正确，评估指标是用户输入一个数据批量下的索引查询的成功率。

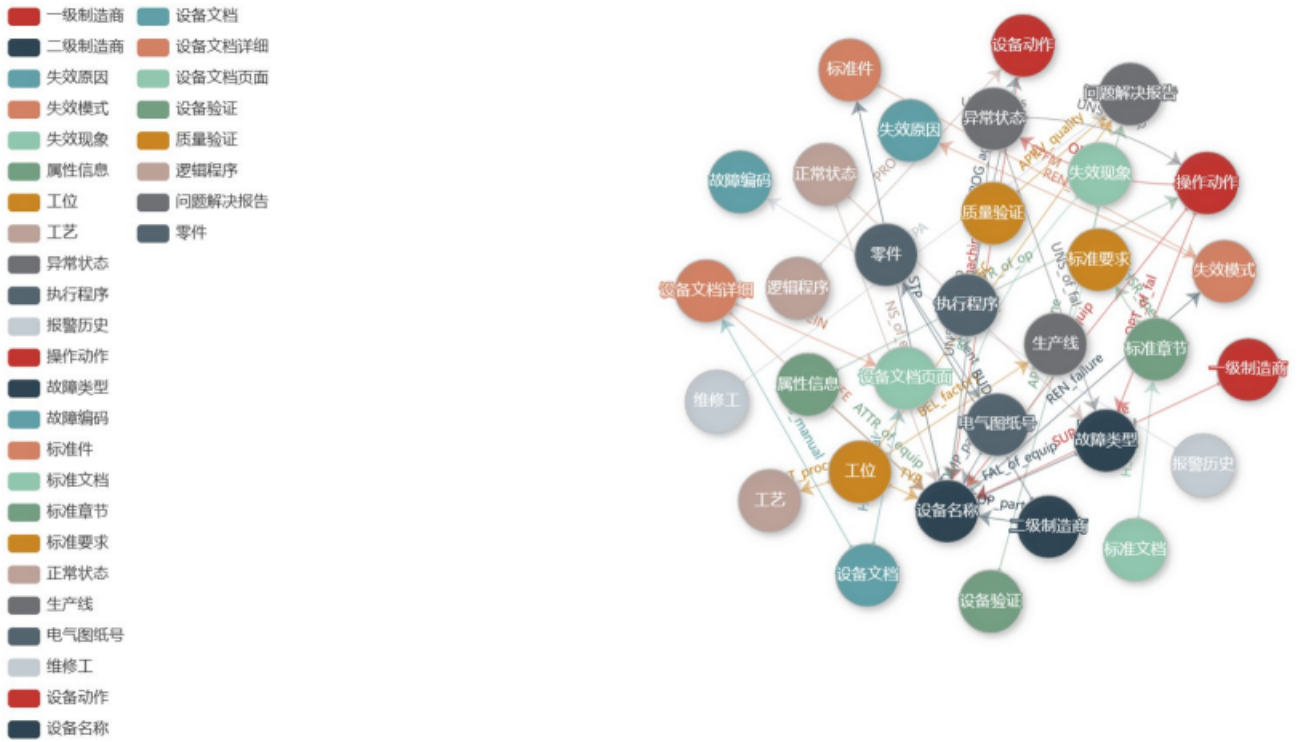
表 2 数据集介绍

测试项目	数据集名称	数据集介绍
百万级索引检索性能测试	保隆数据集	包含36514条质量溯源信息，共涉及15条产线，保存了大量有关设备故障的信息，如设备故障的起止时间、故障原因以及处理手段、造成的影响等等。
	江中数据集	1.江中决策库：包含105370条生产决策信息，主要有工单信息、产品的配方信息、物料信息以及工艺状态信息等等。 2.江中MES库：包含799901条生产进程信息，主要有组件信息、PI信息以及执行状态等等。
	商飞数据集	包含5067条飞机运行故障信息，主要涉及飞机故障的时间、故障描述和处理措施以及相关历史记录等等。
	上汽数据集	包含424840实体、84457关系信息，主要用于上汽通用设备维修问答辅助。

表 3 结构化数据字段展示

测试项目	数据集名称	字段展示
百万级 索引检索 性能测试	保隆数据集	"date": "2021.01.02", "line_name": "发射机05线", "device_name": "组装工位", " 故障起止时间": "20:15-20:19", "累计时间（分钟）": "4", "累计次数": "1", "是否停线": "是", "phenomenon": "胶用完", "method": "换胶", "label": "7更换耗材", "department": "生产部", "记录人": "刘宝宝"
	江中数据集 (仅展示决策库)	"工单号": "W210128031", "工单状态": "生产终止工单", "产品名称": "初元牌复合肽营养饮品(II型)半成品", "配方编码": 2077, "配方状态": "P", "配方名称": "初元牌复合肽营养饮品(II型)半成品", "物料编码": 2077, "物料状态": "V", "物料名称": "初元牌复合肽营养饮品(II型)-半成品", "设备名称": null, "设备编码": null, "设备状态": null, "工艺编码": "HRJZ_CYFHT2_MAIN_MID", "工艺名称": "初元复合肽2型半成品主PI1.0", " 工艺状态": "T", "工艺参数编码": "MES_GEN_TEMP", " 工艺参数值": 1, "工艺参数名称": "临时文本参数", "formula_code": 2077,
	商飞数据集	"发生日期": "2018/3/19", "机号": "B-3387（108）", "标题": "B-3387飞机R PACK FAIL", "故障描述": "B-3387执行成都-合肥航班，机组反映成都航前地面启动双发后，关APU时出现R PACK FAIL的EICAS信息，伴随CMS信息：ECS FAILED/WIRING FCV2 FAIL CLOSED，重置IASC1&2跳开关故障信息消失，返程正常。", "采取的措施": "航后机务按手册更换FCV2，并与B-3386对串IASC2计算机，发动机试车正常，故障未复现，计划后续飞行观察。", " 历史情况及其他信息": null, "风险描述": null, 是否重要故障": null," 多发性故障": null
	上汽数据集	"id": 541288, "entity_type_id": 3446, "entity_type_name": "异常状态", "name": " X轴脱开部分已无法收回复位", "attr": "{}", "color": null

图 1 图谱数据schema图展示



测试过程:

假设：由于保隆、江中以及商飞数据集之间存在场景、属性差异，各个测试均在各自数据集上进行测试。

输入:

- a) 场景标签label;
- b) 目标属性值f1、f2、f3;

输出: 根据准确度排名返回其结果；否则，返回一个指示未找到的值。

步骤:

1. 根据label进入对应的库，准备检索。
2. 初始化一个变量 i 为 0，表示当前查找的索引。
3. 从列表的第一个元素开始，逐个与目标属性进行比较。
4. 使用字符串相似度计算三个目标属性值与表中属性值之间的相似度L1、L2和L3,通过公式 $\text{Sim} = w1 * L1 + w2 * L2 + w3 * L3$ 计算得到相似度Sim(有关权重算法 Δ 具体的细节见下方)
5. 重复步骤 2-5 直到找到相似元素或遍历完整个库。
6. 如果遍历完整个库都没有找到相似元素，则返回没有。
7. 如果存在若干记录，则在一个数量范围内按照相似度Sim高低对结果进行排序返回。
8. 综合以上过程得到准确率公式:

$$\text{precision} = \sum_{i=1}^n \text{Sim}(f1,f2,f3) * \Delta(f1,f2,f3) (n:\text{a default value})$$

测试样例：

以下主要展示检索结果的示意图，考虑到数据集的庞大，故将整体展示分为三个数量级：千级、万级以及十万级，各自选用数据集为商飞数据集（5000左右）、保隆数据集（3万至四万）以及江中数据集（80万左右）。

图2 千级数据规模展示（商飞数据集）

B-602T
备用磁罗盘
采取措施（选择或输入...）
Search

从上到下，数据依次代表飞机号、故障主题、故障处理情况

针对查询条件【（机号：B-602T）（标题：备用磁罗盘）（故障描述：）】的结果为：

平均准确率为0.8799999999999999

针对ID为1409的记录，发生于2020/3/12，飞机号为B-602T

（117），故障标题为备用磁罗盘表盘有裂纹，故障描述为航前发现备用磁罗盘表盘有裂纹（大约2点钟位置），从表盘边缘延续至中间位置，裂纹周围无受力撞击迹象。，采取的措施为MMEL无备用磁罗盘损伤放行条目，待航材到件后更换。。，历史情况及其他信息为2020/1/1，B-603M（120）出现过一次相同问题，更换了备用磁罗盘。，风险描述为None。精度为0.9999999999999999。

针对ID为2323的记录，发生于2020/12/30，飞机号为B-602T

（117），故障标题为备用磁罗盘表盘玻璃有裂纹，故障描述为航后检查发现备用磁罗盘表盘玻璃有裂纹，裂纹周围无受力撞击迹象，裂纹处表盘光滑，无液体渗漏等其他损伤迹象。，采取的措施为按AMM34-23-01进行备用磁罗盘检查，检查正常。。，历史情况及其他信息为由于目前航材无货，按服务请求答复单临时放行。1月1日120架机表盘破损、3月12日117架机表盘裂纹、6月3日119架机背景灯不亮、12月11日147架机指示误差较大，风险描述为None。精度为0.9999999999999999。

针对ID为775的记录，发生于2019/6/27，飞机号为B-602T

（117），故障标题为登机门临近传感器碰伤，故障描述为登机门临近传感器碰伤，采取的措施为设计人员观察碰伤情况，并记录关门行程中标靶与传感器位置关系，待分析。7月7日登机门临近传感器航材到货，机务按手册更换后，测试正常。。，历史情况及其他信息为经查，未收到运营机队反映登机门临近传感器碰伤问题。，风险描述为None。精度为0.7999999999999999。

针对ID为800的记录，发生于2019/7/7，飞机号为B-602T

（117），故障标题为CAS信息R TRU FAIL，故障描述为鄂尔多斯短停时，飞机地面重新上电，APU供电时，出现CAS信息R TRU FAIL，CMS信息GCUA R TRU/AGCU/CURRENT WRG FAULT，机务按MEL办理保留。。，采取的措施为航后机务对串AGCU和GCU1，故障未转移，与B-602S对串RTRU，故障未转移，量线检查AGCU（U1-2403）电气插座（R5-2408）和R TRU（U1-2145）电气插头（P1-2432B）的PinA-A10至Pin15和PinA-B10至Pin7均导通，计划提服务请求寻求进一步排查方案。8日通过串件判断R TRU接触器故障，待件更换。。10日更换R TRU C（P/N:229CC02Y00,S/N:OFF 3203,ON 4098），历史情况及其他信息为经查，未收到运营机队反映TRU FAIL信息。，风险描述为None。精度为0.7999999999999999。

针对ID为801的记录，发生于2019/7/7，飞机号为B-602T

（117），故障标题为左侧标志灯不亮，故障描述为航后检查发现左侧标志灯不亮。。，采取的措施为由于航材缺件办理保留，计划提相关航材需求，到件后排故。7月25日更换左侧标志灯，测试正常。。，历史情况及其他信息为2018年4月10日，B-3386左右标志灯不亮，更换标志灯电源盒。，风险描述为None。精度为0.7999999999999999。

图3 万级数据规模展示（保隆数据集）

发射机05线

功率测试

测试不良

Search

从上到下，数据依次
代表产线名称、工位名
称、故障情况

针对查询条件【（产线：发射机05线）（工位：功率测试）（故障现象：测试不良）】的结果为：

平均准确率为0.9673564189136126

针对ID为56的记录，发生于2021.01.03，产线名称为发射机05线，工位名称为功率测试，故障起止时间为20:37-22:33，累计时间为21分钟，累计次数为4次，是否停线为是，故障现象为为1#暗箱测试不良，处理手段为扫码器扫码不良，调整扫码器参数，label为1设备故障，责任部门为制造工程部，记录人为刘宝宝。 精度为0.9999999999999999。

针对ID为2521的记录，发生于2021.01.23，产线名称为发射机05线，工位名称为功率测试，故障起止时间为08:25-08:30，累计时间为115分钟，累计次数为23次，是否停线为是，故障现象为为取料失败/2#暗箱测试不良，手柄唤醒，频谱仪峰值不跳动，处理手段为复位重启/排查线路，重启频谱仪，重新点检样件，label为1设备故障，责任部门为制造工程部，记录人为欧林。 精度为0.9999999999999999。

针对ID为3503的记录，发生于2021.02.04，产线名称为发射机05线，工位名称为功率测试，故障起止时间为15:40-15:50，累计时间为10分钟，累计次数为1次，是否停线为是，故障现象为为测试不良，处理手段为重新点检好坏样件，label为6可疑品推出，责任部门为制造工程部，记录人为谢世军。 精度为0.9999999999999999。

针对ID为29603的记录，发生于11-12白，产线名称为发射机05线，工位名称为功率测试，故障起止时间为8:00-20:00，累计时间为20分钟，累计次数为1次，是否停线为是，故障现象为为供应商测量加工件尺寸，处理手段为，label为1设备故障，责任部门为制造工程科，记录人为欧林。 精度为0.94916704081373。

针对ID为18222的记录，发生于2021.07.24，产线名称为发射机05线，工位名称为功率测试，故障起止时间为8:15-8:45，累计时间为30分钟，累计次数为1次，是否停线为是，故障现象为为功率值不在范围，处理手段为重启电脑，label为1设备故障，责任部门为制造工程科，记录人为金仁杰。 精度为0.9488837275810392。

针对ID为6433的记录，发生于2021.3.12，产线名称为发射机05线，工位名称为功率测试，故障起止时间为11:44-11:51，累计时间为18分钟，累计次数为3次，是否停线为是，故障现象为为放料报警，处理手段为调整放料位置，label为1设备故障，责任部门为制造工程部，记录人为金仁杰。 精度为0.9473358964890691。

针对ID为17389的记录，发生于2021.07.16，产线名称为发射机05线，工位名称为功率测试，故障起止时间为20:27-5:07，累计时间为39分钟，累计次数为5次，是否停线为是，故障现象为为功率测试值不在范围，处理手段为重新点检样件，label为4参数补偿，责任部门为工艺部，记录人为欧林。 精度为0.9467896666120881。

针对ID为16925的记录，发生于2021.07.12，产线名称为发射机05线，工位名称为功率测试，故障起止时间为10:53-11:00，累计时间为7分钟，累计次数为1次，是否停线为是，故障现象为为吸产品不良报警，处理手段为复位，label为1设备故障，责任部门为制造工程科，记录人为金仁杰。 精度为0.9466750198129747。

图4 十万级数据规模展示（江中数据集）

W211021067
P33056873732_1
HRJZ_RSJSP_04_MAIN_8_1
Search

从上到下，数据依次代表
工单名称、组件唯一ID号、
PI编码

针对查询条件【（工单号：W211021067）（组件唯一ID：P33056873732_1）（PI编码：HRJZ_RSJSP_04_MAIN_8_1）】的结果为：

平均精度为0.7

针对ID为00000599899aa8810d31e8b4851d02b3的记录，工单号为W211021067，环境ID为70，组件唯一ID为P44620541911_225140，PI编码为JZ_SLCKFY_M_NSY，PI版本号为0，step当前状态为0，新状态为1，执行次数为1，循环次数为225140，组件类型为PageStep，PI路径为P36338959458_1；，编辑时间为2021-11-07 18:04:45。 精度为0.7。

针对ID为00001f04492977a76d8d3dcf984c7cc的记录，工单号为W211021067，环境ID为70，组件唯一ID为P44620541911_171448，PI编码为JZ_SLCKFY_M_NSY，PI版本号为0，step当前状态为0，新状态为1，执行次数为1，循环次数为171448，组件类型为PageStep，PI路径为P36338959458_1；，编辑时间为2021-11-05 21:20:09。 精度为0.7。

针对ID为000048df61917bd8ad9c028a18144fa5的记录，工单号为W211021067，环境ID为70，组件唯一ID为P44620541911_220649，PI编码为JZ_SLCKFY_M_NSY，PI版本号为0，step当前状态为0，新状态为1，执行次数为1，循环次数为220649，组件类型为PageStep，PI路径为P36338959458_1；，编辑时间为2021-11-07 14:20:12。 精度为0.7。

针对ID为00009b0b6f374b137a79b679f869197a的记录，工单号为W211021067，环境ID为70，组件唯一ID为P44620541911_267573，PI编码为JZ_SLCKFY_M_NSY，PI版本号为0，step当前状态为0，新状态为1，执行次数为1，循环次数为267573，组件类型为PageStep，PI路径为P36338959458_1；，编辑时间为2021-11-09 05:26:24。 精度为0.7。

针对ID为0000b34f7cc802cd7d007c26eb2a2b51的记录，工单号为W211021067，环境ID为70，组件唯一ID为P44620541911_57591，PI编码为JZ_SLCKFY_M_NSY，PI版本号为0，step当前状态为0，新状态为1，执行次数为1，循环次数为57591，组件类型为PageStep，PI路径为P36338959458_1；，编辑时间为2021-11-01 22:27:18。 精度为0.7。

表 3 工业领域索引算法性能检测数据记录表

测试项目	技术要求	检验结果	判定
任务 1	在结构化数据集上索引检测准确率高于65%		
任务 2	在图谱化数据集上索引检测准确率高于65%		

2.2.3 数据推荐模型

针对推荐数据的准确性测试，要求将用户真正在意的数据推断出来。为了评价基于相关权重支撑的数据推荐的性能，采用命中率作为定量指标。

Hit Rate：命中率，即命中次数（推荐出来的数据中包含相关数据的次数）/测试次数。将总测试次数设为t，其中命中数据的次数设为h，则命中率为($h/t \times 100\%$)。

测试中对拟达到的性能指标采取下述测试步骤完成测试：

- （1） 无相关权重支撑的推荐方法：输入测试数据，使用Bert^[3]进行编码，计算相似系数，仅利用相似系数判断是否进行推荐，计算命中率。
- （2） 有相关权重支撑的推荐方法：输入测试数据，使用Bert进行编码，计算相似系数，利用相似系数以及相关权重判断是否进行推荐，计算命中率。
- （3） 若有相关权重支撑的推荐命中率相较于无相关权重支撑的推荐命中率提高5%，则测试通过。

表4 数据推荐的准确性测试数据记录表

测试项目	技术要求	检验结果	判定
数据推荐的准确性	相比无相关权重支撑的推荐算法命中率提高5%		

2

● 江中数据集推断的准确率

测试过程:

(1) 从江中提供的配方数据中随机抽取1万条数据进行测试, 其他数据作为训练集。

(2) 使用训练集训练无相关权重支撑的推荐算法, 使用测试集对训练后的算法进行测试, 计算算法命中率 a ; 使用训练集训练有相关权重支撑的推荐算法, 使用测试集对训练后的算法进行测试, 计算算法推断命中率 b 。计算有相关权重支撑的推荐算法推荐命中率 b 与无相关权重支撑的推荐算法推荐命中率 a 的差 c 。

(3) 过程1、2, 重复3次, 分别记录 a_1 、 a_2 、 a_3 、 b_1 、 b_2 、 b_3 、 c_1 、 c_2 、 c_3 。

预期结果:

3次提升效果 ($c/a \times 100\%$) 均大于等于5%。

● 保隆数据集推荐准确率

测试过程:

(1) 从保隆提供的故障描述数据中随机抽取2万条数据作为测试集, 其他数据作为训练集。

(2) 使用训练集训练无相关权重支撑的推荐算法, 使用测试集对训练后的算法进行测试, 计算算法命中率 a ; 使用训练集训练有相关权重支撑的推荐算法, 使用测试集对训练后的算法进行测试, 计算算法推断命中率 b 。计算有相关权重支撑的推荐算法推荐命中率 b 与无相关权重支撑的推荐算法推荐命中率 a 的差 c 。

(3) 过程1、2, 重复3次, 分别记录 a_1 、 a_2 、 a_3 、 b_1 、 b_2 、 b_3 、 c_1 、 c_2 、 c_3 。

预期结果:

3次提升效果 ($c/a \times 100\%$) 均大于等于5%。

● 商飞数据集推断的准确率

测试过程:

(1) 从商飞提供的故障描述数据中随机抽取2万条数据进行测试。

(2) 使用训练集训练无相关权重支撑的推荐算法, 使用测试集对训练后的算法进行测试,

计算算法命中率a；使用训练集训练有相关权重支撑的推荐算法，使用测试集对训练后的算法进行测试，计算算法推断命中率b。计算有相关权重支撑的推荐算法推荐命中率b与无相关权重支撑的推荐算法推荐命中率a的差c。

(3) 过程1、2，重复3次，分别记录 a1、a2、a3、b1、b2、b3、c1、c2、c3。

预期结果：

3次提升效果（c/a×100%）均大于等于5%。

2.2.4 工业领域意图识别模型

对工业故障溯源数据集进行随机测试，通过用户的输入进行意图识别，针对文本的意图和其中的实体，进行识别，将最终的意图类别输出，将最终得到的意图识别的准确率作为性能指标。

表 5 数据集介绍

测试项目	数据集名称	数据集介绍
意图识别	保隆和上汽	共包含12007条数据，共有8种意图类别，分别是：失效模式、失效原因、失效影响、处理方式、故障类别、故障起止时间、故障现象、责任部门。

在工业领域数据集（保隆和上汽）数据中，按照比例6:4，随机选择4803条语句作为验证集。模型将对文本进行意图识别，使用意图识别准确率作为评价指标，准确率即为预测正确的结果占总样本的百分率，计算公式如下式所示：

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

公式中 TP 为预测结果为正例，实际结果也为正例； TN 表示预测结果为负例，且实际结果也为负例； FP ， FN 分别代表预测结果为正例结果为负例和预测结果为负例结果正例。使用Pycharm软件，在dataset目录下选择processed,可以使用data_trains.py文件，随机挑选4803条数据传入validtsv文件，也可以手动将随机挑选传入文件中。

测试过程：

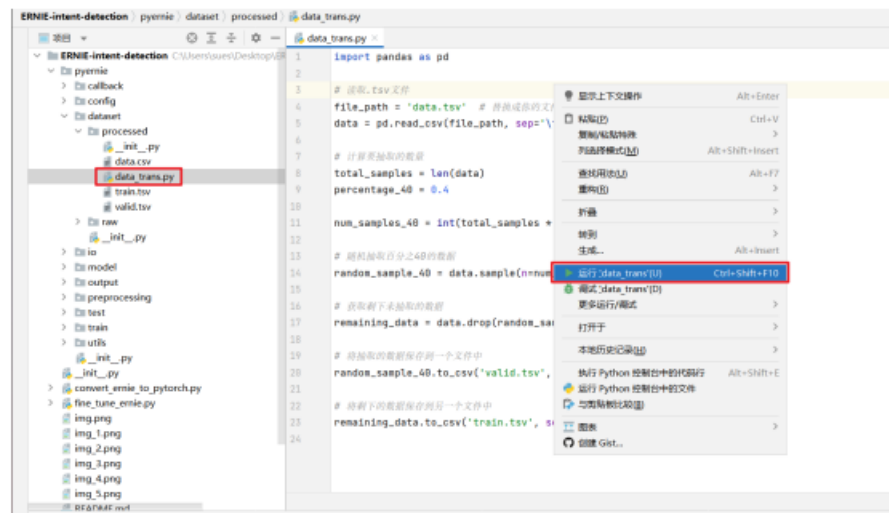


图5 生成验证集示意图

测试数据集准备完成后，点击fine_tune_ernie.py文件，并右键点击运行。

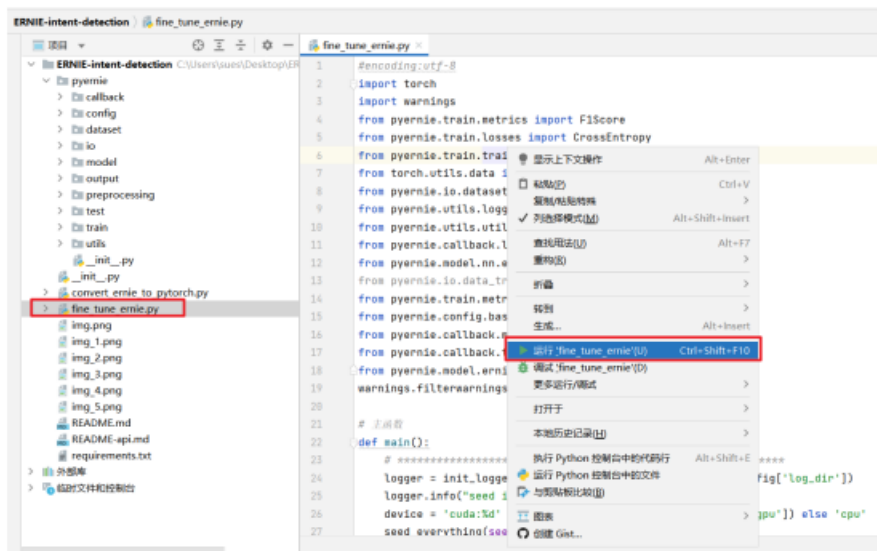


图6 意图识别结果示意图

在下方根据评价指标显示测试结果。

运行: fine_tune_ernie	1.00	1.00	1.00	1284
故障起止时间	1.00	1.00	1.00	3388
故障现象	1.00	1.00	1.00	3325
责任部门				
accuracy			0.99	7284
macro avg	0.98	0.98	0.98	7284
weighted avg	0.99	0.99	0.99	7284

-----valid result -----

classification report:

precision

recall

f1-score

support

失效模式

0.97

0.94

0.95

149

失效原因

0.93

0.82

0.87

161

失效影响

0.84

0.93

0.88

160

处理方式

1.00

1.00

1.00

860

故障类别

1.00

1.00

1.00

858

故障起止时间

1.00

1.00

1.00

862

故障现象

1.00

1.00

1.00

892

责任部门

1.00

1.00

1.00

861

accuracy

0.99

4803

macro avg

0.97

0.96

0.96

4803

weighted avg

0.99

0.99

0.99

4803

[2023-12-06 12:00:27]: ernie_trainer.py[line:179] INFO

Epoch: 3 - loss: 0.0253 - f1: 0.9794 - valid_loss: 0.0388 - valid_f1: 0.9624

[2023-12-06 12:00:27]: ernie_modelcheckpoint.py[line:48] INFO

Epoch 3: valid_loss improved from 0.04014 to 0.03877

-----valid result -----						
classification report:			precision	recall	f1-score	support
失效模式	0.97	0.91	0.94	148		
失效原因	0.70	0.73	0.71	169		
失效影响	0.95	0.64	0.77	151		
处理方式	1.00	1.00	1.00	867		
故障类别	1.00	1.00	1.00	891		
故障起止时间	1.00	1.00	1.00	828		
故障现象	0.95	1.00	0.97	903		
责任部门	1.00	1.00	1.00	846		
accuracy			0.98	4803		
macro avg	0.95	0.91	0.92	4803		
weighted avg	0.98	0.98	0.97	4803		

-----valid result -----						
classification report:			precision	recall	f1-score	support
失效模式	0.97	0.91	0.94	148		
失效原因	0.86	0.83	0.84	169		
失效影响	0.86	0.89	0.87	151		
处理方式	1.00	1.00	1.00	867		
故障类别	1.00	1.00	1.00	891		
故障起止时间	1.00	1.00	1.00	828		
故障现象	0.99	1.00	1.00	903		
责任部门	1.00	1.00	1.00	846		
accuracy			0.99	4803		
macro avg	0.96	0.95	0.96	4803		
weighted avg	0.99	0.99	0.99	4803		

图7 意图识别算法模型测试结果

表 6 工业领域意图识别算法性能检测数据记录表

测试项目	技术要求	检验结果	判定
工业领域意图识别算法性能检测	1、意图识别的准确率达到65%以上。		

2.2.5 工业领域意图识别模型

对决策库数据进行随机测试，通过用户的输入进行关键词匹配，针对数据中包含的产品名称或工单号等具有标识性的信息进行检索，将结果进行召回及排序，将最终的结果输出，将是否能输出具体的推荐方案作为性能指标。

表 7 数据集介绍

测试项目	数据集名称	数据集介绍
工业领域推荐算法性能测试	江中决策库	包含105370条生产决策信息，主要有工单信息、产品的配方信息、物料信息以及工艺状态信息等等。

测试过程：

1. 在江中数据的决策库中，筛选具有唯一性的标识信息作为每个工单的特征（如产品名称），并以此作为关键词匹配的依据。当用户输入完成后，模型将对数据进行召回，因为数据量过大难以进行更加精确的偏好和质量评估，因此，在排序阶段需要对召回结果进行统一准确的打分排序。
2. 用户输入待查的关键词，算法通过关键词匹配的方式对数据进行查询。

请输入产品名称：**健胃消食片**

图8 关键词输入示意图

3. 系统将根据用户的输入在各词条对应的数据中进行比对，返回查询到的结果数量并将这些结果进行排序，然后将最合适的方案输出。

请输入产品名称：**乳酸菌素片**
与该产品相关的信息共有 71299 项
以下是该产品的推荐方案：
工单号为W23051053的工单状态为生产终止工单，该工单生产的产品是乳酸菌素片(两盒装)0.4g*8片*6板*150盒/箱，相应的配方名称为乳酸菌素片0.4
方案完善度：较完善
准确率：72.56

图9 推荐结果及完善度评价



图10 测试样例（健胃消食片）

表 8 工业领域推荐算法性能检测数据记录表

测试项目	技术要求	检验结果	判定
工业领域推荐算法性能测试	1、推荐结果的数量； 2、输出推荐方案并评价推荐结果的准确率达到65%以上。		

2.2.6 工业领域知识图谱推理算法

对知识推理算法的推理补全效果进行评估，推理得到的新三元组规模越大说明推理效果越好，在有限的原始输入的情况下，得到的结果越多，说明规则库中的规则质量越高。

表9 数据集介绍

测试项目	数据集名称	数据集介绍
工业领域知识图谱推理算法	保隆数据集	包含14万条三元组信息，4万多个实体和10种类型的关系。

测试过程：

1. 通过对保隆故障溯源知识图谱与故障溯源业务的分析，抽取出先验规则。
2. 使用数据处理技术将规则处理成推理引擎需要的标准格式。
3. 通过意图模块传入的主实体信息，查找目标子图，然后使用系统的推理模块进行推理。

4. 系统返回经过规则补全后的目标子图数据。
5. 将原始数据规模与推理之后的数据规模进行计算的到新推出的三元组数据数量，该数量越高表明推理效果越好。

在不同的数据规模下测试结果如下：

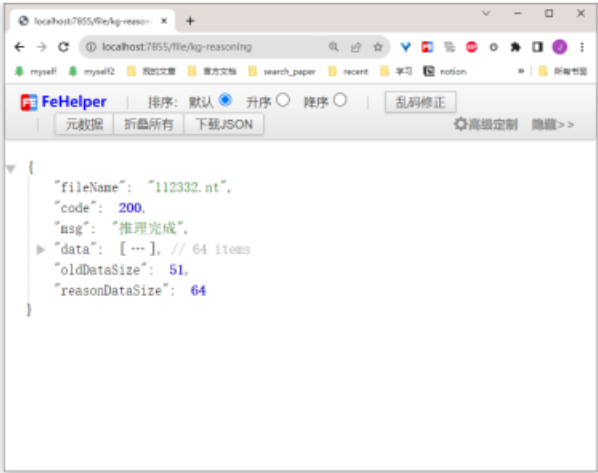


图11 推理产品ID为112332的目标图谱

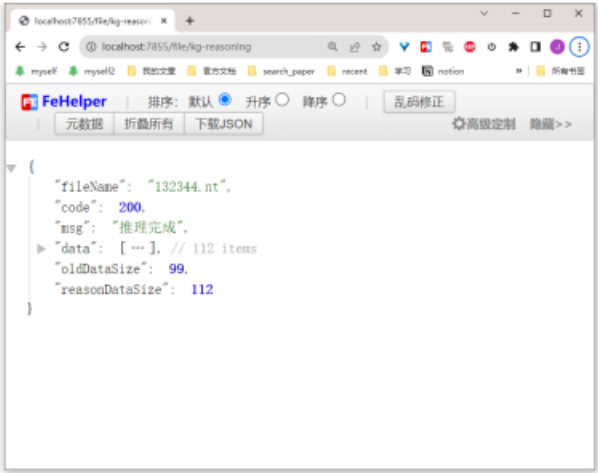


图12 推理产品ID为132344的目标图谱

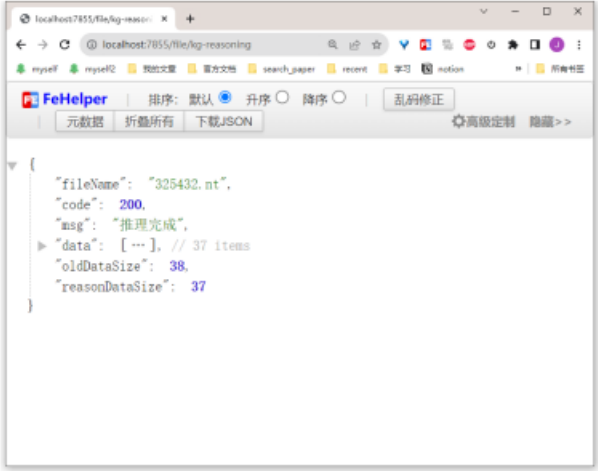


图13 推理产品ID为325432的目标图谱

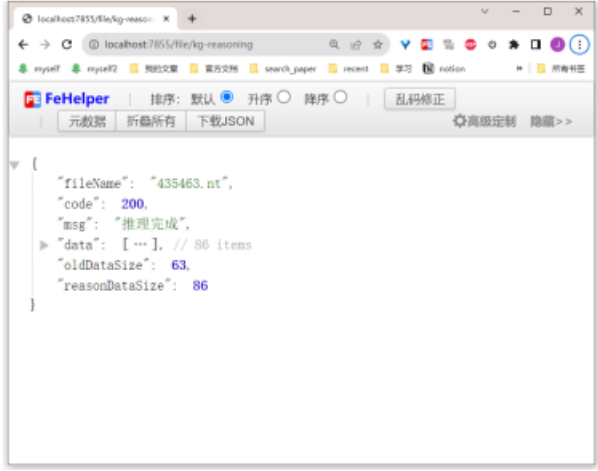


图14 推理产品ID为435463的目标图谱

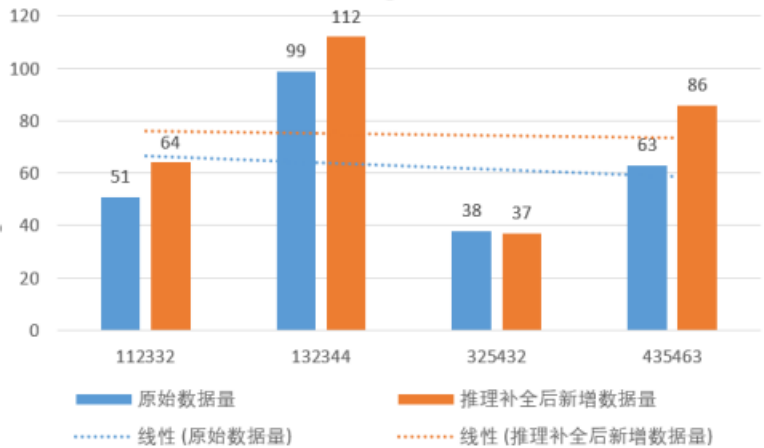


图15 不同的产品数据推理得到的结果对比图

预期结果：

推理新增数据量超过原有数据的55%。

表10 工业领域推理算法测试结果数据记录表

测试项目	技术要求	检验结果	判定
任务	在图谱化数据集上推理补全得到新三元组数量高于原始数据量的55%		

3 测试环境

3.1 工业领域百万级数据分布式存储

3.1.1 硬件环境

本测试中，硬件环境主要是指计算机的硬件配置，所采用的计算机为多核 CPU 与众核 GPU 共存的 CPU/GPU 异构并行移动工作站，其具体配置如下。

① CPU 处理器：Intel Xeon E5-2620 16 核 64 位，具体配制参数见下表。

表 5 CPU 处理器配制参数

CPU 系列	Intel Xeon E5-2620
CPU 架构	64 位
核心数量	十六核心
主频	2.10GHz

② GPU 加速部件：NVIDIA GeForce GTX Titan X，其具体配制参数见下表。

表 6 NVIDIA GeForce GTX Titan X 显卡配制参数

参数名	参数值
核心数量	3072
显存容量	12GB
显存类型	GDDR5X
显存位宽	384bit
显存带宽	336.5GB/s
最大功耗	250W

③ 内存：存储量 64GB。

3.2 工业领域索引算法性能检测

3.2.1 硬件环境

本测试中，硬件环境主要是指计算机的硬件配置，所采用的计算机为多核 CPU 与众核 GPU 共存的 CPU/GPU 异构并行移动工作站，其具体配置如下。

(1) CPU 处理器：Intel Xeon E5-2620 16 核 64 位，具体配制参数见下表。

表 7 CPU 处理器配制参数

CPU 系列	Intel(R) Xeon(R) Gold 6226R
CPU 架构	64 位
核心数量	十六核心
主频	2.90GHz

(2) GPU 加速部件：NVIDIA Corporation TU102，其具体配制参数见下表。

表 8 NVIDIA Corporation TU102 显卡配制参数

参数名	参数值
核心数量	4608
显存容量	24GB
显存类型	GDDR5X
显存位宽	384bit
显存带宽	336.5GB/s
最大功耗	280W

(3) 内存：存储量 64GB。

3.2.2 软件环境

软件环境主要包括操作系统版本、测试需要的依赖库、配套软件

设备等。

(1) 操作系统: Ubuntu 20.04.4 LTS 64 位。

(2) 软件依赖库:

1.GPU 运行环境动态链接库:

cuda64_11.dll,cublas64_11.dll,cudnn64_8.dll,curand64_11.dll,pthreadGC2.dll,pthreadVC2.dll。 2.其

他依赖动态链接库:

Pytorch

3.整体 Python 环境依赖:

Python 3.8.0+Flask

4 测试时间地点

(1)测试时间: 2023 年 12月。

(2)测试地点: 上海工程技术大学松江校区

5 参考文献

[1] Rendle S. Factorization machines[C]//2010 IEEE International conference on data mining. IEEE, 2010: 995-1000.

[2] Bordes A, Usunier N, Garcia-Duran A, et al. Translating embeddings for modeling multi-relational data[J]. Advances in neural information processing systems, 2013, 26.