

基于关联规则算法的检查器组优化建议机制

李 昕, 谢颖华

(东华大学 信息科学与技术学院, 上海 201620)

摘 要: 工业生产过程中, 图纸设计是最关键的一步, 如何实现效率与质量并存的设计流程是生产制造业中的热点问题。介绍了一种基于关联规则挖掘算法的检查器组优化建议机制, 该机制将关联规则挖掘算法用于检查器组的优化, 以实现在保证检查器组覆盖率的前提下, 缩短检查用时, 提高工作效率。实验表明, 该优化系统中优化后的检查器组覆盖率满足实际工程的需求, 能有效地缩减检查器组的规模, 从而节约检查时间, 减少系统资源占用, 提高用户的工作效率。

关键词: 检查器组优化建议; 关联规则挖掘; 模型缺陷分析

中图分类号: TP39

文献标识码: A

DOI: 10.19358/j.issn.2096-5133.2020.07.012

引用格式: 李昕, 谢颖华. 基于关联规则算法的检查器组优化建议机制[J]. 信息技术与网络安全, 2020, 39(7): 67-71.

The optimization suggestion mechanism of checkers based on association rule algorithm

Li Xin, Xie Yinghua

(School of Information Science and Technology, Donghua University, Shanghai 201620, China)

Abstract: Drawing design is the most essential part in industrial production. How to realize a design process with great efficiency and quality is a hotspot issues in manufacturing. This paper introduces an optimization suggestion mechanism of checkers based on association rule mining algorithm, which can reduce the time of check-mate, enhance work efficiency under the premise of ensuring the coverage of the checkers by applying the association rule mining algorithm into the optimization of checkers. Experimental results show that the coverage rate of the optimized checkers in the optimized system can meet the requirements of the actual engineering, which can effectively reduce the scale of the checkers, so as to reduce the time of check-mate, reduce the system resource occupation, and improve the working efficiency of users.

Key words: checkers optimization recommendations; association rule mining; model defect analysis

0 引言

工业产品生产过程离不开图纸设计, 一个功能完善、细节满分的设计平台可以为图纸设计的过程提供极大的便利, 甚至可以简化或加快产品开发过程。Siemens PLM Software 公司出品的 NX 就是一个交互式 CAD/CAM (Computer Aided Design and Computer Aided Manufacturing) 系统, 它不止提供模型设计平台, 还提供功能强大全面的 Check-Mate 模块用于产品模型的质量检查^[1]。检查结果将通过 QDB (Quality Dashboard) 得以显示, 这一套功能形成一条闭环, 为产品设计流程提供极大的便利。

第二代 QDB 更名为 Cloud QDB, 摆脱了第一代

单机版的设计, 实现了在线网页版的转变, 引用云技术实现设计者范围之内的数据共享和设计监控, 不仅满足了自定义报告展示的需求, 还提供了更加美观的用户交互界面, 更直观展示 Check-Mate 检查报告。NX 的用户较多从事汽车行业设计, 设计的模型文件十分复杂, 所以 Check-Mate 的运行时间也随之变长, 同时在运行 Check-Mate 的时候系统资源被占用, 用户工作效率不高。

针对 Check-Mate 用时过长这一问题, 本文设计引入检查器组优化建议机制, 通过关联规则挖掘算法, 挖掘存在关联的检查器, 删除冗余检查器, 在满足有效检查器组覆盖率的基础上为用户提供检查

器组优化建议,减少检查时间,提高用户的工作效率,节约生产制造中的成本。

1 系统总体架构及功能模块

1.1 总体架构

检查器组优化建议系统总体架构由数据库端、数据提取端、数据分析单元、报告及优化建议输出端四部分构成。检查器组优化建议系统可以为用户提供检查结果报告展示;将报告保存在云端,实现用户群体之间的数据共享;并提供检查结果分析,根据关联规则挖掘算法探索检查器之间的关联,为用户提供可以节省检查时间的检查器组优化建议。总体架构如图 1 所示。

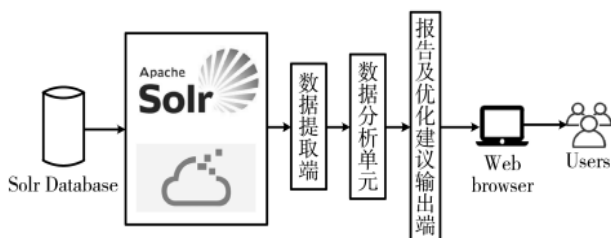


图 1 检查器组优化建议系统总体架构

数据库端使用的 Solr 负责保存以及管理数据引擎,具有在线保存并提供搜索 XML 格式的检查结果文件的功能;数据提取端提供解析服务,通过解析检查结果文件,提取检查器的关键信息,包括检查器状态、运行用时、历史状态以及检查器与模型文件的属性特征信息;数据分析单元提供关联规则挖掘功能,绑定关联规则挖掘算法,挖掘检查器组中的关联规则,为优化建议提供理论依据;报告及优化建议输出端可以整合数据分析单元输出的规则,与前端的部件组合生成可视化报告,提供检查器组优化建议,在保证检查器组覆盖率的同时合理建议优化检查器组;优化建议和报告将通过浏览器在前端展示,提供可视报告,为用户的工作提供便利。

1.2 系统功能模块

检查器组优化建议系统包含 5 个模块,分别是检查结果信息获取模块、数据预处理模块、特征分析模块、优化建议过滤模块以及报告输出模块。

检查结果信息获取模块主要是用于提取检查结果中的检查器以及模型文件的关键信息,包括检查器在该模型文件的状态、运行用时、历史状态、检查器本身的设计者及功能;数据预处理模块可以综

合检查器和模型两部分信息,根据二者关联部分进行检查器组初次过滤;特征分析模块的功能是分析模型文件特征,决定选取哪些检查器的关键信息来作为关联规则挖掘所需的数据;优化建议过滤模块绑定了关联规则算法的运算,将特征分析模块选取的数据加以运算生成规则后,对冗余检查器进行筛选罗列;报告输出模块整合数据,根据用户自定义格式的模板生成整体报告。图 2 所示为系统功能模块图。

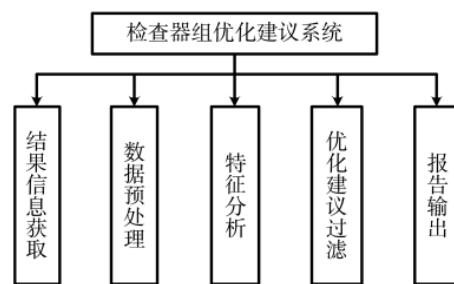


图 2 系统功能模块

2 关联规则挖掘的应用

2.1 关联规则挖掘

数据挖掘技术主要是指从大量的、模糊、随机等数据流中,通过技术手段和智能分析获取隐藏在庞杂数据中的有用信息^[2],本质是数据库内的知识发现。这些信息反映了在一定规模的数据集中各元素之间的有效联系,或许它们不能在所有领域中通用,但一定可以效力于某一特定领域。

目前运用最普遍的数据挖掘技术是关联规则,关联规则是指从指定数据集中迅速、有效地找到各个事务项目之间存在的、隐藏的、有价值的关联关系,从而为用户提供有力的决策支持^[3]。最初它由 Agrawal 等专家学者于 20 世纪 90 年代初提出,目的是用于发现顾客的商品消费模式,经过了十几年的发展,已日益成熟和完善^[4]。

在实现关联规则挖掘的过程中,为了最大限度找到有用的规则,需要使用某些度量来约束,其中最常用的是支持度(Support)和置信度(Confidence)。前者表示数据库中项集出现的频率,而后者则用来度量规则的可信程度,它们分别表示发现规则的有用性和确定性^[5]。

此外,也经常会使用提升度(Lift)来度量相互依赖程度。

提升度有三种情况,分别是等于 1、大于 1 和

小于 1。以 1 为分界线,其依赖程度随提升度值呈正相关;当等于 1 时,则表明前提和结论对应的事件相互独立。

至今关联规则挖掘已经衍生出多种算法,包括 Apriori、FP-growth 等算法。传统关联规则挖掘包含两部分:生成频繁项集以及生成关联规则。生成有效规则的前提是存在可靠的频繁项集,所以选取合适的数据来生成频繁项集是算法的关键。

2.2 算法设计

进行检查器组优化分析的目的是缩短 Check-Mate 运行用时,而运行用时与检查器组规模息息相关,故需要在保证 Check-Mate 可靠运行的同时尽可能缩减检查器组规模。

关联规则算法中的 Apriori 算法是一种基于频繁项集理论的递归方法^[6],通过生成候选项集和情节的向下封闭检测两部分来产生频繁项集,从数据库中挖掘支持度大于最小支持度阈值以及置信度大于最小置信度阈值^[7]的关联规则。另一种名为 FP-growth 的关联规则算法的原理是深度优先搜索,相比于 Apriori 算法,FP-growth 算法不会产生候选项集,仅仅遍历数据集两次,就能够完成频繁项集的发现。故 FP-growth 算法比 Apriori 算法效率高^[8]。

已知 Apriori 算法需要多次扫描数据库,每次利用候选项集产生频繁项集;而 FP-growth 算法则利用树形结构,无需产生候选项集而是直接得到频繁项集,大大减少扫描数据库的次数,从而提高了算法的效率。另一方面 Apriori 算法在扩展性方面又优于 FP-growth 算法,两种算法各有千秋,故本文为了实现检查器组针对模型的覆盖率与运行速度的平衡,选择 FP-growth 算法和 Apriori 算法组合来进行规则挖掘。

2.2.1 数据预处理

数据来源于每次运行 Check-Mate 之后生成的检查结果文件,将它们按批次输入,以批次数识别。将文件通过 Cloud QDB 中的解析功能从 XML 格式解析转化成易于阅读的 JSON 格式,重点提取当前状态为 fail 的检查器名称,对应检查用时、检查器应用、特色功能、设计特征以及继承关系,甚至包括设计者信息、设计时间等边缘信息。将这些信息制成数据集的标签,整合成为数据集;之后进行数据预处理,综合模型文件结构和检查器功能两方信

息,对检查器组的组建进行分析,判断检查器功能是否与模型文件结构要求吻合,对不符合要求的检查器进行初次过滤;接下来进行特征分析,分析模型文件装配层次,根据层次分类检查器组信息,分层运算也可以帮助并行运算,从而缩短运行用时;接下来进入关联规则挖掘步骤,第一步是生成频繁项集,参考 Apriori 和 FP-growth 两种算法的特点,这一步选择效率较高的 FP-growth 算法。

2.2.2 频繁项集挖掘

FP-growth 算法首先根据批次以及分层对数据编号,根据批次及分层提取数据集中状态为 fail 的检查器组作为相应事务组,整合成清单。接下来对每个事务进行计数,设定最小支持度,按计数结果降序排列每一个事务,将不符合支持度的事务删除并重新整理清单。然后到了关键的一步:构建 FP 树。创建一个虚拟的根节点,加入第一个事务组,即将根节点指向事务组中第一个事务,由于此时已经过降序排列,排序靠前的事务也相应具有较高的权重。第一个事务将指向排序第二的事务,由此类推,将事务组内所有事务按序完成指向。根据规则加入后续的事务组,如果出现相同的节点则进行累加,最终得到 FP 树。挖掘频繁项集需要获取每一个事务的条件模式基,它是集合形式,包含以所查找事务为结尾的路径。根据条件模式基构造条件 FP 树,和被分析的事务并集得到支持度大于 2 的所有频繁项集。

2.2.3 挖掘规则

频繁项集生成后,需要使用 Apriori 算法进行规则挖掘,将频繁项集根据项数进行降序排列,发现最大频繁项集。Apriori 算法有一条定理,即频繁项集的子集也是频繁项集。所以最大频繁项集中一定包含全部的频繁项集。通过上一步 FP-growth 算法已经得到了支持度计数,计算置信度和提升度就可以立即提取规则。具体流程如图 3 所示。

3 实验分析

基于关联规则算法的检查器组优化建议分析需要从大量的验证结果文件中获取检查器的特征信息。常规模型文件与检查器组如表 1 所示,其中 Part 代表模型文件,其编号表示一次工程中所涉及到的模型文件,文件大小从 KB 到 GB 不等,据模型复杂度而定;Checkers 表示检查器组,Checkers(n)表示组内检查器个数,模型越复杂,所需检查器越多,同时运行 Check-Mate 的时间也越长。

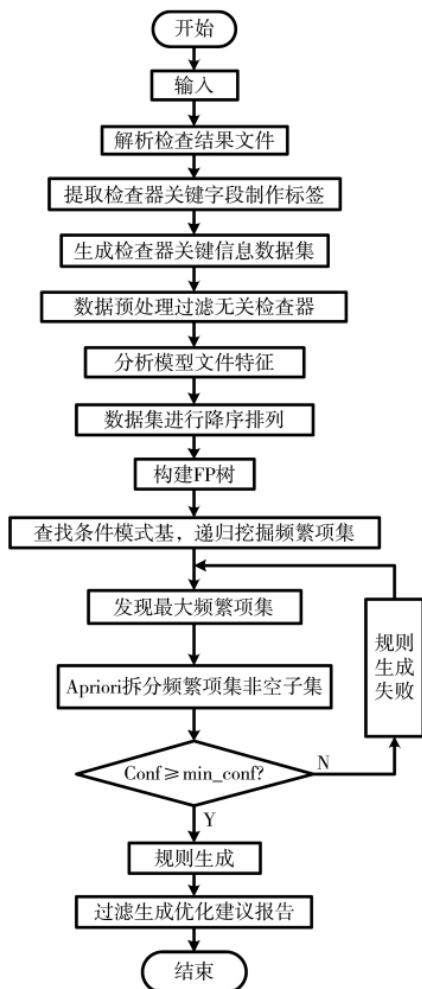


图3 算法流程

本文采用的共计 13 599 份验证结果文件取自

表 1 模型与检查器组

Part	Checkers	Checker (n)
P00001	inappropriatecu, multisolid, narrow_f, fragmenteded, narrow ...	406
P00002	high_dege, tiny_d, embedded_s, sharpangle, volume_solid, o ...	506
P00003	intersect_loop, check_pref, check_sheet_an ...	357
P00004	ValiForExport, facepatch, check_drafting, ValidateFailedExs ...	432
...	...	...
P13599	check_percent_fill, ValidatePleConnections, spell_changle ...	388

Siemens PLM Software 公司内部测试文件。数据集的 20% 作为训练集, 80% 用于测试。

本文实验环境: 设备型号: HP Z240 Tower Work-station; CPU: Intel® Core™ i5-6400 CPU@2.70 GHz; Windows 版本: Windows 7 旗舰版。

3.1 运行时间对比

设置对照试验, 首先记录 Check-Mate 在无优化建议状态下的运行时间 (bf), 获取同一检查结果文件数据库在经过检查器组优化系统后得出的规则, 用户根据规则推算出的优化建议来删减冗余检查器, 而后再对同一模型文件运行 Check-Mate 并记录时间 (af), 观察两次运行时间的对比。

优化前后 Check-Mate 占时对比如图 4 所示, 可以看出在 10 次实验中, 优化后 (af) 对比优化前 (bf) 的 Check-Mate 用时有了显著的缩短。

3.2 覆盖率

Num1 为 Check-Mate 在无优化建议状态下结果为 fail 的检查器个数, 接下来通过检查器组优化建议系统, 删除冗余检查器, 用此时被优化后的检查

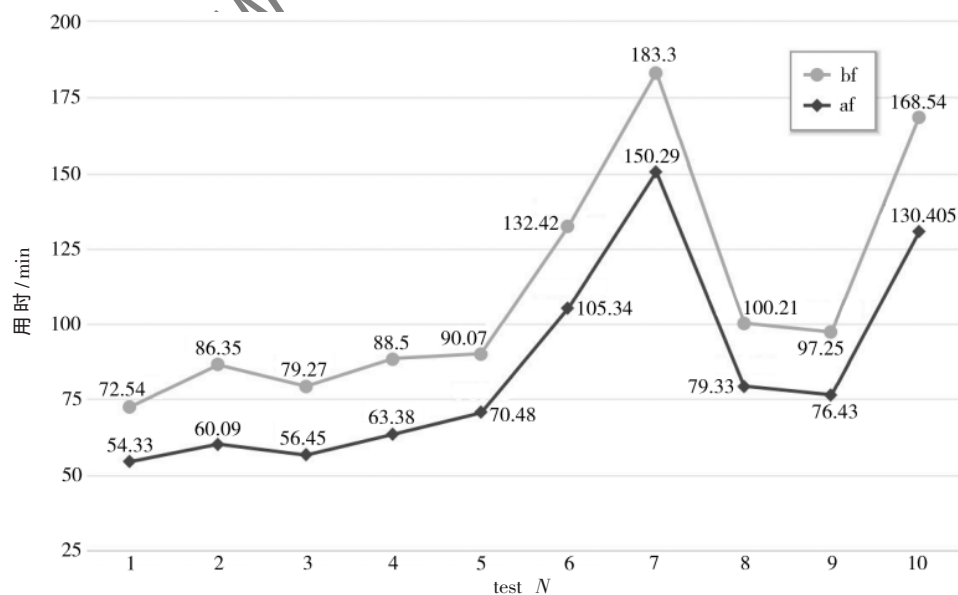


图4 优化前后占用时间对比

器组运行 Check-Mate, 参考检查结果, 修复模型文件中存在的设计问题。下一步使用被删除的检查器再次对该模型文件运行 Check-Mate, 记录此时状态仍为 fail 的检查器个数为 Num2, 即被优化建议误删除的检查器个数。

覆盖率由式(1)确定:

$$C_v = \frac{\text{Num1} - \text{Num2}}{\text{Num1}} \times 100\% \quad (1)$$

完整的检查器组由于不筛选检查器, 所以检查器组对于模型文件的覆盖率可以达到 100%, 其代价就是用时过长, 所以如果缩短用时, 势必要牺牲检查器组的覆盖率。考虑到工业设计中对于模型文件的质量需求, 覆盖率应考虑在 95% 以上, 才能认为是有效的优化系统。

如图 5 所示, 检查器组优化建议系统可以在保证覆盖率在 95% 的基础上对用时尽可能缩短, 使覆盖率和用时处于良性平衡。该优化系统覆盖率满足实际工程的需求, 由此系统可靠性得到了保证, 证明检查器组优化建议系统值得推广。

4 结论

本文提出的基于关联规则挖掘算法的检查器组优化建议机制, 在保证覆盖率的情况下, 可以为用户提供优化建议, 用户可根据需要删减检查器组, 缩短 Check-Mate 的运行时间, 提高工作效率。

参考文献

[1] 苏媛媛, 毛宏图. UG NX 平台模型质量检查应用与

扩展技术研究[J]. 航空发动机, 2015, 41(1): 62-65.

[2] 张鸿雁. 基于 Apriori 算法的校园教学质量评价系统设计[J]. 电子技术与软件工程, 2019(18): 188-189.

[3] 周雯. 大数据环境下关联规则算法的研究[D]. 西安: 西安邮电大学, 2016.

[4] 林明方. 大数据环境下关联规则挖掘研究[J]. 自动化与仪器仪表, 2017(6): 36-37, 40.

[5] 马勇. 基于 Apriori 算法的 Web 日志分析方法[C]. 中国计算机学会计算机安全专业委员会: 第 27 次全国计算机安全学术交流会论文集, 2012: 183-185.

[6] 曾子贤, 巩青歌, 张俊. 改进的关联规则挖掘算法——MIFP-Apriori 算法[J]. 科学技术与工程, 2019, 19(16): 216-220.

[7] 陈海军. 基于 Apriori 数据挖掘算法的信息推荐图书管理系统设计[J]. 现代电子技术, 2019, 42(23): 115-119, 124.

[8] 杨彩, 喻铁朔, 石月凤, 等. 基于 FP-growth 算法的课程关联性分析[J]. 中国教育信息化, 2019(17): 35-38.

(收稿日期: 2020-03-05)

作者简介:

李昕(1996-), 女, 硕士研究生, 主要研究方向: 大数据、关联规则挖掘。

谢颖华(1972-), 女, 硕士, 副教授, 主要研究方向: 大数据、数据挖掘等。

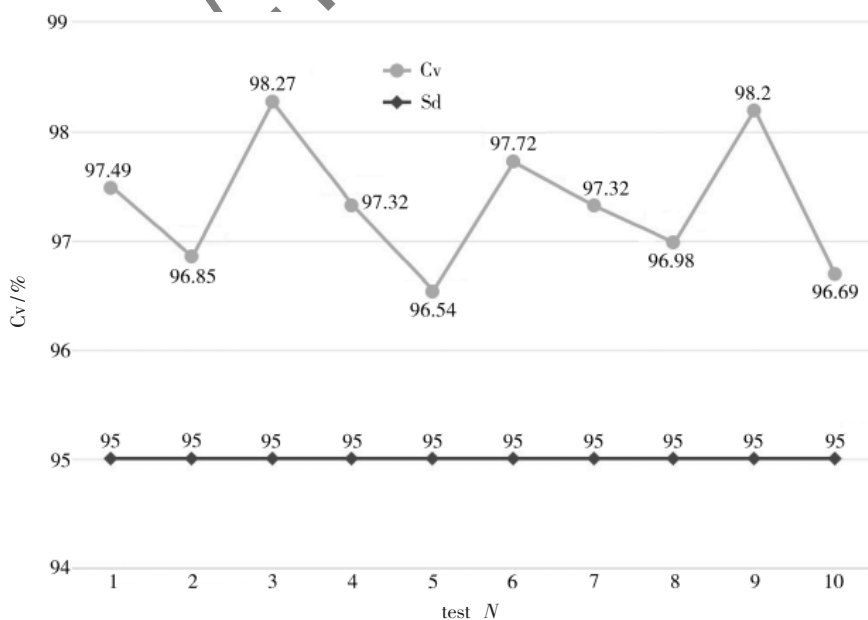


图 5 优化后检查覆盖率(Cv)

版权声明

经作者授权，本论文版权和信息网络传播权归属于《信息技术与网络安全》杂志，凡未经本刊书面同意任何机构、组织和个人不得擅自复印、汇编、翻译和进行信息网络传播。未经本刊书面同意，禁止一切互联网论文资源平台非法上传、收录本论文。

截至目前，本论文已经授权被中国期刊全文数据库（CNKI）、万方数据知识服务平台、中文科技期刊数据库（维普网）、JST 日本科技技术振兴机构数据库等数据库全文收录。

对于违反上述禁止行为并违法使用本论文的机构、组织和个人，本刊将采取一切必要法律行动来维护正当权益。

特此声明！

《信息技术与网络安全》编辑部
中国电子信息产业集团有限公司第六研究所