

质量嵌入的大数据产品生产系统超图模型及其生产线决策研究

王 旻¹ 蔡淑琴¹ 邹新文¹ 陈梓桐²

(华中科技大学管理学院 武汉 430074)¹ (华中科技大学软件学院 武汉 430074)²

摘 要 大数据产品(Big Data Product, BDP)在原材料、用户需求、加工工艺等方面具有不同于实体产品的特征,而现有 BDP 生产系统的研究仍停留在概念模型阶段。为了解决该问题,提出 BDP 生产线的概念,基于生产线特征研究了生产线决策要素,强调了质量作为关键决策要素在 BDP 生产中的作用机理;采用超图理论建立了嵌入质量、质量传递函数和质量聚集函数的 BDP 生产系统模型,设计了 BDP 生产线决策流程;提出了供给侧稳定和需求侧稳定的 BDP 生产线决策模式。实例验证结果表明,所提出的模型和决策方法能够满足用户对 BDP 质量的要求。

关键词 数据质量,大数据产品,生产线,超图模型

中图法分类号 TP391, G203

文献标识码 A

DOI 10.11896/j.issn.1002-137X.2019.02.002

Quality-embedded Hypergraph Model for Big Data Product Manufacturing System and Decision for Production Lines

WANG Yang¹ CAI Shu-qin¹ ZOU Xin-wen¹ CHEN Zi-tong²

(School of Management, Huazhong University of Science and Technology, Wuhan 430074, China)¹

(School of Software Engineering, Huazhong University of Science and Technology, Wuhan 430074, China)²

Abstract Due to the different characteristics of the Big Data Product (BDP) in terms of raw materials, user requirements, and processing techniques, research on BDP production system still stays at the conceptual model stage. To solve this problem, this paper purposed the concept of BDP production lines, discussed the decision factors for production line selection based on the characteristics of production line, and emphasized the action mechanism of quality as a key decision element in BDP production. This paper established the BDP production system model embedded with quality, quality transfer function and quality aggregation function by using the hypergraph theory, designed the decision processes for BDP production line, and proposed the decision mode of the BDP production line with supply-side-stable and demand-side-stable. The results show that the proposed model and decision method can meet users' requirements for BDP quality.

Keywords Data quality, Big data product, Production line, Hypergraph model

1 引言

作为一类极其重要的战略性信息资源,大数据正引领着创新、竞争和生产力发展,在电子商务、市场营销、危机管理、医疗健康、网络安全等领域展现出了巨大的应用价值^[1]。与自然资源一样,大数据资源需要被加工来实现其价值增值^[2],这一加工过程可以看作是信息产品的生产过程^[3]。大数据产品是一类以大数据资源为原材料,经过清洗、聚合、挖掘、分析等加工操作生产出的信息产品^[4]。作为大数据时代信息资源的主要形式,BDP 已成为数字经济的基础性要素,在经济社会发展 and 生产生活中扮演着重要角色。

BDP 的重要性与日俱增,其设计、生产等问题也越来越受到业界和学术界的关注^[5]。现有研究根据实体产品的设计和生产理论提出了 BDP 设计和生产系统的概念模型,其原型

是 Ballou 等提出的信息生产系统 IMS,包括数据源、加工、存储、质量控制和消费者 5 个模块^[6]。Lee 等在 IMS 的基础上补充了决策、系统边界、业务边界、系统-业务边界模块,构建了更完善的生产系统模型 CEIP maps^[7]。Meyer 等认为一个典型的数据产品生产过程应包括数据获取、提炼、存储、分发、展示等阶段^[8]。Davenport 等基于大数据环境下及时响应用户的应用情境需求,提出在 BDP 生产过程模型中加入产品概念设计和用户反馈两个阶段,构建了 BDP 生产系统七阶段闭环模型^[4]。蔡淑琴等以社会化媒体平台的用户生成内容(User-generated Content, UGC)为例,基于公理设计理论(Axiomatic Design Theory)提出了 UGC 产品族的设计映射模型,将 BDP 的设计过程转化为用户域、功能域、物理域、工艺域的映射过程^[9]。上述模型对于 BDP 的设计与生产具有战略指导意义。

在大数据资源管理中,合理安排大数据资源的生产并保

到稿日期:2018-03-27 返修日期:2018-05-25 本文受国家自然科学基金项目(71071066, 71371081)资助。

王 旻(1986—),男,博士生,主要研究方向为信息质量、大数据产品, E-mail: wangyanghim@hust.edu.cn; 蔡淑琴(1955—),女,博士,教授,博士生导师,主要研究方向为信息系统、商务智能, E-mail: caishuqin@hust.edu.cn (通信作者); 邹新文(1995—),男,硕士生,主要研究方向为信息管理; 陈梓桐(1995—),男,硕士生,主要研究方向为自然语言处理。

障高质量的 BDP 输出,成为大数据生产和服务主体面临的最大挑战之一^[10]。质量即“合用性”(Fitness for Use),用以衡量产品满足消费者实际使用需求的程度^[11]。在当前大数据资源大量涌现和质量参差不齐的情况下,BDP 质量已成为使用者、生产者、研究者所关注的重要问题^[12-14]。然而,现有大多 BDP 生产系统的研究仅停留在概念模型阶段,常常忽略了质量这一重要因数,或仅将质量控制作为加工中的一个环节,未考虑 BDP 质量问题本身的复杂性及其影响因素特征。事实上,数据产品是一类不同于实体产品的信息、知识或情报^[15],其质量与原材料、生产工艺、用户对产品的需求等因素相关^[16-17]。如表 1 所列,与实体产品相比,BDP 在原材料、生产工艺、产品需求上都具有不同特征。首先,实体产品生产中的原材料是以稳定的种类、特征、速率投入生产的,而 BDP 的原材料是具有规模大、种类多、速度快、可靠性低等特征的大数据资源^[18];其次,实体产品的需求是相对稳定的,BDP 的需求是根据使用者的情境动态变化的,不同需求的用户对 BDP 的感知有用性存在显著差异^[19];最后,实体产品的生产线较为单一且短期内一般不发生变化,而 BDP 生产线是多样且具有可替代性的。例如,产品推荐既可基于用户购买记录实现“以产品为中心的”推荐,也可基于相似用户的兴趣实现“以用户为中心”的推荐^[20];又如,对 UGC 进行情感分析,既可以采用基于情感词表的方法,也可以采用机器学习技术,但不同的生产工艺会产生质量有差异的 BDP^[21]。因此,BDP 的质量具有高度动态性,如何根据消费者的不确定的应用情境需求,合理组织异质的大数据资源,从多样化的加工工艺中选取合适的加工工艺组合,保障稳定地生产出高质量的 BDP,进而为用户持续提供优质的信息服务,成为亟待解决的关键难题。

表 1 实体产品与 BDP 的质量及其影响因素的对比
Table 1 Comparison of quality and its influence factors of physical products and BDPs

		实体产品		BDP
产品质量		静态		动态
产品质量 影响因素	原材料	同质	异质	
	生产工艺	单一	多样	
	产品需求	确定	不确定	

本文首先界定 BDP、BDP 的生产工艺、生产线、生产系统的概念,研究 BDP 生产系统中的质量传递和质量聚集机理,并基于超图理论设计质量嵌入的 BDP 生产系统模型,研究基于超图模型的 BDP 生产线决策流程、方法和模式,并通过实例进行验证。

2 BDP 生产工艺、生产线和生产系统

为建立质量嵌入的 BDP 生产系统模型,本文首先对 BDP、生产工艺、生产线、生产系统等主要概念进行界定。Lee 等将信息产品定义为能满足信息消费者需求的数据元素的集合^[22];Batini 等从产品生命周期的视角,依据数据元素所处的阶段将其分为原材料、零部件和成品^[23]。

定义 1(数据元素) 数据元素是最基本的数据单元,它包括未经加工的原始大数据资源 R ,以及对 R 加工后产生的

大数据零部件 C ,还包括对 C 进行再加工后产生的新的 C 。用 ϵ_i 表示任意数据元素, $E=\{\epsilon_1,\epsilon_2,\cdots,\epsilon_M\}$ 是数据元素的全集,则有 $E=R\cup C$ 且 $R\cap C=\emptyset$ 。

定义 2(BDP) BDP 是满足用户特定应用情境需求的数据元素 ϵ_j 的集合,记为 P ,即 $P=\{\epsilon_1^j,\epsilon_2^j,\cdots,\epsilon_N^j\}$ 。显然 $P\subseteq E$,组成 P 的数据元素既可能是 R 中的元素,也可能是 C 中的元素。

BDP 的生产可分为流程型以及离散型两个阶段。在流程型生产阶段,数据元素被加工为更高级的数据元素;在离散型生产阶段,根据用户的应用情境需求选择特定的数据元素,这些数据元素的集合 P 即交付用户的 BDP,如图 1 所示。

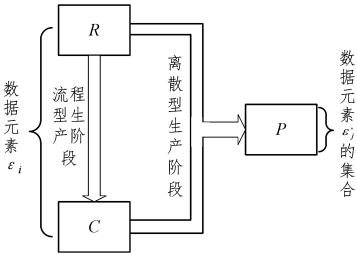


图 1 BDP 相关概念之间的关系
Fig. 1 Relations among BDP-related definitions

由于离散型生产阶段就是一个简单的数据元素集合过程,而流程型生产阶段常常包含多步、复杂、交错的加工过程,因此本文定义的生产工艺特指流程型生产,即数据元素间的工艺。

定义 3(生产工艺) 对于生产工艺 f ,设被加工的数据元素为 ϵ_{in} ,加工产出的数据元素为 ϵ_{out} ,则有:

$$\epsilon_{out} = f(\epsilon_{in}) \tag{1}$$

对于 BDP 的生产,一个加工可能存在多个输入和多个输出的情况,例如,在采用 SentiWordNet 进行基于语义的情感分析生产操作中,输入数据有待分析的文本内容和 SentiWordNet 情感词典,输出数据则有文本中包含的情感词以及量化的文本内容情感强度值。

由于 BDP 是用户所需数据元素的集合形式,每个数据元素 ϵ_j 都有其生产流程,因此 BDP 生产线的定义应建立在数据元素 ϵ_j 生产线定义的基础之上。

定义 4(数据元素的生产线) BDP 所包含的数据元素 ϵ_j 的生产线 $PL(\epsilon_j)$ 是以所需的大数据资源集 R 为输入,以 P 中特定的数据元素 ϵ_j 为输出,所采取的一系列生产工艺的有序组合。

定义 5(BDP 的生产线) 大数据产品 P 的生产线 $PL(P)$ 是 P 所包含的所有数据元素的生产线 $PL(\epsilon_j)$ 的组合。

定义 6(生产系统) 生产系统是该系统可以生产出的所有大数据产品的生产线的组合。

生产线是生产系统的核心要素,它融合了原材料加工和产品生产工艺的规范性知识。与实体产品相比,BDP 的生产线具有特殊性。首先,实体产品的生产线能同时承载的工作量极为有限,但 BDP 的生产本质是数据加工处理,因而一般情况下生产线的工作量还远远未达到饱和;其次,BDP 的生产线本质是一类可转变成软件的规范性知识,具有易移植性;

然后, BDP 的生产线多样且具有可替代性, 通常以多条生产线组成的生产系统形式存在, 具体采用哪条生产线投入生产则取决于生产线、原材料、用户需求这三者的匹配程度; 最后, 由于生产工艺以虚拟的软件方法形式存在, BDP 生产线的转换成本很低, 不涉及对基础设施的任何操作, 这一特征支持生产系统对用户需求和数据资源特征的变化做出快速响应。

3 质量嵌入的 BDP 生产系统超图模型

3.1 BDP 生产中的质量传递和聚集

在 BDP 的生产过程中, 生产工艺 f 除了将输入数据转化为特定的输出外, 对数据的质量也存在影响, 且不同的加工工艺对质量的影响也不尽相同。这个影响过程被称为质量传递 (Quality Transfer)^[12], 其影响机制可以用质量传递函数表示。

定义 7(质量传递函数) 假设数据元素 ϵ_{in} 经过了生产线操作 f 后产出了数据元素 ϵ_{out} , 则在输入和输出数据元素的质量之间存在质量传递函数 Q_{trans} , 使得:

$$Q_{eval}(\epsilon_{out}) = Q_{trans}(Q_{eval}(\epsilon_{in})) \tag{2}$$

其中, Q_{eval} 是质量测度函数。

质量传递函数体现了生产工艺 f 对数据元素质量的作用, 即流程型生产阶段中的质量变换机理。在离散型生产阶段, BDP 的质量是由用户所需的多个数据元素的质量聚集而成的, 其机理可用质量聚集 (Quality Aggregation) 函数表示^[24]。

定义 8(质量聚集函数) 假设用户所需的数据元素 ϵ_1^* , ϵ_2^* , $\dots$, ϵ_N^* 集合后产出了大数据产品 P , 则在数据元素和 BDP 的质量之间存在质量聚集函数 Q_{agg} , 使得:

$$Q_{eval}(P) = Q_{agg}(Q_{eval}(\epsilon_1^*), Q_{eval}(\epsilon_2^*), \dots, Q_{eval}(\epsilon_N^*)) \tag{3}$$

其中, Q_{eval} 是质量测度函数。

质量传递函数和聚集函数的具体形式可通过对生产工艺、产品结构、用户需求进行理论分析获得, 对于很多较复杂的加工工艺, 例如基于机器学习的方法, 很难通过简单分析得到精确的函数形式, 此时可以通过统计学方法获得函数的近似表示或经验公式。由于本文的研究重点是基于质量的 BDP 生产系统模型及其生产线决策, 因此不深入研究质量传递和聚集函数的计算方法。

3.2 超图模型

由于生产线常常包含成百上千道工序, 且工序间存在多种交织, 为实现生产线的管理和优化, 学者们采用多种技术对其进行建模, 其中图与网络的方法被广泛应用^[25-26]。然而, 在 BDP 的生产中, 数据并不一定是单输入、单输出 (一对一) 的, 常存在多对一、一对多, 甚至是多对多的情况。对于不同的输出元素, 其质量传递的机制和结果都是有区别的。传统的基于图论的方法不能有效地表达 BDP 生产的多输入-多输出、数据元素聚集等关系。超图是对普通图的推广, 具有一般性的离散结构, 对关系的表达能力优于普通图。近年来, 超图在一些新的领域, 如计算机视觉、自然语言处理、商务智能等, 得到了应用和拓展^[27-30]。本文引入超图 (Hypergraph) 理论, 建立 BDP 生产的超图模型, 并基于该模型生产线决策进行分析。

定义 9(超图模型) BDP 生产系统的超图模型是一个六元组: $HM = (V, E, E', Q, T, A)$ 。其中, 顶点集 $V = \{v_1, v_2, \dots, v_n\}$ 是 BDP 数据 ϵ_j 元素的集合, V 既包括未经加工的原始数据元素, 也包括原始数据元素经过加工后生成的新的数据元素; 超弧集 $E = \{e_1, e_2, \dots, e_m\}$ 是加工单元的集合, 超弧 e_i 的尾 $T(e_i) \subset V$ 表示加工单元的输入, 超弧 e_i 的头 $H(e_i) \subset V$ 表示加工单元的输出; 超边集 $E' = \{e_1', e_2', \dots, e_n'\}$ 表示 BDP 由数据元素到产品的装配过程, 对任意的 BDP, 有且仅有一条超边与之对应; $Q = \{q_1, q_2, \dots, q_n\}$ 是顶点集 V 对应的数据元素质量的集合, 即 $q_i = Q_{eval}(v_i)$, 质量 q_i 通常以 m 维列向量的形式存在, 其中的元素对应 m 个不同维度的质量度量值, 则数据元素的质量可推广为 $m \times n$ 的矩阵; $T = \{t_{ij}\}$ 是生产工艺的质量传递函数集合, $i = 1, 2, \dots, m$ 且 $j = 1, 2, \dots, |H(e_i)|$, 每个超弧 e_i 对应 $H(e_i)$ 个质量传递函数; $A = \{a_k\}$ 是质量聚集函数的集合, BDP 生产系统中质量聚集函数的个数即为该系统可以生产出的 BDP 的数量。

定义 10(超路径) 称顶点-超弧交错序列为一条超路径 $Hyperpath$, 即:

$$Hyperpath(v_1, v_{q+1}) = \langle v_1, e_1, v_2, \dots, e_q, v_{q+1} \rangle \tag{4}$$

超路径 $Hyperpath(v_1, v_{q+1})$ 的长度为 q , 若 $v_1 \in R, v_{q+1} = \epsilon_j^*$ 且 $v_{q+1} \in P$, 则该超路径是数据元素的一条生产线 $PL(\epsilon_j^*)$, 因此, 产品生产线和生产系统可视为超路径的组合。

与现有大多 BDP 生产系统模型相比, 本文提出的超图模型不再停留在概念建模阶段, 且具有以下优点: 1) 引入质量要素。通过引入数据质量、质量传递函数和质量聚集函数, 强调并解释了质量在 BDP 生产中的重要作用。2) 扩展描述能力。相比普通有向图, 有向超图能通过准确表达多个数据元素输入-多个数据元素输出的关系来描述 BDP 在流程型生产中的工艺逻辑; 相比普通无向图, 无向超图能通过表达多个数据元素对应多个 BDP 的集合映射关系来描述 BDP 在离散型加工阶段中的装配逻辑。3) 支持生产决策。超图模型的建立, 将 BDP 生产线的决策问题转化为超图中最优超路径的计算问题。

4 基于超图模型的 BDP 生产线决策

在建立了 BDP 的超图模型之后, 便可根据实际用户的应用需求并参考数据资源的情况进行 BDP 生产线的决策活动。

产品生产线的优化决策是生产管理中的经典问题。实体产品生产中有有限的生产设施是稀缺资源, 须采取合理的调度策略获取利用率、生产与等待时间、效率等性能参数, 对瓶颈设备和生产线能力进行评估, 科学地布局与调度生产线。BDP 生产线的特征, 使得其瓶颈不再是有限的生产设施, 而是“资源-工艺-产品”三者间的匹配程度, 这一匹配是通过质量实现的; 其决策也不再是一个工作量调度问题, 而是基于平台上现有的大数据资源及其特征, 动态地响应资源供应和用户需求的动态变化, 在生产线系统中恰当地选择生产线来保障生产出最优 BDP 的问题。其流程如图 2 所示。

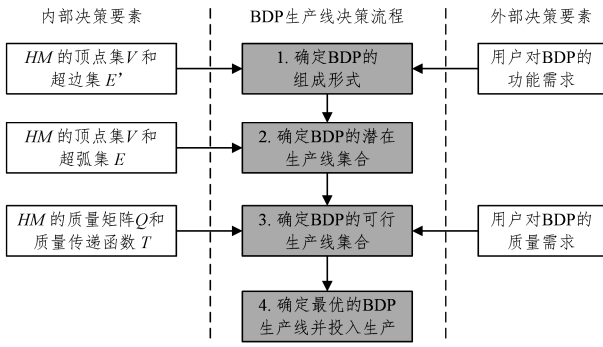


图 2 BDP 生产线的决策流程

Fig. 2 Decision process for BDP production lines

具体而言,首先根据用户的应用情境与现实需求,结合大数据资源及生产加工工艺知识,确定最终成品 BDP 的组成结构。在此基础上,确定 BDP 的潜在生产线集合。

定义 11(潜在生产线) BDP 对应的潜在生产线是指利用现有大数据资源和生产工艺能生产出满足特定用户功能需求的生产工艺的序列,所有潜在生产线的集合被称为潜在生产线集。

显然,潜在生产线不考虑数据元素和产品的质量要求,是一个纯粹的产品功能概念。每条潜在生产线都可以用超图的一个子超图来表示。对于包含多个数据元素的 BDP,其潜在生产线由其中每个数据元素的生产线(超路径)组合而来。若数据元素 v_i 的生产线的条数为 $NumHyperpath(v_i)$,则产品 BDP 的潜在生产线的条数为 $\prod_{v_i \in BDP} NumHyperpath(v_i)$ 。

定义 12(可行生产线) BDP 对应的可行生产线是指利用现有大数据资源和生产工艺能生产出满足特定用户功能和需求的生产工艺的序列,所有可行生产线的集合被称为可行生产线集。可行生产线集是潜在生产线集的子集。

本文采用递归思想设计了计算 BDP 中数据元素潜在生产线集合的算法 $CalculatingHyperpath(HM, v_p)$ 。

算法 1 $CalculatingHyperpath(HM, v_p)$

输入:超图模型 HM ,数据元素 $v_p, v_p \in BDP$

输出:数据元素 v_p 的潜在生产线集合 $\{PL(v_p)\}$

1. Initial $v_i = v_p$; /* 顶点初始化为 v_p */
2. $CalculatingHyperpath(HM, v_i)$
3. for v_i
4. find hyperarc e_j , where $H(e_j) = v_i$; /* 寻找以 v_i 为头的超弧 */
5. calculate $|e_j|$; /* 计算超弧的数量 */
6. if $|e_j| = 1$
7. set $t_j = \langle T(e_j), e_j, v_i \rangle$ as underlying; /* 工艺 t_j 是生产 v_i 的唯一路径,它必然在生产线上 */
8. $v_i = T(e_j)$; /* 沿路径回溯 */
9. else /* 以 v_i 为头的超弧多于一条 */
10. for each k where $v_k = T(e_j)$
11. $CalculatingHyperpath(HM, v_k)$ /* 递归 */
12. until $|e_j| = 1$, end; /* 直至不存在以 v_k 为头的超弧 */
13. multiplicative principle $\{t_j\}$; /* 按照乘法原理组合所有潜在的加工工艺 */
14. end;
15. end algorithm

BDP 生产者在潜在生产线的基础上进一步比较质量数

值,选择最优质的可行生产线并投入生产。显然,产品形式、用户质量需求、加工工艺、大数据资源质量都是影响可行生产线的重要因素。在众多可行生产线中,企业可以根据预期产品质量、生产效率、成本等因素决定投入其中一条或多条生产线进行 BDP 的生产活动。

在生产线决策活动的第三步中,相关主体综合考虑用户对 BDP 的质量需求、原材料的质量现状和生产线上每一个工艺的质量传递函数 3 个要素,在已确定的生产线集合中寻找具有生产能力且能满足用户需求的 BDP 生产线。这一过程可以根据实际情况在以下两种模式中选择。

1) 供给侧稳定的决策模式

当大数据资源质量特征相对稳定,而用户需求变化较多时,适合采取供给侧稳定的决策模式。该模式按照生产线“自顶向下”计算生产线上数据单元的质量参数,即首先测度未经加工的数据元素的质量,然后依据质量传递函数依次计算下游每个数据元素的质量,直到生产线结束。供给侧的质量特征较为稳定,通常并不需要对大数据资源的质量进行实时反复评估,而是更多地响应用户动态的质量需求。对于运营情况较好、用户数较多的大数据平台和较为成熟的 BDP,由于相关的数据资源能保证较高质量的持续生成,其供给侧质量特征稳定,这类决策模式较为适用,能为用户提供多元化的选择,满足其不同应用情境下的个性化需求。

2) 需求侧稳定的决策模式

当用户需求相对稳定,但大数据资源质量特征相对变化较多时,适合采用需求侧稳定的决策模式。该模式与生产方向相反,按照“自底向上”的方向计算数据元素的质量,即首先以用户对最终 BDP 成品中各元素的质量期望为基准,根据生产线上的加工工艺期望输出质量和传递函数计算期望输入质量,直到计算出所有原始数据元素的期望质量,再通过比较期望质量和实际质量,确定可行的生产线集合。在该模式下,将生产线原有的质量传递函数的输入和输出进行了调换,因此计算质量传递时需要先求解传递函数的反函数,再以超弧的头为输入、超弧的尾为输出进行计算。此外,当多个超弧的头对应同一个尾时,为保证所有工艺的质量,还需要用最大化函数进行修正。该模式适用于处于发展初期的大数据平台及相关 BDP。此时,用户对平台和 BDP 的认识和需求较为单一且稳定,但由于平台用户数十分有限,供给侧的大数据资源规模较小、质量参差不齐,因此需要侧重对供给侧资源的质量特征进行实时监控。

5 实例研究

本文通过一个社会化媒体平台 A 的 BDP 生产实例,进一步解释上述模型与方法对 BDP 生产线决策的支持机理。整个实例研究共分为 4 步。

Step1 建立社会化媒体平台 A 的超图模型

该平台上现有 5 种未经加工的原始大数据元素,分别表示为 $v_1 - v_5$,平台的数据资源处理知识能对 $v_1 - v_4$ 进行相应的加工,因此 v_5 是该平台上存在但尚未被发掘的数据资源。平台对 $v_1 - v_4$ 加工后会生成新的数据元素。最终,根据平台所掌握的生产加工技术及资源情况,得到了该平台超图模型

HM_A 的顶点集和超弧集,如图 3 所示。其中,顶点集 V_A 包含有 v_1-v_{13} 共 13 类数据元素,超弧集 E_A 包含有 e_1-e_7 共 7 种加工工艺。数据元素中, v_1-v_5 为原始资源, v_6-v_{11} 为中间资源, v_{12} 和 v_{13} 为叶子节点即成品资源。

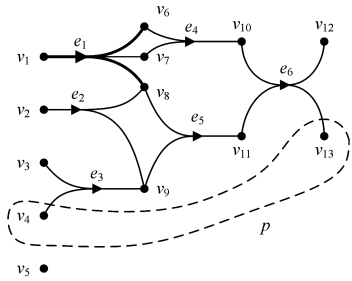


图 3 A 企业 BDP 生产系统的超图模型

Fig. 3 Enterprise A’s hypergraph model for BDP manufacturing syste

为简化问题,假设该平台只考虑准确性和实时性两个维度(记为 d_1 和 d_2),因此数据元素的质量 Q_A 是一个 2×13 的矩阵。平台现有的 7 个加工单元共存在 10 个质量传递函数,这些函数的形式是平台根据加工逻辑分析或对实际加工效果统计后得到的,假设已获得了表 2 中的传递函数。

表 2 A 企业 BDP 生产系统的质量传递函数

Table 2 Quality transfer functions for A’s BDP manufacturing system

名称	内涵	函数形式
t_{11}	v_1 通过 e_1 生产 v_6	$\begin{cases} q_6^{(1)}=q_1^{(1)} \\ q_6^{(2)}=q_1^{(2)} \end{cases}$
t_{12}	v_1 通过 e_1 生产 v_7	$\begin{cases} q_7^{(1)}=q_1^{(1)} \\ q_7^{(2)}=0.95 \cdot q_1^{(2)} \end{cases}$
t_{13}	v_1 通过 e_1 生产 v_8	$\begin{cases} q_8^{(1)}=0.8 \cdot q_1^{(1)} \\ q_8^{(2)}=(q_1^{(2)})^2 \end{cases}$
t_{21}	v_2 通过 e_2 生产 v_8	$\begin{cases} q_8^{(1)}=0.95 \cdot q_2^{(1)} \\ q_8^{(2)}=\sqrt{q_2^{(2)}} \end{cases}$
t_{22}	v_2 通过 e_2 生产 v_9	$\begin{cases} q_9^{(1)}=0.95 \cdot q_2^{(1)} \\ q_9^{(2)}=(q_2^{(2)})^2 \end{cases}$
t_{31}	v_3, v_4 通过 e_3 生产 v_9	$\begin{cases} q_9^{(1)}=\min(q_3^{(1)}, q_4^{(1)}) \\ q_9^{(2)}=\sqrt{\frac{(q_3^{(2)})^2+(q_4^{(2)})^2}{2}} \end{cases}$
t_{41}	v_6, v_7 通过 e_4 生产 v_{10}	$\begin{cases} q_{10}^{(1)}=\sqrt{\frac{(q_6^{(1)})^2+(q_7^{(1)})^2}{2}} \\ q_{10}^{(2)}=\sqrt{\frac{(q_6^{(2)})^2+(q_7^{(2)})^2}{2}} \end{cases}$
t_{51}	v_8, v_9 通过 e_5 生产 v_{11}	$\begin{cases} q_{11}^{(1)}=q_8^{(1)} \\ q_{11}^{(2)}=q_9^{(2)} \end{cases}$
t_{61}	v_{10}, v_{11} 通过 e_6 生产 v_{12}	$\begin{cases} q_{12}^{(1)}=\frac{q_{10}^{(1)}+q_{11}^{(1)}}{2} \\ q_{12}^{(2)}=q_{11}^{(2)} \end{cases}$
t_{62}	v_{10}, v_{11} 通过 e_6 生产 v_{13}	$\begin{cases} q_{13}^{(1)}=\frac{q_{10}^{(1)}+q_{11}^{(1)}}{2} \\ q_{13}^{(2)}=\sqrt{q_{11}^{(2)}} \end{cases}$

Step2 确定满足用户需求的 BDP 的结构

BDP 被定义为能满足用户需求的数据元素的集合,因而该模型中的信息产品 $BDP_i=V_i, V_i \subseteq V_A$ 且 $V_A \neq \emptyset$ 。在本实

例中,假设用户需要的信息产品仅包含 v_4 和 v_{13} 两个元素,即 $BDP_i=\{v_4, v_{13}\}$,在图 3 中用 p 将这两个元素圈起,为与超弧相区别, p 用虚线表示,至此已确定出产品的组成结构。由于 BDP_i 包含两个元素,其最终产品的质量由这两个元素的质量决定。本算例中采用最常见的加权平均质量,即 $Q(BDP_i)=w_4 \cdot q_4 + w_{13} \cdot q_{13}$,假定两个元素的权重分别为 $w_4=0.2$, $w_{13}=0.8$,且加权方法对 d_1 和 d_2 两个维度质量的影响效果相同。

Step3 计算 BDP 的潜在生产线

v_4 是平台上现存的大数据资源的一种,作为成品的组成部分,它不需要经过任何生产加工过程即可进入最终的产品装配流程,因此,本例中只须研究数据元素 v_{13} 的潜在生产线形式。采用路径求解算法 $\text{CalculatingPath}(HM, v_p)$,以最终需要的叶子节点 v_{13} 为初始节点,从对应超弧的头回溯到超弧尾,依次向上。若节点 v_i 作为头只对应一条超弧,则表明该元素只能通过一种工艺生产,说明该工艺是 BDP 生产线中必备的环节,将该超弧即超弧的尾(可能多个尾)一同放入生产线中;若节点 v_i 作为头对应了多条不同的超弧,说明生产该元素的工艺多于一种,对应的超弧及超弧的尾可作为备选生产环节。用该方法求得的共有生产线如图 4 所示。

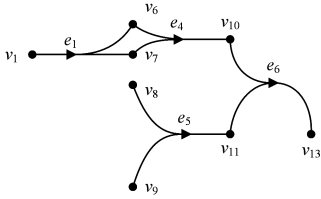


图 4 数据元素 v_{13} 的共有生产线

Fig. 4 Public production lines for data element v_{13}

图 4 的共有生产线通过位于最上端的路径已回溯到原始资源 v_1 ,但在节点 v_8 和 v_9 两处回溯时遇到阻碍,因为分别以这两个点为头的超弧不止一条。以 v_8 为头的超弧有 e_1 和 e_2 ,以 v_9 为头的超弧有 e_2 和 e_3 ,分别在这 4 条超弧上回溯,不再出现超弧不止一条的情况,故通过排列组合,共有 4 种生产线组合方式。

$$PL_1:(v_1, e_1, v_6, e_4, v_{10}, e_6, v_{13}) \cup (v_1, e_1, v_7, e_4, v_{10}, e_6, v_{13}) \cup (v_1, e_1, v_8, e_5, v_{11}, e_6, v_{13}) \cup (v_2, e_2, v_9, e_5, v_{11}, e_6, v_{13});$$
$$PL_2:(v_1, e_1, v_6, e_4, v_{10}, e_6, v_{13}) \cup (v_1, e_1, v_7, e_4, v_{10}, e_6, v_{13}) \cup (v_2, e_2, v_8, e_5, v_{11}, e_6, v_{13}) \cup (v_2, e_2, v_9, e_5, v_{11}, e_6, v_{13});$$
$$PL_3:(v_1, e_1, v_6, e_4, v_{10}, e_6, v_{13}) \cup (v_1, e_1, v_7, e_4, v_{10}, e_6, v_{13}) \cup (v_1, e_1, v_8, e_5, v_{11}, e_6, v_{13}) \cup (v_3, e_3, v_9, e_5, v_{11}, e_6, v_{13}) \cup (v_4, e_3, v_9, e_5, v_{11}, e_6, v_{13});$$
$$PL_4:(v_1, e_1, v_6, e_4, v_{10}, e_6, v_{13}) \cup (v_1, e_1, v_7, e_4, v_{10}, e_6, v_{13}) \cup (v_2, e_2, v_8, e_5, v_{11}, e_6, v_{13}) \cup (v_3, e_3, v_9, e_5, v_{11}, e_6, v_{13}) \cup (v_2, e_2, v_9, e_5, v_{11}, e_6, v_{13}).$$

图 5 用图形的方式对生产线进行了表示,粗线条表示被选定的生产线,细线条表示存在但没有被选用的生产线,星状顶点表示产品 BDP_i 中的数据元素。

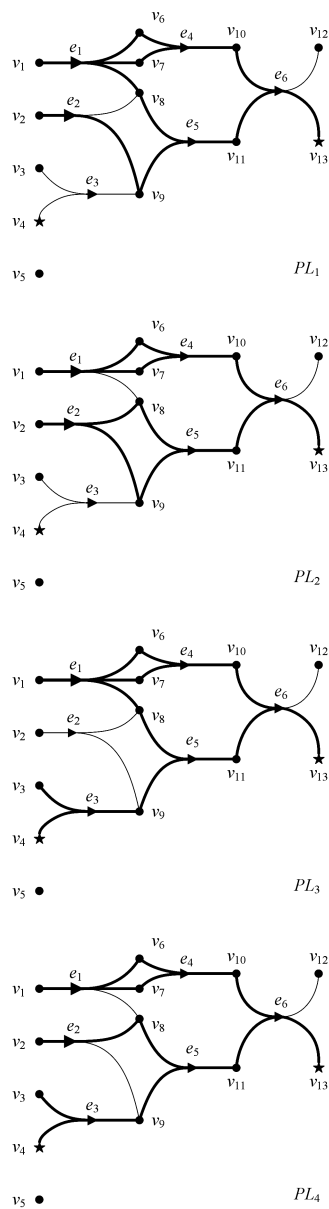


图 5 BDP 的 4 条潜在生产线

Fig. 5 Four underlying production lines for BDP

Step4 根据质量参数确定可行生产线

本步骤引入质量参数,包括用户的质量需求和数据源 v_1-v_5 的质量,再基于生产线中各环节的质量传递函数,依次判定这 4 条生产线是否可行。本文假定用户需求是稳定的,用户对最终产品的 d_1 和 d_2 两个维度上的需求分别为 0.800 和 0.900,即产品要达到的最低质量要求为 $(0.800, 0.900)^T$ 。假设数据源的质量是动态变化的,在时间 t_1 , v_1 到 v_5 的质量分别为 $(0.925, 0.950)^T, (0.825, 0.900)^T, (1.000, 0.925)^T, (0.900, 1.000)^T, (1.000, 0.900)^T$ 。在时间 t_2 ,由于某些原因,数据元素 v_2 在 d_1 维度上的质量骤降,例如此时平台上突然出现大量垃圾广告内容,导致内容质量在准确性维度上下降显著,但在另一维度实时性上没有变化,假设此时 v_2 的质量由 $(0.825, 0.900)^T$ 变为 $(0.650, 0.900)^T$,其他数据元素的质量不变,用户需求不变。表 3 和表 4 分别列出了在时间 t_1 和 t_2 下,4 条生产线产出的 BDP 的质量。

表 3 t_1 时 4 条潜在生产线的

Table 3 Data quality of four underlying production lines at t_1

	PL_1	PL_2	PL_3	PL_4
q_1	$(0.925, 0.950)^T$	$(0.925, 0.950)^T$	$(0.925, 0.950)^T$	$(0.925, 0.950)^T$
q_2	$(0.825, 0.900)^T$	$(0.825, 0.900)^T$	$(0.825, 0.900)^T$	$(0.825, 0.900)^T$
q_3	$(1.000, 0.925)^T$	$(1.000, 0.925)^T$	$(1.000, 0.925)^T$	$(1.000, 0.925)^T$
q_4	$(0.900, 1.000)^T$	$(0.900, 1.000)^T$	$(0.900, 1.000)^T$	$(0.900, 1.000)^T$
q_5	$(1.000, 0.900)^T$	$(1.000, 0.900)^T$	$(1.000, 0.900)^T$	$(1.000, 0.900)^T$
q_6	$(0.925, 0.950)^T$	$(0.925, 0.950)^T$	$(0.925, 0.950)^T$	$(0.925, 0.950)^T$
q_7	$(0.925, 0.903)^T$	$(0.925, 0.903)^T$	$(0.925, 0.903)^T$	$(0.925, 0.903)^T$
q_8	$(0.740, 0.903)^T$	$(0.784, 0.949)^T$	$(0.740, 0.903)^T$	$(0.784, 0.949)^T$
q_9	$(0.784, 0.810)^T$	$(0.784, 0.810)^T$	$(0.900, 0.963)^T$	$(0.900, 0.963)^T$
q_{10}	$(0.925, 0.927)^T$	$(0.925, 0.927)^T$	$(0.925, 0.927)^T$	$(0.925, 0.927)^T$
q_{11}	$(0.740, 0.810)^T$	$(0.784, 0.810)^T$	$(0.740, 0.963)^T$	$(0.784, 0.963)^T$
q_{12}	$(0.833, 0.810)^T$	$(0.854, 0.810)^T$	$(0.833, 0.963)^T$	$(0.854, 0.963)^T$
q_{13}	$(0.833, 0.900)^T$	$(0.854, 0.900)^T$	$(0.833, 0.981)^T$	$(0.854, 0.981)^T$
Q	$(0.846, 0.900)^T$	$(0.864, 0.900)^T$	$(0.846, 0.981)^T$	$(0.864, 0.981)^T$
(BDP)				

表 4 t_2 时 4 条潜在生产线的

Table 4 Data quality of four underlying production lines at t_2

	PL_1	PL_2	PL_3	PL_4
q_1	$(0.925, 0.950)^T$	$(0.925, 0.950)^T$	$(0.925, 0.950)^T$	$(0.925, 0.950)^T$
q_2	$(0.650, 0.900)^T$	$(0.650, 0.900)^T$	$(0.650, 0.900)^T$	$(0.650, 0.900)^T$
q_3	$(1.000, 0.925)^T$	$(1.000, 0.925)^T$	$(1.000, 0.925)^T$	$(1.000, 0.925)^T$
q_4	$(0.900, 1.000)^T$	$(0.900, 1.000)^T$	$(0.900, 1.000)^T$	$(0.900, 1.000)^T$
q_5	$(1.000, 0.900)^T$	$(1.000, 0.900)^T$	$(1.000, 0.900)^T$	$(1.000, 0.900)^T$
q_6	$(0.925, 0.950)^T$	$(0.925, 0.950)^T$	$(0.925, 0.950)^T$	$(0.925, 0.950)^T$
q_7	$(0.925, 0.903)^T$	$(0.925, 0.903)^T$	$(0.925, 0.903)^T$	$(0.925, 0.903)^T$
q_8	$(0.740, 0.903)^T$	$(0.618, 0.949)^T$	$(0.740, 0.903)^T$	$(0.618, 0.949)^T$
q_9	$(0.618, 0.810)^T$	$(0.618, 0.810)^T$	$(0.900, 0.963)^T$	$(0.900, 0.963)^T$
q_{10}	$(0.925, 0.927)^T$	$(0.925, 0.927)^T$	$(0.925, 0.927)^T$	$(0.925, 0.927)^T$
q_{11}	$(0.740, 0.810)^T$	$(0.618, 0.810)^T$	$(0.740, 0.963)^T$	$(0.618, 0.963)^T$
q_{12}	$(0.833, 0.810)^T$	$(0.771, 0.810)^T$	$(0.833, 0.963)^T$	$(0.771, 0.963)^T$
q_{13}	$(0.833, 0.900)^T$	$(0.771, 0.900)^T$	$(0.833, 0.981)^T$	$(0.771, 0.981)^T$
Q	$(0.846, 0.900)^T$	$(0.797, 0.900)^T$	$(0.846, 0.981)^T$	$(0.797, 0.981)^T$
(BDP)				

由表 3 可知,4 条生产线都能达到用户对 BDP 质量的要求,因此这 4 条生产线均为可行生产线。由于 BDP 生产线具有非饱和性,若忽略不同生产线的成本差异,应选择产品质量最高的生产线即 PL_4 进行生产。当 v_2 的 d_1 维度质量降低时,情况发生了显著变化,如表 4 所列,原来可行的 4 条生产线有 2 条出现质量下降的问题,尤其是在 t_1 时间被认为是“最优”的 PL_4 的质量无法达到要求,而 v_2 质量下降并没有对 PL_1 和 PL_3 生产线上的产品质量造成影响,这 2 条生产线是 t_2 时间的可行生产线。相对而言, PL_3 的产品质量更高,是此时最合适的选择。

结束语 以大数据为源的数字信息生产和服务正逐渐成为信息服务平台、信息管理机构、网络社区等实体的重要活动。由于大数据的特殊性,BDP 的生产与实体产品有显著区别。本文研究表明,在用户需求变化大、资源特征变化频繁时,相关实体要合理选择生产线来完成持续高质量信息产品的生产和服务,从而提升服务质量。本文提出应将“质量”作为 BDP 生产线决策的关键要素,研究了质量在 BDP 生产过程中的传递和聚集机制,提出了质量传递函数和质量聚集函数。基于超图理论,建立了质量嵌入的 BDP 生产系统超图模型,研究了超图模型下的生产线决策流程,提出了超图模型中潜在生产路径的求解算法,以及供给侧稳定和需求侧稳定的两类决策模式,它们适用于不同的应用情况。最后,通过一个实例对模型和决策方法的有效性进行了验证。本文提供了在需求、资源、工艺多变的动态服务环境中,恰当组织大数据资源、生产优质的 BDP、提升用户服务质量的管理理论与方法,其具有应用参考价值。

参 考 文 献

- [1] CHEN H,CHIANG R H L,STOREY V C. Business intelligence and analytics:From big data to big impact [J]. MIS Quarterly,2012,36(4):1165-1188.
- [2] JIN X,WAH B W,CHENG X,et al. Significance and challenges of big data research [J]. Big Data Research,2015,2(2):59-64.
- [3] WANG R Y,LEE Y W,PIPINO L L,et al. Manage your information as a product [J]. Sloan Management Review,1998,39(4):95-105.
- [4] DAVENPORT T H,KUDYBA S. Designing and developing analytics-based data products [J]. MIT Sloan Management Review,2016,58(1):83-89.
- [5] CHAO L M,XING C X,ZHANG Y. Data science studies:State-of-the-art and trends [J]. Computer Science,2018,45(1):1-13. (in Chinese)
朝乐门,邢春晓,张勇. 数据科学研究的现状与趋势 [J]. 计算机科学,2018,45(1):1-13.
- [6] BALLOU D,WANG R,PAZER H,et al. Modeling information manufacturing systems to determine information product quality [J]. Management Science,1998,44(4):462-484.
- [7] LEE Y,CHASE S,FISHER J,et al. CEIPMaps:Context-embedded information product maps[C]// Proceedings of the 13th Americas Conference on Information Systems (AMCIS 2007). Association for Information Systems,2007.
- [8] MEYER M H,ZACK M H. The design and development of information products [J]. Sloan Management Review,1996,37(3):43-59.
- [9] CAI S Q,MA Y T,XIAO Q,et al. Research on mapping for designing product family of user-generated content based on hypergraph design model [J]. Journal of the China Society for Scientific and Technical Information,2011,30(4):387-394. (in Chinese)
蔡淑琴,马玉涛,肖泉,等. 基于超图设计模型的用户创造内容产品族设计映射研究 [J]. 情报学报,2011,30(4):387-394.
- [10] YAQOOB I,HASHEM I A,GANI A,et al. Big data:From beginning to future [J]. International Journal of Information Management,2016,36(6):1231-1247.
- [11] JURAN J M. Juran on quality by design:The new steps for planning quality into goods and services [M]. New York:The Free Press,1992.
- [12] LIU J,ZHAO S Z. Research on dynamic development model information product quality assesment [J]. Computer Science,2015,42(1):244-248. (in Chinses)
刘婧,赵嵩正. 信息产品质量测量动态衍变模型研究 [J]. 计算机科学,2015,42(1):244-248.
- [13] HUANG D M,ZHAO D F,WEI L F,et al. Managing marine data as big data:uprising challenges and tentative solutions [J]. Computer Science,2016,43(6):17-23. (in Chinese)
黄冬梅,赵丹枫,魏立斐,等. 大数据背景下海洋数据管理的挑战与对策 [J]. 计算机科学,2016,43(6):17-23.
- [14] SAHA B,SRIVASTAVA D. Data quality:The other face of big data[C]// Proceedings of the IEEE 30th International Conference on Data Engineering. IEEE,2014:1294-1297.
- [15] PARSSIAN A,SARKAR S,JACOB V S. Assessing data quality for information products:Impact of selection, projection, and cartesian product [J]. Management Science,2004,50(7):967-982.
- [16] SHANKARANARAYANAN G,BLAKE R. From content to context:The evolution and growth of data quality research [J]. Journal of Data and Information Quality,2017,8(2).
- [17] KLEINDIENST D. The data quality improvement plan:Deciding on choice and sequence of data quality improvements [J]. Electronic Markets,2017,27(4):387-398.
- [18] WATSON H J. Tutorial:Big data analytics:Concepts, technologies, and applications [J]. Communications of the Association for Information Systems,2014,34(1):1247-1268.
- [19] WOODALL P. The data repurposing challenge:New pressures from data analytics [J]. ACM Journal of Data and Information Quality,2017,8(3-4):11.
- [20] RICCI F,LIOR R,BRACHA S. Introduction to recommender systems handbook [M]. US:Springer,2011:1-35.
- [21] PANG B,LILLIAN L. Opinion mining and sentiment analysis [J]. Foundations and Trends in Information Retrieval,2008,2(1-2):1-135.
- [22] LEE Y W,PIPINO L L,FUNK J D,et al. Journey to data quality [M]. Cambridge,USA:The MIT Press,2006.
- [23] BATINI C,CAPPIELLO C,FRANCALANCI C,et al. Methodologies for data quality assessment and improvement [J]. ACM Computing Surveys,2009,41(3):1-52.
- [24] TODORAN I G,LECORNU L,KHENCHAF A,et al. A methodology to evaluate important dimensions of information quality in systems [J]. Journal of Data and Information Quality,2015,6(2-3):11:1-11:23.
- [25] RAJAGOPALAN R,BATRA J. Design of cellular production systems a graph-theoretic approach [J]. The International Journal of Production Research,1975,13(6):567-579.
- [26] LIU P,LI Y,ZHANG K F. Layered orientation graph modeling method for assembly of complex products [J]. Journal of Machine Design,2007,24(4):30-32. (in Chinese)
刘平,李原,张开富. 复杂产品装配的分层有向图建模方法 [J]. 机械设计,2007,24(4):30-32.
- [27] PURKAIT P,CHIN T J,SADRI A,et al. Clustering with hypergraphs:The case for large hyperedges [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2017,39(9):1697-711.
- [28] KONSTAS I,LAPATA M. Unsupervised concept-to-text generation with hypergraphs[C]// Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies. Association for Computational Linguistics,2012:752-761.
- [29] WANG Y,WU Y,CAI S. A hypergraph model for market opportunity discovery[C]// Proceedings of the Fourth International Joint Conference on Computational Sciences and Optimization. IEEE,2011:322-326.
- [30] CHENG W,WANG B,LI Z H,et al. Study on Optimization Matching Method Oriented to E-platform for Investment and Financing[J]. Journal of Chongqing University of Technology (Natural Science),2011,25(5):75-81. (in Chinese)
成卫,王奔,李战华,等. 一种面向电子商务投融资平台的优化匹配方法[J]. 重庆理工大学学报(自然科学),2011,25(5):75-81.