

基于产品生命周期的质量数据分析与数据挖掘

高建荣^{1,2}, 高琦^{1,2}, 高菲^{1,2}

¹ 山东大学 高效清洁机械制造教育部重点实验室, 山东 济南 250061

² 山东大学 机械工程学院, 山东 济南 250101

摘要: 针对传统质量管理理念约束下, 离散制造企业质量管理体系条块分割, 管理活动分散且对大规模质量数据分析匮乏的问题, 文章从产品生命周期的角度对质量数据进行了分析整理, 在进行数据挖掘之前对数据的获取、数据的预处理以及数据的转化进行了讨论, 构建了 ETL 数据预处理模型。然后对具体的数据挖掘应用方法进行了研究, 提出了产品生命周期质量数据挖掘整体体系结构, 并对结构中包含的数据层、方法层、知识层、应用层进行了分析阐述。

Quality data analysis and data mining based on product life cycle

Gao Jianrong^{1,2}, Gao Qi^{1,2}, Gao Fei^{1,2}

(1 Shandong University, Key Laboratory of High Efficiency and Clean Mechanical Manufacture Shandong University, Ministry of Education, Jinan 250061, China

2 Shandong University, School of Mechanical Engineering, Jinan 250061, China)

Abstract: To solve the problem of the quality management system of the discrete manufacturing enterprise is segmented, decentralized and lacking of large-scale quality data analysis under the traditional concept, this paper analyzes the quality data from the perspective of the whole life cycle, it have discussed the data acquisition, data preprocessing and data conversion before data mining, and have constructed the ETL data preprocessing model. Then, studied the concrete data mining application method, put forward the overall data mining architecture of product life cycle quality data, and analyzed the data layer, method layer, knowledge layer and application layer contained in the architecture.

Keywords: product life-cycle management; quality data; data preparation; data mining architecture

收稿日期:

作者简介: 高建荣 (1991—), 女, 陕西延安人, 山东大学机械工程学院在读硕士, 研究方向: 数据挖掘、知识管理, leading soul@163.com; 高琦 (1970—), 女, 山东东营人, 山东大学机械工程学院教授, 博士, 研究领域为产品数据管理 PDM/ 产品生命周期管理 PLM、知识工程、智能设计, gaoqi@sdu.edu.cn。

基金项目: “十三五”装备预研领域基金项目 (61409230102)

1 前言

制造业是国民经济的重要组成部分，制造业的技术水平不仅决定了其他产业竞争力强弱，而且关系到整个国民经济运行的质量和效益。离散制造业是现代制造业发展的重要方向，离散制造业不仅要满足客户个性化定制、品种多、批量少、交货周期短的需要，还要满足高质量产品的要求。离散制造行业竞争非常激烈，主要表现在产品质量和服务上，提高产品质量和服务水平已成为离散制造业提高竞争力的重要手段和途径^[1]，为了适应行业新形势的发展要求，借助先进的质量管理方式来提高企业的科学质量管理水平，增强企业核心竞争力是非常必要的。

随着现代信息技术的迅速发展，多数离散制造企业已实现了信息化管理，日常生产活动中产生的数据爆炸式增长，形成了覆盖产品全过程的数据海洋^[2]，但大多数企业对数据的分析还停留在简单的统计分析、日常数据汇总、趋势分析等浅层的处理，虽然也有少部分企业有对数据进行进一步挖掘的尝试，但基本都停留在表面和局部问题的分析，没有一个系统的方法和技术理论支持。因此，如何有效地管理和利用这些从企业生产、经营中产生的庞杂的质量数据成为企业长远发展亟待解决的问题。近年来，各个行业逐渐认识到大数据能够给企业带来的巨大价值，通过数据挖掘技术来提高运作效率、提升精细化和智能化水平，提高质量管理水平，可以有效地推动中国制造走向中国创造。数据挖掘以其强大的数据分析能力从大量的、杂乱的数据中提取有价值的信息，它能开采出数据的潜在模式，从而辅助企业决策，增强企业竞争力^[3]。集诸多优点于一身的数据挖掘技术作为一种先进且强大的数据分析工具，为离散制造企业实现高效质量管理提供了全新的科学手段，因此将数据挖掘技术与制造业质量管理有机结合具有重大现实意义，也是未来智能制造发展的必然趋势。

文章以离散制造业为背景提出了基于产品生命周期的质量数据分析与挖掘方法。首先对产品质量数据源进行了分析，讨论了ETL技术在质量数据获取与整合过程中的应用，为后续进行数据挖掘做准备。然后在对典型数据挖掘方法和算法分析的基础上，提出了产品生命周期质量数据挖掘体系结构，并阐述了各层次基本原理，说明了将数据挖掘策略用于产品生命周期质量数据分析的重大现实意义。

2 产品生命周期质量数据源分析

2.1 生命周期质量数据源

在传统制造模式和质量理念的约束下，我国制造企业的质量管理通常局限于质量管理部门的独立活动，质量管理体系条块分割，产品生产的各部门，各阶段以及各环节的质量管理活动相对封闭且分散，单纯以满足产品设计规范为目标，难以实现产品生命周期全过程质量水平的控制，导致企业各类质量信息资源独立，最终形成一个个“信息孤岛”，大量的业务数据杂乱堆积，无法从系统角度出发对产品质量进行全方位的评价^[4]。目前对质量数据的获取多是通过手工整合来实现局部质量数据整理，还没有实现质量数据生命周期整体分析。在分析产品生命周期质量数据的过程中，对不同阶段、不同位置的质量数据的抽取、转换和装载是实现全面质量数据整理的关键。

影响产品质量的因素贯穿于产品的整个生命周期，产品的形成过程也就是产品质量的形成过程，也伴随着质量数据生成的过程，如图1所示，产品生命周期由需求分析、设计（包括概念设计、详细设计、工艺设计）、材料准备（零部件的采购供应）、生产装配（包括零件加工、部件装配、整机装配、整机测试等）、使用维修（包括仓储、销售、配送、安装调试、运行、检修、待工）和转化再生（包括产品报废、零部件再用、废件再生制造、原材料回收再利用、废料降解处理等）等阶段组成。产品质量数据根据其产生阶段不同可以划分为需求阶段、设计阶段、采购阶段、生产阶段和服务阶段五大部分，如图2所示。而这些数据又分别来自不同的应用软件系统。这些系统一般是由不同的软件厂商设计开发的，它们的数据库开发分属不同架构，即使同样架构的系统不同模块间也使用不同字段，不同系统架构、不同表结构，不同数据形态，为质量数据的综合利用带来挑战。要对产品生命周期各阶段的质量数据进行数据挖掘，就必须将其转换为可供分析调用

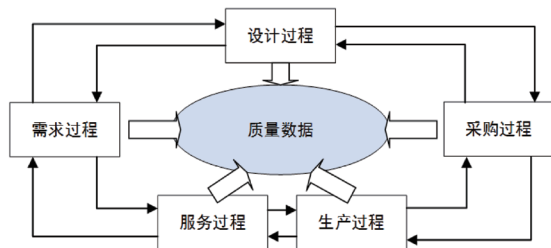


图1 产品生命周期质量数据形成过程

的统一形式的数据。

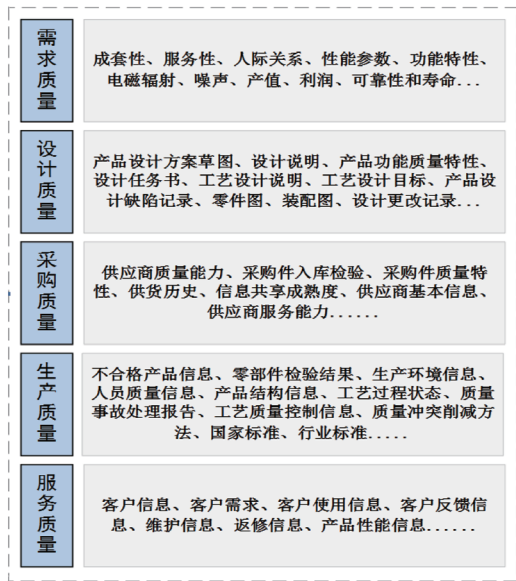


图 2 产品生命周期质量数据源

2.2 异构数据的获取与转化

从产品设计到制造的整个过程中涉及到的质量数据通常包括需求阶段、设计阶段、采购阶段、生产阶段和服务阶段等五大类质量数据。这些数据基本可以分为产品设计模型数据，Word/Excel 等文本、文档数据，以及数据库中的数据^[5]。在进行数据挖掘之前，需要对各类非结构化、半结构化的数据进行结构化处理，对结构化数据需要进一步讨论其获取存储方法。如图 4 对不同存储方式的数据，采取的结构化方法也不相同。

首先是对于与产品设计相关的 STEP 格式文档的结构化处理方法。STEP 文件以 B—rep 方式表达几何模型，分为数据段和头部段两部分组成，数据段包括几何实体的定义及其数据要素，头部段包括 STEP 转换器的版本和建立三维模型的相关信息。采用属性邻接图对 STEP 格式文档进行特征提取，并将提取到的特征数据以结构化的方式存储到数据库中。

其次是对于表格类文件（如 Word、Excel 和工艺卡片等）信息抽取并转换为结构化数据库数据的方法。表格类文件通常是依托 Word 软件创建一维表格，属于典型的非结构化数据类型。这种形式的数据可以采用 XML 的半结构化的数据格式作为过渡数据类型，以 Word 等文档为数据源，针对不同的类型和格式，构建不同的数据抽取策略，在数据输

出时，选择不同的 XML 模板，将表格类文件以指定的 XML 文档格式进行输出；在半结构化数据到结构化数据的过程中，通过解析 XML 文档，将数据转换为结构化数据并存储到数据库或知识库中。

再次是对于文本类文档的信息抽取及结构化处理方法。对文本文档的信息抽取过程包括对文档的合并、分词和标注，其中最主要的部分为标注，采用中国科学院计算技术研究所开发的汉语词法分析系统 ICTCLAS(Institute of Computing Technology, Chinese Lexical Analysis System) 对文档进行分词处理，再利用英国 Sheffield 大学开发的 GATE (General Architecture for Text Engineering, 文本工程通用框架) 为主体实现对文本的标注，最后将标注过的文本生成标注集，存储到数据库中。

最后是对数据库数据的信息抽取。与前面不同，数据库中数据已经是结构化的数据，我们需要做的只是如何将数据库中的数据抽取并加载到数据仓库中去，经过数据清洗、转换等预处理操作后为后续数据挖掘做准备。包括前面的模型数据，表格文档类数据及文本类数据最终都将是源数据库与数据仓库之间的关系。

3 质量数据挖掘的体系结构

图 3 为基于产品生命周期的质量数据挖掘整体体系结构，主要由数据层、方法层、知识层和应用层组成。数据层由可获得的与质量有关的各个数据源（如三维模型、文本、CAPP、MES、PDM 等信息化软件）组成的原始数据，以及经过结构化处理的为数据挖掘做好准备的结构化的质量数据^[6]。在此基础上建立数据挖掘方法层，基于可能挖掘出的知识类型提出了

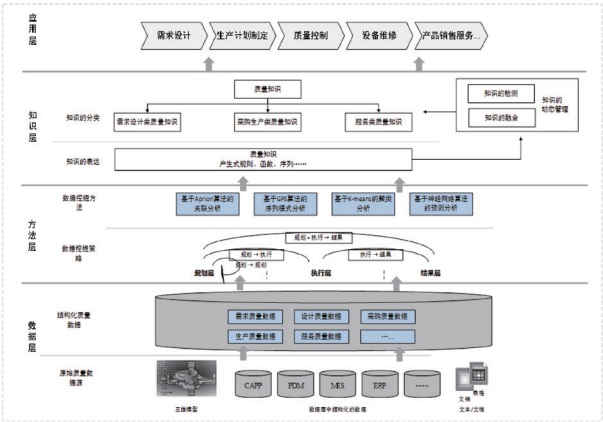


图 3 质量数据挖掘体系结构

对应的挖掘算法，并将这些算法进行封装，只提供相关的参数接口和数据接口。知识层对数据挖掘产生的所有知识进行分类和动态管理。应用层从需求分析到设计再到制造执行过程中充分运用数据挖掘得到的知识来支持上层业务决策。

3.1 数据层

设计阶段、采购阶段、生产制造阶段及服务阶段等五大类质量数据，这些数据来源于各种信息系统如 PDM、CAPP、MES、ERP 等。由于数据的来源不同、用途不同，因此具有不同的存储方式。一些数据可以存放在数据库中，为完全结构化数据；一些模型、图纸类的的数据，如产品的三维 CAD 模型通常以 STEP 标准实例文件 .stp/.step 的文件格式进行存储；还有一些数据是以文本、表格等办公文档的形式存储，如 Excel/Word 形式的工艺卡片、统计文件等。通常这些半结构化或者非结构化数据，无法直接对其进行数据挖掘，因此在进行数据挖掘之前，首先需要进行数据预处理操作使其满足数据挖掘的要求，这里就用到了数据仓库技术，如图 4 所示，在数据仓库搭建过程中，任务最重、工作最复杂的当属数据的抽取、转换、清洗和装载，即 ETL (Extract/Transform/Loading) 过程（数据从业务系统抽取、转换、装载到数据仓库，同时提供数据质量管理，完成数据处理与调度的过程），对于不同存储方式的数据，采取不同的数据抽取措施和不同的结构化处理方法，其主要功能包括：（1）数据的抽取：从不同的网络、不同的操作平台、不同的数据库及数据格式、不同的应用系统中抽取数据；（2）数据的转换：数据的转化（数据的合并、汇总、过滤等）、数据的重新格式化和计算、关键数据的重新构建；（3）数据的加载：跨网络、跨平台，将数据按照物理数据模型定义的表结构装入目标数据库。

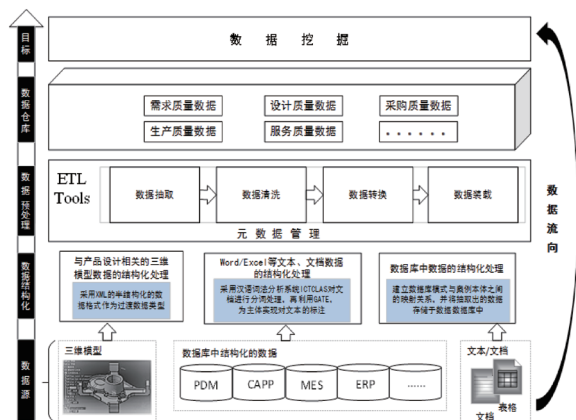


图 4 质量数据预处理模型

3.2 方法层

产品从需求分析、概念设计、结构设计、工艺设计、到生产制造出来，最后到维修报废利用的过程^[7]，就是经历了一个“规划→执行→结果”的过程，这期间会产生规划层数据（需求、设计质量数据）、执行层数据（生产制造数据）和结果层数据（服务维修报废反馈数据）。于是提出“规划—规划”、“规划→执行”、“执行→结果”、“规划+执行→结果”四种数据挖掘策略。

基于“规划—规划”的策略，挖掘常用的工序序列模式，挖掘结果为某种序列。基于“规划→执行”的策略，挖掘制造特征数据与功能需求之间的相互关系，挖掘加工方法和可靠性要求之间的相互关系，挖掘结果可能为某种规则。基于“执行→结果”的策略，挖掘加工方法对产品销售服务报废结果产生影响，挖掘资源能力对产品返修进度效率产生影响，挖掘结果可能为某种规则或函数。基于“规划+执行→结果”的策略，挖掘在某种性能要求、采用某种加工方法、选择某种资源能力的前提下，预测最终的产品维修概率，报废年限，挖掘结果可能为某种规则或函数^[8]。

分析上述几种数据挖掘的内容、结果的表现形式，本文采用关联分析、序列模式分析、聚类分析以及预测分析四种数据挖掘方法，当前在计算机和数学领域已经有较为完善的数据挖掘算法，而且这些算法已经在一些领域被证明为是非常有效的^[9]。因此，本文采用已经较为成熟的挖掘算法，即基于 Apriori 算法的关联规则挖掘、基于 GSP 算法（generalized sequential pattern）的序列模式分析、基于 k-means 算法的聚类分析、以及基于神经网络算法的预测分析。如图 5 所示为基于数据挖掘经典算法的数据分析模型，为了便于这些复杂的挖掘方法更方便的使用，这里对数据挖掘算法封装，只提供相关参数接口和数据接口，为后续的质量数据挖掘做准备。

① 基于 Apriori 算法的关联分析模型

如图 5 中①所示为基于 Apriori 算法的关联分析模型。将离散化的字符型或数值型数据输入到模型中，对最小支持度和最小置信度进行设置，得到具有一定支持度和置信度的关联规则。

② 基于 GSP 算法的序列模式分析模型

序列模式挖掘的大部分方法都采用了类 Apriori 算法的变种，GSP 算法是基于 Apriori 算法实现序列模式挖掘的一

种经典算法。其分析模式如图 5 中②所示。

③ 基于 K-means 算法的聚类分析模型

如图 5 中③所示为基于 K-means 算法的聚类分析模型。将离散化的字符型或数值型数据输入到模型中，对聚类数或相似性阈值进行设置，得到按照不同规则聚集的类及其相应的属性变量。

④ 基于神经网络算法的预测分析模型

如图 5 中④所示为基于神经网络算法的预测分析模型。将历史的时序及数值型数据输入到模型中，得到预测值和相对误差。

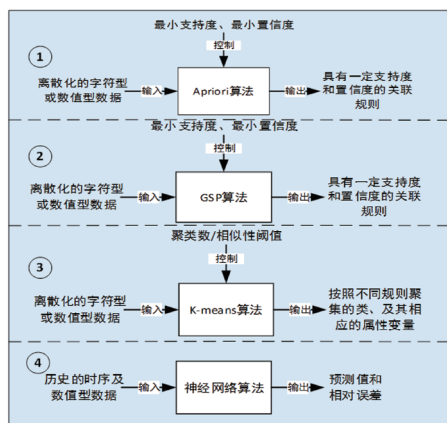


图 5 基于数据挖掘经典算法的数据分析模型

3.3 知识层

质量知识的分类是对知识的静态描述，然而随着数据量的不断增大、挖掘出的质量知识的不断增多，需要对质量知识进行动态的管理。动态管理分为知识的检测和知识的融合两个步骤。

一般来说，随着知识数量与种类的不断增多，知识库知识储备越来越多，解决问题的能力越来越强，但知识库的知识数量达到某个节点后，知识数量的增加反而会削弱知识库解决问题的能力，这是因为在知识动态管理过程中，随着知识储备量的增加，用来查找、定位和整合有用知识的周期越来越长，导致知识和其他资源的浪费。因此对于知识库实时动态检测在知识库构建中的作用非常重要，首先要针对已有的知识库进行知识检测，判断知识库中是否已经存在挖掘出来的新知识，或者相似的知识。其次，基于知识的推理过程，对这些知识之间的相互关系进行分析，如冗余关系、从属关系、矛盾关系等。基于这个原则，将推理有用的知识不断地添加融合到知识库中将能使其发挥更好的作用，这样的动态

管理也是企业信息智能发展的必然趋势。

3.4 应用层

将知识层储存和分类的知识有效地应用到需求设计、生产计划制定、质量控制、设备维修、产品销售等实际应用中，支持上层业务决策，实现基于产品生命周期的质量数据管理。

4 结论

以现在离散制造企业质量管理体系落后及产品生命周期质量数据分析的重要性为出发点，对产品生命周期质量数据来源进行研究，构建质量数据预处理模型，在此基础上提出基于产品生命周期的质量数据挖掘体系结构，对结构中的数据层、方法层、知识层、应用层的原理进行讨论，说明了该体系结构对于将数据挖掘策略用于产品生命周期质量数据分析的实际应用具有重要意义。后续还将对挖掘策略的具体实现过程进一步深入分析，并尝试不同算法结合运用及具体数据类型的精准挖掘等，使该体系内容更深入充实。

参考文献

- [1] Kusiak A. Optimising product configurations with a data-mining approach[J]. International Journal of Production Research. 2009, 47(7): 1733-1751.
- [2] Kretschmer R, Pfouga A, Rulhoff S, et al. Knowledge-based design for assembly in agile manufacturing by using Data Mining methods[J]. Advanced Engineering Informatics. 2017, 33.
- [3] Zhang K, Li G, Gong J Z, et al. Research on Assembly Data Mining Technology of Complex Mechanical System [J]. Procedia Cirp. 2016, 44: 97-101.
- [4] 郑院生, 王美清. 基于 ETL 思想的产品全生命周期质量数据整合 [J]. 成组技术与生产现代化. 2013, 30(4): 35-41.
- [5] Eraslan E. The Estimation of Product Standard Time by Artificial Neural Networks in the Molding Industry[J]. Mathematical Problems in Engineering. 2009, 2009(2): 266-287.
- [6] 曾芬芳, 陈万领, 严晓光. 基于 PDM 系统的产品生产过程中质量管理系统的研究和实现 [J]. 机械设计与制造. 2010(5): 256-257.
- [7] 高伟, 殷国富, 成尔京. 机械制造工艺序列中的知识发现方法研究 [J]. 机械工程学报. 2004, 40(5): 121-125.
- [8] 翟敬梅, 应灿, 徐晓. 知识建模和数据挖掘融合的粗糙度预测新方法 [J]. 计算机集成制造系统. 2012, 18(5): 1046-1053.
- [9] 李利强. 支持精益生产的数据挖掘技术的研究与应用 [D]. 东华大学, 2010. 