

基于差分隐私 Skyline 算法的智能云平台中的数据隐私保护研究

周锴炼, 鲍 敏

(浙江理工大学机械工程学院, 浙江 杭州 310018)

✉ 2415699291@qq.com; mbao@zstu.edu.cn



摘 要:针对工业互联网云平台中不同子系统数据割裂和隐私保护问题,文章提出基于 Bubble-Skyline 算法的解决方案,该方案包括数据收集模块、数据预处理模块、数据分析建模模块。在数据收集模块和数据预处理模块中使用协议加密和哈希映射转化表,安全地传输与整合不同系统之间的数据;在数据分析模块中使用 Bubble-Skyline 算法获得最优解。采用某工厂的历史数据验证表明,在工厂排产中,该方案能够在保护数据隐私的同时优化排产方案,平均生产时间减少了 17.15%。

关键词:智能制造;云平台;Bubble-Skyline;数据割裂

中图分类号:TP311.5 **文献标志码:**A

Research on Data Privacy Protection in Intelligent Cloud Platform Based on Differential Privacy Skyline Algorithm

ZHOU Kailian, BAO Min

(College of Mechanical Engineering, Zhejiang Sci-Tech University, Hangzhou 310018, China)

✉ 2415699291@qq.com; mbao@zstu.edu.cn

Abstract: Aiming at issues of data fragmentation and privacy protection in different subsystems of the industrial internet cloud platform, this paper proposes a solution based on the Bubble-Skyline algorithm, which includes data collection module, data preprocessing module, and data analysis modeling module. Protocol encryption and Hash Map conversion table are used in the data collection module and data preprocessing module to securely transmit and integrate data between different systems. The Bubble-Skyline algorithm is used in the data analysis module to obtain optimal solutions. The validation of the historical data from a certain factory shows that the solution can optimize the scheduling scheme while protecting the data privacy, and it reduces the average production time by 17.15%.

Key words: intelligent manufacturing; cloud platform; Bubble-Skyline; data fragmentation

0 引言(Introduction)

在当今工业互联网快速发展的时代,GE Predix、Siemens MindSphere、IBM Watson IoT、Microsoft Azure IoT、Alibaba Cloud IoT 等大型企业都部署了智能云平台^[1]。由于数据具有隐私性,所以企业的物料需求计划^[2](MRP)系统、高级计划和调度(APS)系统、企业资源计划^[3](ERP)系统等专业子系统虽然能各自实现功能,但是造成了不同系统之间的数据割裂问

题。在内部网络环境下,企业可以通过放弃一定程度的隐私权缓解数据割裂问题,但在云平台上,企业的数据直接暴露给第三方,存在数据泄露和安全风险。如何在保护数据隐私的同时实现系统间数据的流通和整合,已成为当下工业界必须认真考虑的重要问题^[4]。针对上述背景,本文提出了一种基于 Bubble-Skyline 算法的解决方案,旨在解决工业互联网中的数据割裂问题,实现精确的数据分析,并以智能制造的排产优化为例,验证该方案的可行性和实用性。

1 数据割裂问题与隐私保护(Data fragmentation and privacy protection)

1.1 数据割裂问题

企业的 MRP 系统、APS 系统、ERP 系统等涵盖了生产计划、物料管理、供应链协调等关键功能,在工业互联网中发挥了至关重要的作用。MRP 系统主要关注物料需求计划,包括原材料、零部件的采购计划和库存管理;APS 系统负责高级计划和调度,用于优化生产计划和资源分配;ERP 系统是一个综合性的企业资源管理系统,涵盖了财务、人力资源、采购、销售等多个方面的数据。系统各自的独立性运行,导致不同系统中的数据无法直接交换和共享,造成信息的孤立和分散,加剧了数据割裂问题。

1.2 隐私保护

随着数字化和互联网技术的普及,数据隐私保护变得至关重要,特别是涉及企业商业机密、生产工艺等方面的数据。由于 MRP 系统、APS 系统、ERP 系统等中的数据格式和结构存在差异,导致其数据的共享和整合非常困难。此外,由于涉及敏感信息,例如客户数据和产品设计数据等,因此保证数据的安全性和隐私性尤为重要。

为了保证数据的安全性和隐私性,数据收集模块使用微服务架构收集不同云平台中的数据,同时在传输数据阶段采用了协议加密和哈希映射转化表等技术,实现不同系统之间数据的安全共享和整合。采用协议加密可以确保数据在各系统间传输的安全性,防止出现未经授权的访问行为。同时,通过哈希映射转化表,可以将不同系统的数据进行转化和映射,使其具备一致的格式和结构,从而实现数据的整合和共享。

2 Bubble-Skyline 算法(Bubble-Skyline algorithm)

2.1 Skyline 算法

Skyline^[5]算法是一种被广泛应用于解决多目标查询问题的多目标优化技术,它在数据挖掘和数据库查询等领域得到了较好的应用,对实际场景中的多目标决策有显著影响。Skyline 算法的核心思想是根据多个目标函数对数据进行筛选,选择出那些在所有目标上优于其他数据的数据点,形成 Skyline 集合。这些数据点在多个目标上都具有优势,因此可以作为决策的候选解。

以如表 1 所示的生产任务执行时间和成本数据为例,某制造厂希望找到执行时间较短且成本较低的任务进行优先排产。图 1 展示了这些任务在二维空间中的分布情况。在执行时间的选择中,任务 T4 用时最短,因此它是最合适的选择;而在成本的选择中,任务 T6 具有较低的成本,因此它是最合适的选择。

表 1 生产任务执行时间与成本数据

Tab.1 Production task execution time and cost data

任务 ID	执行时间/h	成本/元
T1	8	200
T2	4	170
T3	6	180
T4	2	320
T5	7	220

续表

任务 ID	执行时间/h	成本/元
T6	3	100
T7	6	160
T8	5	140

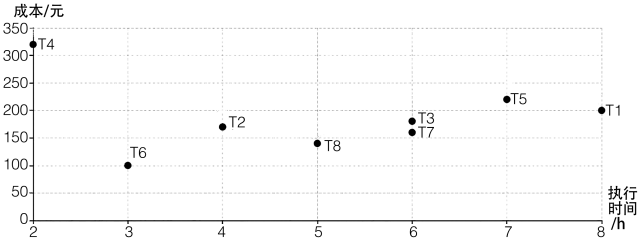


图 1 生产任务成本-执行时间

Fig. 1 Production task cost-execution time

2.2 Bubble-Skyline 算法

基于 Bubble 的 Skyline 算法是一种高效的多目标优化方法,它结合了分支定界技术与效用搜索^[6]的特点,能够在具有多个属性的数据集中找出不被其他数据点支配的点,即 Pareto 优势点。为了进一步优化 Skyline 集合的生成,Bubble-Skyline^[7]算法引入了冒泡排序和索引 ID 列表的思想。如表 2 所示,Bubble-Skyline 算法在执行 Skyline 计算时,使用的是数据索引 ID 列表,而不是真实数据,从而在保护数据隐私的前提下执行算法和比较数据。通过冒泡操作,较好的数据点逐步上移,排除次优的数据点,从而使最终的 Skyline 集合更具优化效果。使用基于 Bubble 的 Skyline 算法,可以在保护数据隐私的同时,找到多目标优化问题中的最优解。该算法的应用可以帮助决策者在面对多个目标和约束时,做出更准确、更全面的决策,同时该算法也为数据挖掘和决策支持系统等领域提供了一种高效的解决方案。

表 2 生产时间 ID 列表

Tab.2 Productive time ID list

任务 ID	执行时间排序 ID
T4	1
T6	2
T2	3
T8	4
T3	5
T7	5
T5	7
T1	8

3 设计方案(Design scheme)

本方案包括数据收集模块、数据预处理模块和数据分析建模模块,方案架构如图 2 所示。

3.1 数据收集模块

数据收集模块的主要职责是从多个源头,包括不同的云平台读取 MRP 系统、APS 系统、ERP 系统等,采集与生产排产相关的数据,包括原材料采购、生产计划、物流配送等关键环节产生的数据。但是,这些数据包含商业机密和生产设计等敏感内容,因此在数据收集过程中必须重点考虑数据隐私保护的问

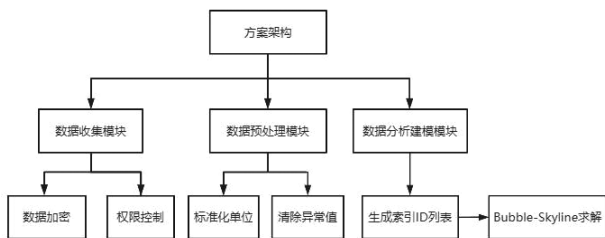


图2 方案架构

Fig. 2 Solution architecture

题。本文采用传输层的安全性协议,如安全套接层(SSL)和传输层安全性(TLS)对数据进行加密^[8],防止数据在传输过程中被截取、窃听或篡改。安全协议使用加密算法和证书确保数据在传输过程中的机密性和完整性,只有被授权的接收方能够解密和访问数据内容,确保了数据的安全性。此外,为了确保数据的合规性,数据收集模块采用权限校验机制。每次发起数据访问请求时,系统都会进行严格的权限检查,确保只有被授权的人员才能访问特定数据。

3.2 数据预处理模块

数据预处理模块负责对从各系统收集到的数据进行净化、转化和规范化,方便后续的数据分析和建模步骤。数据预处理主要聚焦于两个关键任务:异常值清除和单位标准化。

异常值清除:异常值通常指那些明显与其他观测值有差异或超出正常数据范围的数据点。在生产排产过程中,可能由于机器故障、备料不及时、数据采集误差或其他未知因素而产生异常数据。异常值需要被识别并清除,以防止它们对后续分析和建模产生负面影响。

单位标准化:在工业互联网中,不同的设备和传感器可能使用不同的单位和量纲表示数据,导致数据的不一致性。因此需要将数据转换为相同的单位和量纲,以便于后续的数据分析和建模。本文建立一个单位转换的映射表标准化单位(图3),该表包含了不同单位之间的转换关系,并且易于更新和维护。

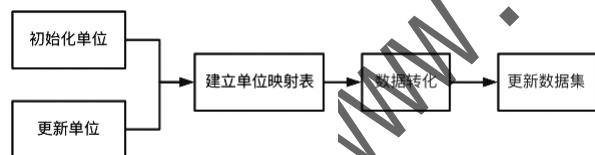


图3 标准化单位

Fig. 3 Standardized units

3.3 数据分析建模模块

数据分析建模模块旨在利用收集到的数据进行深入分析,并构建合适的模型解决排产问题。通过对数据进行统计分析和可视化展示,揭示数据中存在的模式、趋势和关联性。同时,模型可以通过调整生产参数和资源分配实现最佳的排产,制订合理的生产计划和调整资源分配,以应对生产过程中的变化。在工业互联网中,系统的独立性导致不同系统之间的数据无法直接交换和共享,造成信息的孤立和分散。Bubble-Skyline算法是一种数据分析方法,它能从大量数据中筛选出最优的数据子集,同时满足数据保密性的要求。

基于 Bubble 的 Skyline 算法,能在保护数据隐私的同时,为生产部门优化排产方案。该算法的基本步骤如图4所示。

步骤1:初始化。首先初始化一个空的 Skyline 集合(S),

用于存储最终的 Skyline 数据点。

步骤2:建立搜索空间。使用多维数据的属性建立搜索空间,其中每个维度对应一个属性。为每个搜索空间节点分配一个 ID 列表,用于存储与该节点相关的数据点 ID。

步骤3:分支定界搜索。利用 Bubble 算法的分支定界搜索技术,从搜索空间的根节点开始搜索。在搜索过程中,每次访问一个节点时,根据该节点的 ID 列表与当前 Skyline 集合中的所有数据点进行比较。若该节点不被 Skyline 集合中的任何点支配,则继续搜索该节点的子节点;否则,该节点及其所有子节点都将被剪枝,并从搜索空间中被删除。

步骤4:更新 Skyline 集合。当找到一个不被其他节点支配的数据点时,将其添加到 Skyline 集合中。同时,从 Skyline 集合中移除所有被新数据点支配的点。

步骤5:终止条件。当搜索空间中的所有节点都被访问或剪枝时,搜索过程结束。此时, Skyline 集合 S 包含了所有不被其他数据点支配的点,即最终的 Skyline 结果。

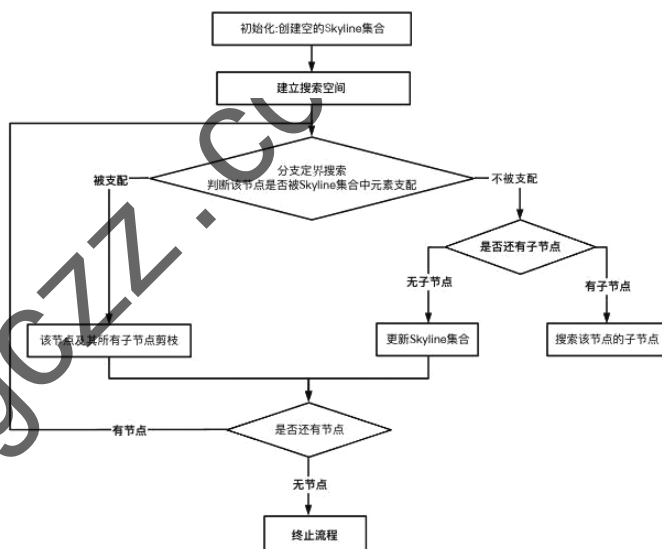


图4 Bubble-Skyline 算法的基本步骤

Fig. 4 Basic steps of Bubble-Skyline algorithm

Bubble-Skyline 的部分代码如下。

```
def dominates(row1, row2):
    return all(r1 <= r2 for r1, r2 in zip(row1, row2)) and any
(r1 < r2 for r1, r2 in zip(row1, row2))

def bubble_skyline(df):
    sky = []
    for _, row1 in df.iterrows():
        dominated = False
        to_remove = []
        for sky_row in sky:
            if dominates(row1, sky_row):
                to_remove.append(sky_row)
            elif dominates(sky_row, row1):
                dominated = True
                break
        if not dominated:
            sky.append(row1)
    sky = [r for r in sky if r not in to_remove]
    return pd.DataFrame(sky, columns=df.columns)
```

4 方案验证(Scheme validation)

本文建立订单信息 ID 表(表 3),用来保存订单在不同系统中产生的信息,包括订购的产品信息、交付日期等,构建 ID 列表,基于 Bubble 的 Skyline 算法计算排产方案。通过与实际使用的历史排产计划^[9]进行对比,评估新的排产^[10]方案在交货期限、最大产量等方面的表现。

表3 订单信息 ID 表

Tab.3 Order information ID table

字段名	功能描述
OrderFormModel	型号
OrderFormSpec	规格
OrderFormNum	订单数量
OrderFormRealNum	实际投产数量
OrderFormCompleteNum	完成数量
OrderFormCreatTime	生效时间
OrderFormTerm	交货时间
OrderFromBearingID	轴承型号 ID
OrderFormMenoy	订单利润
CustomerID	客户 ID
OrderFormID	订单 ID
OrderFormStatus	状态

排产流程图如图 5 所示。

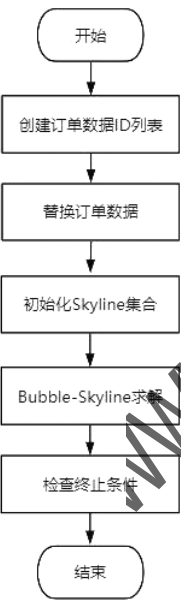


图 5 排产流程图

Fig. 5 Production scheduling flowchart

利用基于 Bubble 的 Skyline 算法对历史订单数据进行分 析,生成新的排产计划数据,将这个新的排产计划与实际使用的历史排产计划进行对比分析。图 6 展示了通过基于 Bubble 的 Skyline 算法生成的排产计划(虚线)与实际使用的历史排产计划(实线)中不同订单的生产时间对比。从图中可以观察到,新生成的排产计划在时间点上低于历史排产计划,生产时间明显减少。通过原排产计划所需总天数与新排产计划所需总天数的差值,再除以原排产计划的总天数,计算出平均生产时间减少了 17.15%,证明了通过采用基于 Bubble 的 Skyline 算法,在保护数据隐私的同时能更好地优化排产方案,减少了生产时间,提高了生产效率。

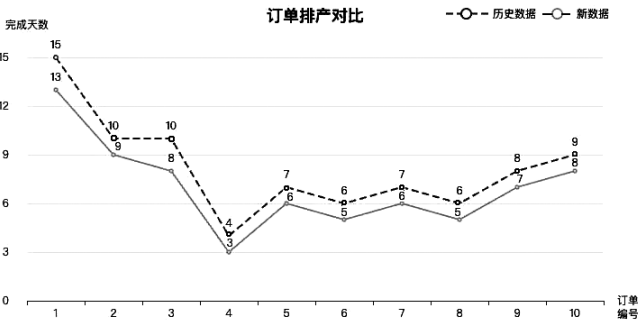


图 6 订单日期图

Fig. 6 Order date chart

5 结论(Conclusion)

本文分析了在工业互联网中,由于不同系统之间的数据无法直接交换和共享导致的数据孤立和分散问题。针对智能云平台的数据隐私保护和数据割裂问题,设计基于 Bubble-Skyline 算法的方案,该方案使用协议加密和哈希映射转化表实现了不同系统之间数据的安全共享与整合。使用 Bubble-Skyline 算法,可以从大量的数据中筛选出最优的数据子集。采用历史订单数据对方案进行验证,并与实际使用的历史排产计划进行对比分析,结果表明该方案在保护数据隐私的同时能更好地优化排产方案,平均生产时间减少了 17.15%,证明了该方案具有可行性和实用性。

参考文献(References)

[1] 王伟,苏耀东. 智能工厂工业控制系统安全体系构建和思考[J]. 石油化工自动化,2021,57(3):1-5.

[2] 王培露,曾强. 开环、广义物料需求计划算法研究及系统设计[J]. 工业控制计算机,2022,35(2):129-132.

[3] 仇晓筠. ERP 主生产计划在钢铁热轧生产的应用研究[D]. 上海:上海交通大学,2011.

[4] 刘晓曼. 浅谈工业互联网数据安全现状与形势[J]. 保密科学技术,2021(9):10-14.

[5] XIAO X X, LI J Z. Sampling-based approximate skyline calculation on big data[J]. Discrete mathematics, algorithms and applications, 2022, 14(7): 32-46.

[6] 申培萍,吴殿晓,王亚飞. 全局求解线性多乘积规划的分支定界算法[J]. 应用数学,2023,36(2):290-294.

[7] LIU R, LI D. Dynamic dimension indexing for efficient skyline maintenance on data streams[C]//NAH Y, CUI B, LEE S W, et al. Database Systems for Advanced Applications: 25th International Conference, DASFAA 2020. Cham: Springer International Publishing, 2020: 272-287.

[8] 叶会成,刘光杰. 基于身份公钥体制的新型 SSL 协议设计与实现[J]. 计算机与现代化,2017(9):50-55.

[9] 石乔梓. Y 日用瓷企业排产优化研究[J]. 河南科技,2022, 41(11):147-151.

[10] 杨丰赫. 基于云边协作的工业互联网智能排产方法研究与实现[D]. 南京:东南大学,2021.

作者简介:

周锴炼(1999-),男,硕士生。研究领域:数据隐私保护及排产优化。

鲍 敏(1977-),男,博士,副教授。研究领域:智能制造,生产过程数据分析。