DOI:10.13530/j.cnki.jlis.2024024

数据故事的内涵、生成及应用研究

朝乐门

摘　要　故事是古老的艺术和文学体裁,而数据故事是大数据时代新兴的一门科学与工程技术。数据故事的公式化定义揭示了数据故事已有定义之间的区别与联系,聚焦数据故事化中的主要矛盾,加深了对数据故事的理解层次,较好地支持数据故事的自动生成。数据故事的两个主要阶段、三类核心科学问题、四个基本特征以及五个关键要素的提出,进一步明确了数据故事的知识体系。数据故事的生成过程模型——DAIS 的提出不仅明确了数据故事生成过程中的四个关键要素——数据、分析、洞见和故事,而且深入探讨了每个阶段的工作要点。数据故事具有体验、解释和启发三种主要功能,是现实世界和虚拟世界之间的桥梁。数据故事将成为元宇宙为代表的虚实结合型应用问题研究的关键课题之一。图 8。表 3。参考文献 58。

关键词　数据故事　数据科学　数据洞见　数据分析　叙述

分类号　G203

Data Story:Definition,Methods and Applications

CHAO Lemen

ABSTRACT

Stories are an ancient genre of art and literature, while data stories are a new technique of science and technology. As a brand-new application area, data stories have attracted much attention in industry, but the academic community is also in dire need of groundbreaking research on its key issues.

The formulation of the definition of data stories unifies the differences among existing definitions of data stories, focuses on the key contradictions in data storytelling, deepens the understanding of data stories, and better supports the automatic generation of data stories. At the same time, the proposal of the two-stage theory of data storytelling not only further explores the definition of data storytelling but also decomposes data storytelling into two relatively independent activities of generation and description. Usually, there are three basic scientific questions behind data stories and data storytelling: what-if questions, why-not questions, and how-to questions, which focus on exploratory analysis, explanatory analysis, and instructive analysis, respectively. Data stories have four main features, namely data, story, business and science. These four features of data stories are proposed to better describe the differences between data stories and literary stories and data visualization works. A data story usually consists of five elements, namely the character, the event, the plot, the data insight and the business purpose. The proposal of the above five elements of data stories has eliminated the confusion of sources and components in previous research, corrected the misunderstanding of the one-sided emphasis on the status of data visualization in data storytelling, and realized the creation and narration of data stories.

通信作者:朝乐门,Email:chaolemen@ruc.edu.cn,ORCID:0000-0001-8963-7507(Correspondence should be addressed to CHAO Lemen,Email:chaolemen@ruc.edu.cn,ORCID:0000-0001-8963-7507)

The DAIS model for data story generation clarifies not only the four key stages of the data story generation process—data, analysis, insights and story—but also the work content and procedures in each stage. From a methodological perspective, data story generation methods can be divided into four types: model-agnostic global storytelling, model-specific local storytelling, model-agnostic global storytelling, and model-specific local storytelling. Data storytelling has three main functions: to experience, to explain, and to enlighten. Data storytelling is currently used primarily in data analysis and model interpretation, metaverse application and linking virtual and real worlds, teaching and training, brand advertising and digital marketing, data-driven management and decision making, and content creation and product design. Data stories bridge the real and virtual worlds, and the exploration of virtual-real synthetic data stories is becoming an important topic in the study of the metaverse. 8 figs. 3 tabs. 58 refs.

KEY WORDS

Data story. Data science. Data insight. Data analysis. Narrative.

0 引言

作为一种古老的叙述方式,故事在人类认识世界和改造世界中发挥着重要作用。一方面,故事是人类传递知识、思想、文化、传统以及价值观的主要方式之一,而故事叙述能力是演讲[1]、教学[2]、领导力[3]的基本功;另一方面,通过故事有效连接数据与受众的想法、兴趣和生活,可为受众带来愉悦与惊喜,激发他们的创造力[4]。数据故事化(Data Storytelling)将成为数据分析、商务智能、数据科学和决策支持类软件平台的一种新功能。根据 Gartner 公司的报告,到 2025 年,数据故事(Data Story)将成为最为流行的分析结果的利用方式,而且 75%的数据故事将基于增强分析技术自动生成[5]。

目前,数据故事受到数据可视化领域的高度重视,已成为数据可视化领域的前沿话题,其背后的主要原因有三个。①可视化和故事化的融合——可视故事化(Visual Storytelling)已成为数据可视化理论研究领域的一个新趋势。与此同时,数据故事化功能模块也开始出现在数据可视化类软件工具之中。数据可视化软件 Tableau 副总裁 Jock Mackinlay 和可视化分析师 Robert Kosara 提出:“带有故事元素的呈现是可视化研究的下一步,与数据探索和数据分析同等重要。”[6]②相对于文本和语音等其他呈现形式,可视化的实现难度较小,相关的理论成果和实践经验最为丰富。因此,数据可视化成为现阶段数据故事化的主要叙述方法。③与数据可视化相比,数据故事化具备更易于支持受众的记忆、认知和体验的特征[7],对受众的影响更为深刻,较容易激发受众的实际行动。在一项分组对照实验中,研究人员为同一个被捐赠对象分别制作可视化版本和故事化版本的宣传手册,实验结果显示,拿到数据故事化版本的捐赠者的平均捐赠金额比拿到数据可视化版本的对照组高出两倍以上[8]。

如何将业务中的客观数据与受众的主观认知规律相结合是开发利用数据要素的一个关键课题。通常,数据的客观性和故事的主观性之间存在互补性。数据的客观性为人们带来了真实的客观世界,但并不符合人类与生俱来的认知规律;故事的主观性具备更好的认知和体验特征,但容易缺失客观性。随着数据开发利用水平的提高,数据与故事的融合是一个必然趋势。本文结合笔者团队的学术

研究和工程实践经验,探讨数据故事化的若干核心问题,旨在进一步推动相关理论研究和工程开发。首先,探讨数据故事的五个关键问题——数据故事的公式化定义、数据故事化的两个阶段、数据故事化的三类基本科学问题、数据故事的四个主要特征以及数据故事的五个组成要素;其次,分析数据故事化过程模型——DAIS,分别提出了数据、分析、洞见和故事四个节点的工作要点及注意事项;再次,结合数据故事化中数据和模型的特点,从模型依赖性和样本指向性两个维度探讨数据故事化的四种方法;从次,结合数据故事的 EEE 作用——体验(Experience)、解释(Explain)和启发(Enlighten),探讨数据故事化的主要应用场景;最后,为相关研究提出对策与建议。

1 数据故事的内涵

深入理解数据故事的术语定义、基本科学问题、主要特征和组成要素是从事数据故事化研究的基础和前提。目前,关于数据故事内涵的研究仍处于起步阶段,相关理论研究停留于概念引进和未来畅想,尚未形成系统性较强的理论体系。为此,本节将深入探讨数据故事理论的五个核心问题。

1.1 数据故事的公式化定义

目前,数据故事或数据故事化的定义有很多,既有共性,又存在差异性。共性体现在原始材料的数据性和叙述方式的故事性上,即一致承认数据故事的原始材料为数据,而叙述方式为故事叙述;差异性体现在对数据故事化过程的关键活动的认识有所不同,可以大致分为三大类。第一类为洞见说,认为数据洞见是数据故事化的关键所在。洞见是从大数据中获得的有意义的发现,洞察是获得洞见的过程。基于数据的洞察过程及洞见结果决定故事叙述的价值、细节和冲突。例如,朝乐门提出数据故事是以满足特定业务需求为目的,以数据为原料,以数据分析和建模方法为手段,从数据中发现有价值的洞见,并以故事形式向目标受众提供的一种数据产品或服务[9]。第二类为行动说,主张数据故事化的最终目的是激发受众的行动,强调的是数据故事化的最终目的。例如,Duarte 认为数据故事的关键是基于故事的形式解释数据和激发行动[10],Dykes 提出当讲述者以数据故事的形式呈现洞见时,更容易影响决策并推动价值创造[11]。第三类为分析说,强调数据分析在数据故事化中的重要地位。例如,微软提出数据故事化是基于复杂数据和辅助分析生成具有引人入胜的叙事并告知与影响特定受众的行为[12],Gartner 公司认为数据故事化是一种将交互式数据可视化与叙述技术相结合的能力,以引人注目且易于理解的形式向决策者呈现和交付分析内容[13]。分析说与洞见说的区别在于:分析说强调的是数据分析的过程,而洞见说强调的是数据分析的结果,尤其是通过分析发现潜在的、有意思的、有价值的新模式和新成果。按照洞见说的观点,数据分析是数据故事化的必要条件,但并非充分条件,只有得出有价值洞见的数据分析活动才能对数据故事化产生有效作用。

可见,现有的对数据故事及数据故事化的定义存在两个方面的问题和缺陷:一是容易导致读者的曲解,不同学说的切入点不同,对主要矛盾和关键因素的刻画单一,不够系统;二是停留在概念层次的描述性定义方法,不利于进行数据故事的自动生成与数据故事化软件的工程实现。为此,本文在充分借鉴和继承现有定义的基础上,提出了一种公式化定义方法。该公式化定义方法可以较好地揭示不同定义之间的区别及内在逻辑关系,突出数据故事化中的主要矛盾和关键要素,为数据故事的自动生成和数据故事化软件的工程实现提供指导。具体定义方法如下:

$$Data\ Story = Insights\ (Data + Analysis) \times Story$$

根据该公式化定义,对于数据故事而言,数据(Data)和分析(Analysis)是数据洞见的输入变量,洞见(Insights)是数据和分析的激活函数,而故事(Story)是洞见结果的调节权重。

①对"Data+Analysis"的解释:在此公式中,加法表示可选关系和弱增强关系。即"Data+Analysis"的含义为:数据故事可以建立在数据集或(和)分析模型上,只要获取二者之一即可进行数据故事化。当原始数据集为敏感数据或分析模型为商业秘密时,数据故事化工作无法同时获得数据集和分析模型。但是,可以通过数据集训练出分析模型或分析模型的代理模型;同样,当分析模型为已知时,可以基于分析模型生成数据集。

②对"Insights(·)"函数的解释:在此公式中,函数代表一种加工处理过程。函数 Insights(·)是参数"Data+Analysis"的激活函数,代表的是从数据或(和)分析中发现的可以用于数据故事的细节、价值和冲突的潜在的、有意思的和有用的发现。

③对"Insights(·)× Story"的解释:在此公式中,乘法表示的是必选关系和强增强关系。即"Insights(·)× Story"的含义为:对于数据故事而言,洞见和故事是两个必要组成部分,缺一不可。相较于数据与分析之间的弱增强关系(+),故事对洞见的增强影响更为显著(×)。对于数据故事而言,故事是洞见的权重系数,对洞见具有选择、过滤、放大和缩放的功能。

在数据故事的定义中,人们往往容易将其与叙事或叙述(Narrative)、数字故事(Digital Story)、数据新闻(Data Journalism)、用数据讲故事(Telling Stories with Data)、数据驱动型故事(Data-driven Story)、可视故事化(Visual Storytelling)、交互故事化(Interactive Storytelling)、个性化故事化(Personalized Storytelling)和动态故事化(Dynamic Storytelling)等概念混淆。图 1 展示了上述术语之间的联系与区别:数据故事的创作——从数据到故事建模是其他术语所代表的活动和成果的基础,而故事模型叙述策略的选择(如可视故事化、交互故事化、个性化故事化和动态故事化)是数据故事化过程中的一个环节;数字故事、数据新闻、用数据讲故事和数据驱动型故事则是采用不同的叙述策略对数据故事模型叙述后的产品或服务形态,处于更高层次;叙事则是一种更为笼统的模糊提法,不仅可以代表故事叙述策略,而且还可以泛指故事的不同产品和服务形态。上述活动均属于数据故事化的具体叙述策略,而上述成果代表了数据故事作为产品和服务的常见形态。

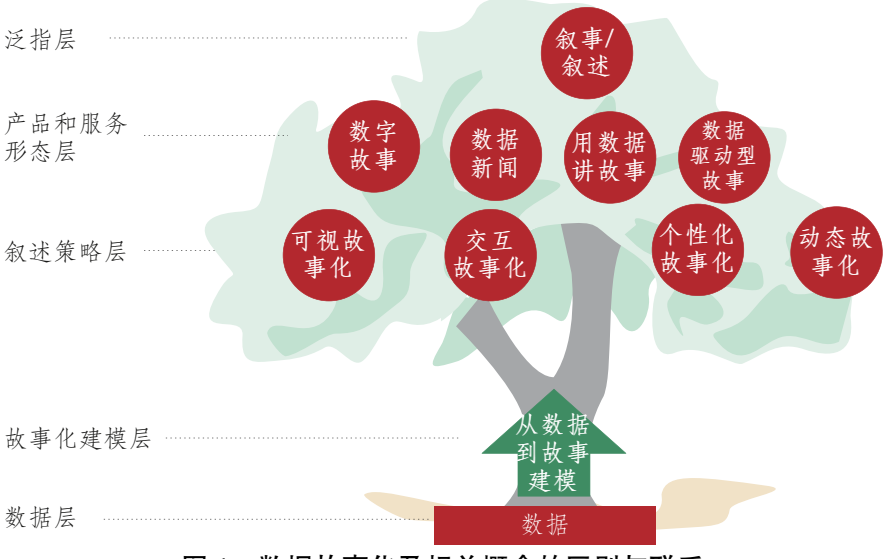

图 1　数据故事化及相关概念的区别与联系

1.2 数据故事化的两个阶段

数据故事化是与数据故事有密切联系的重要术语。数据故事化是指为了提升数据的可理解性、可记忆性及可体验性,将“数据”还原或关联至特定情景,并以叙述方式呈现的过程[14]。从二者的逻辑关系看,数据故事化是从数据到数据故事的转化过程,而数据故事是数据故事化的结果。

对于数据故事化的研究而言,区分术语故事和叙事具有重要意义。二者的区别与联系体现在两个方面:一是从叙事角度看,被叙述的对象不仅限于故事,还可以是事件、历史和人物;二是从故事角度看,针对同一个故事,故事叙述者可以采取不同的叙述策略或方式,如第一人称/第三人称叙述、文字叙述、数字化叙述和可视化叙述等。

从工程视角认识数据故事化是深入理解数据故事的一个重要突破口。为了支持数据故事的自动生成及数据故事化的工程实现,理论研究需要将数据故事化分解为两个相对独立的不同阶段:数据故事的生成与数据故事的叙述,如图2所示。

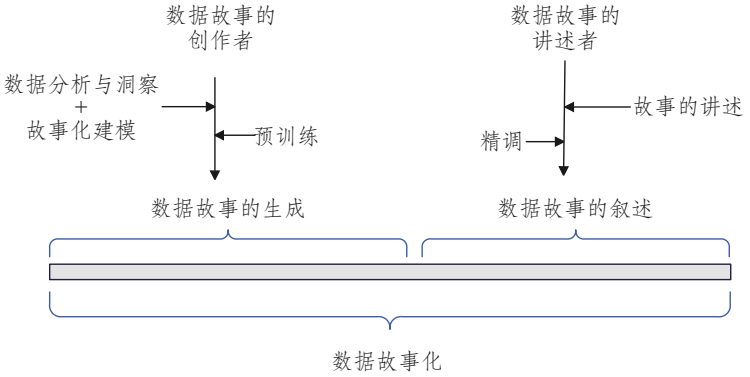

图2 数据故事化的两个阶段

从图2可以看出,数据故事的生成是故事的构思、设计和建模过程;数据故事的叙述是故事模型生成后将其叙述给受众的过程[15]。可见,数据故事生成是数据故事叙述的前提,而数据故事叙述是数据故事的具体讲述和呈现过程。

(1)数据故事的生成

数据故事的生成主体为数据故事的创作者。数据故事的生成是根据数据故事化的业务需求,采集相关数据,对其进行分析,并根据分析得出数据洞见,设计数据故事模型的过程。数据故事模型主要由数据故事的五个要素及其内在联系组成。因此,数据故事模型生成的关键在于如何根据数据洞察结果识别故事要素及要素之间的内在联系。

除了数据故事模型的组成要素之外,选取冲突情节、描写细节和价值倾向是讲好数据故事的主要关注点。然而,数据故事中的冲突、细节和价值应基于数据洞见决定,并非直接由数据或分析获得。相对于数据和分析,数据洞察是一种具有激活作用的处理过程,可以从数据和分析中发现潜在、有趣和有用的洞见,进而奠定数据故事的核心要素,提升数据故事的体验性、吸引力和生动性。

(2)数据故事的叙述

数据故事的叙述主体为数据故事的讲述者。数据故事的生成侧重于数据故事模型的预训练,而数据故事的叙述目的是将数据故事叙述给目标受众。同一个数据故事模型,可以采取不同的叙述方式,并应支持故事讲述者在不同叙述方式之间自由切换。故事讲述者可根据叙述目的及目标受众的

不同,采取不同的方式叙述,如线性/非线性叙述、作者驱动型/受众驱动型叙述、顺叙/倒叙/插叙、第一人称/第三人称叙述、可视图/文本/语音/视频/漫画图/动画/多媒体/富媒体等叙述方式。

数据故事叙述阶段的本质是一个精调的过程,通常将故事受众的数据作为输入,对数据故事模型进行精调,进而达到数据故事化的个性化叙述和互动叙述的目的。从数据故事的叙述策略看,可视化是目前最为流行的叙述策略。从数据故事化角度看,数据可视化的关键在于视觉元素和视觉通道的选择。例如,Strachnyi 在其著作 *ColorWise:A Data Storyteller's Guide to the Intentional Use of Color* 中专门讨论了数据故事的讲述者如何正确使用颜色通道[16]。

数据故事的叙述策略有多种。根据故事讲述中的主动权及互动性,可以将数据故事的叙述分为作者驱动型叙述和受众驱动型叙述,前者信息量大但缺乏互动,后者互动性高但信息量较少[17]。根据叙述策略,数据故事的叙述分为线性叙述和非线性叙述两种,前者按固定或预定的顺序叙述数据故事,叙述策略与目标受众无关;后者不设固定或预定的叙述顺序,叙述策略更加灵活,可以根据目标受众的不同,调整叙述顺序、裁剪故事内容以及与目标受众互动。马提尼酒杯结构(Martini Glass Structure)[18]是一种线性叙述与非线性叙述相结合的数据故事的叙述策略。根据叙述对象的层级,数据故事的叙述策略可分为上钻型叙述和下钻型叙述,前者代表的是自底向上的不断合并和放大过程,后者代表的是自顶向下的逐步分解和缩小过程。

数据故事的生成和叙述的分离是数据故事化工程实现的重要前提。与文学故事中的作坊式创作模式不同,数据故事的生成需要引入工程化思想,将数据故事化活动分解成故事生成和故事叙述两个不同的工程问题,进而实现数据故事的自动生成和个性化叙述,推动数据故事化的产业发展、社会分工和科学研究问题的细化。

1.3 数据故事化的三类基本科学问题

数据故事化的主要作用是实现基于数据的体验、解释和启发。其中,体验是采用故事叙述方法提高受众对数据及决策的体验效果,进而达到受众对数据的深入理解与重新认知的目的;解释是采用故事叙述方法向受众解释数据或基于数据的决策,进而达到识别数据或决策是否有偏见与歧视的目的;启发是通过数据故事激发受众实际行为的过程,进而达到故事作者和叙述者的最终业务目的。Weber 等提出数据故事的类型有三种:用数据讲述的故事、使用数据进行调查的故事、解释数据的故事[18]。Dykes 也提出类似的观点,认为数据故事化的作用是基于"叙事+可视化"的吸引、基于"数据+叙述"的解释以及基于"数据+可视化"的启发[11]。

然而,发挥数据故事的体验、解释和启发作用需要回答三个基本科学问题:What-if 类问题、Why-not 类问题和 How-to 类问题,如图 3 所示。数据故事的最小单位为故事点——由人物和事件组成的故事单元。数据故事是将故事点按照一定情节组织和建构的综合体。在数据故事中,故事事件的计

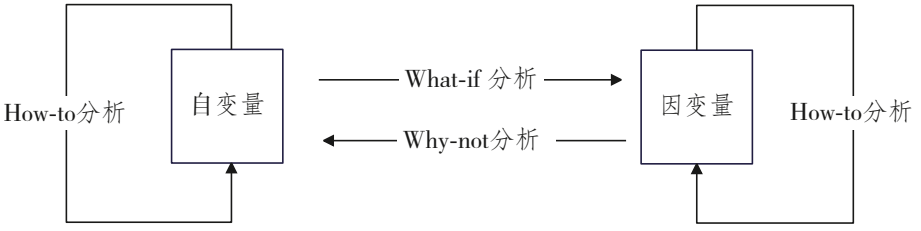

图 3 数据故事化的三类科学问题

算和情节组织应围绕数据点(样本)的特征(自变量)和目标(因变量)之间的映射和影响关系进行。在此,数据点指的是样本点,即数据故事的预训练和精调所需的数据集;特征和目标分别代表的是样本集对于业务系统而言的自变量和因变量,特征是目标的解释变量。

(1)What-if 类问题主要探讨假设性变化的结果或影响,通常以"如果……那么……?"的形式提出。What-if 类问题的回答通常采用探索型分析方法。当自变量的取值为已知而因变量的取值为未知时,数据故事化的重点工作是 What-if 分析。What-if 分析是指将自变量或特征变量的取值作为输入,在原始数据应用系统的真实业务模型或其代理模型上进行预测,得出每个样本所对应的预测值。What-if 分析关注的是目标变量的真实值和预测值之间的残差分析。What-if 分析的重点是定义合适损失函数及其最优化策略。

(2)Why-not 类问题旨在探究预期之外的结果发生的原因,通常以"为什么……?"的形式提出。Why-not 类问题的回答通常采用解释性分析方法。当因变量的取值为已知而自变量及其取值为未知时,数据故事化的重点工作是 Why-not 分析。Why-not 分析是 What-if 分析的逆向操作,即根据已知因变量推断对应的自变量及其取值,并基于自变量特征分布规律,解释某一决策的科学性与公平性。Why-not 分析的关键在于因果分析和特征分布的估计。

(3)How-to 类问题关注达成特定目标所需的策略、方法、手段或步骤,通常以"如何做到、实现或达到……?"的形式提出。How-to 类问题的回答通常采用指导性分析方法。How-to 分析的目的是对自变量和因变量的关系进行最优化分析,为受众提供最优决策建议,进而激发受众的行动,最终达到数据故事化的业务目的。

在数据故事化中,从数据点到故事点转化的关键在于回答上述三类基本科学问题。在此,数据点是指样本空间或训练数据集中一个点,通常所指的是含有自己的特征值和目标值的具体样本。从样本空间而言,数据点不仅代表故事中的人物信息(人物基本信息,如人物姓名、性别和年龄等基础属性),而且还包含一些事件信息(人物的交互行为,如购物记录、购物车数据和浏览日志等)。然而,故事点是数据故事的基本单元,主要记录的是某个事件及其对应人物的特征信息和目标值。因此,数据点和故事点具有相互映射和转化关系:What-if 分析将数据点转换为体验型故事点,Why-not 分析将数据点转换为解释型故事点,How-to 分析将数据点转换为启发型故事点,如图 4 所示。

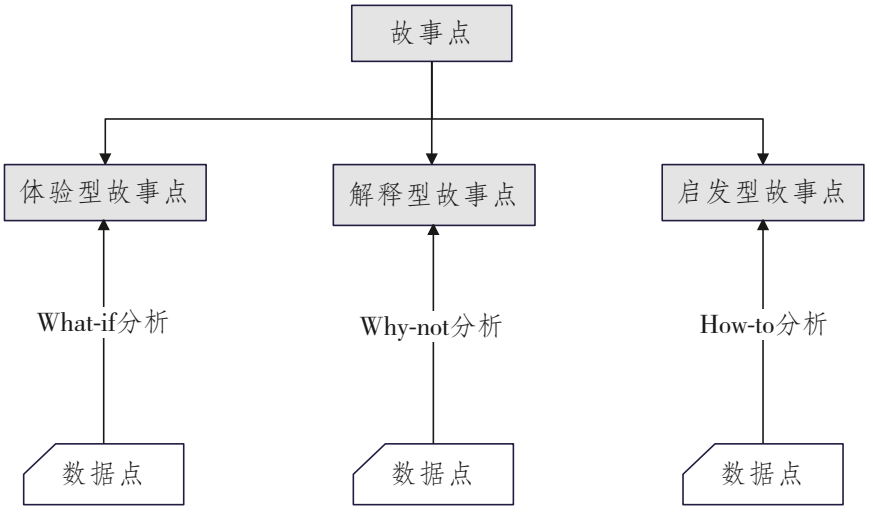

图 4 故事点的类型及与数据点、数据故事的科学问题之间的联系

1.4 数据故事的四个特征

相对于文学故事以及数据可视化作品而言,数据故事同时具备数据性、故事性、业务性和科学性四个显著特征,如图 5 所示。

图 5 数据故事的四个特征

(1)数据性是数据故事在信息来源及原材料方面的特征。从数据故事的形式化定义可以看出,数据是数据故事的原材料,数据故事的生成必须以数据分析及其洞见为基础。数据故事需要具备数据的客观性、真实性、及时性和准确性。数据性是数据故事与文学故事的主要区别之一。与数据故事不同的是,文学故事并不要求必须具备数据性,不需要以真实数据为依据,可采用纯虚构的数据。

(2)故事性是数据故事在呈现和叙述方式方面的特征。数据故事需要具备故事的短篇幅、固定情节、单一情节、重视细节描述和通俗易懂的特点。与数据可视化作品不同的是,数据故事至少包括三个要素——人物、情节(或故事线)以及叙述[19]。故事性和主观性是数据故事和文学故事的共同之处。数据故事的故事性要求应遵循以下四种原则:①数据故事的复杂度应遵循“金发姑娘原则”(Goldilocks Principle),Schramm 强调在故事叙述时应借鉴“金发姑娘原则”,即细节描述不宜过多也不宜过少,应做到恰到好处[20];同理,数据故事不能太复杂,也不能过于简单,需要掌握好“度”。②数据故事的情节设计应遵循“KISS 原则”(Keep It Simple,Stupid Principle),数据故事需要有故事情节,但不要求故事情节的多变及个性化;数据故事化中故事情节类型的选择和情节结构的设计应做到简单通用,符合多数人的思维习惯。③故事叙述时长应遵循“3 分钟原则”,数据故事的叙述不宜过于冗长,一般应控制在 3 分钟左右。④SUCCESs 原则,Heath 等提出了如何提高观点和故事粘性的六条原则,并称之为“SUCCESs 原则”——简单(Simple)、意外(Unexpected)、具体(Concrete)、可信(Credible)、情绪(Emotional)和故事(Story)原则[21]。

(3)业务性是数据故事在生成动因和目的方面的特征。数据故事需要结合实际业务、服务业务或决策需求,脱离所对应的业务活动,数据故事就没有讲述意义。通常数据故事的有效性取决于具体业务生命期,若超出所对应业务的有效期,数据故事的作用和意义也将自动消失。在数据故事化中,数据故事的创作者和讲述者需要理解目标受众,应做到“不同受众,不同故事”[22]22。数据故事的叙述需要结合实际业务,数据内容来自业务活动,数据故事的生成和叙述目的是满足真实的业务需求和决策需要。可见,业务性是数据故事与文学故事的第二个区别。

(4)科学性是数据故事在生成方法和手段方面的主要特征。数据故事需要采用科学方法和工程

技术,强调数据故事的自动生成和数据故事化的工程研发。与文学故事不同的是,数据故事并非采用想象或文学创作方法生成,而是通过科学研究、技术分析和工程实现的方式生成数据故事。故事生成和故事叙述的分离是实现数据故事科学性的重要前提,数据故事的自动生成和数据故事化活动的工程实现需要将数据故事的生成与故事叙述分离开来,并将其作为数据故事化独有的新研究课题。科学性是数据故事与文学故事的第三个区别,文学故事并不要求必须具备科学性。

1.5 数据故事的五个要素

关于数据故事的组成要素,现有的主流观点有两类。一类是五要素说,Dykes 认为数据故事的主要组成部分有五个:中心思想、探索要点、线性结构、叙述技术和视觉效果[23]。另一类是三要素说,Cote 认为数据故事有三个核心要素:数据、叙述和视觉效果[24];微软在其 Power BI 官网上也提出类似的观点,认为数据故事化的三个要素是叙述、视觉效果和数据[12]。

但是,以三要素说和五要素说为代表的现有的数据故事要素论述中普遍存在三个问题。第一,混淆了数据故事的来源与组成要素。对于数据故事而言,数据并非是直接组成部分,而是组成要素的来源和原材料。数据故事化中的"化"字强调的是类似于化学中性质转化过程,而不是物理层次的简单拼凑操作。通常,对数据进行分析,获得有价值的洞见后,才能将洞见进一步转换成故事。第二,过于片面地强调了数据可视化在数据故事化中的地位,而忽略了数据故事化中的其他叙述方式,如文本、语音、动画、漫画和游戏等。目前,多数专家将数据可视化作为数据故事组成要素的主要原因有两个,一是相对于其他叙述方式,数据可视化的实现更容易,相关方法和技术条件相对成熟;二是相关观点的提出者主要来自 Tableau、Microsoft Power BI 等可视化工具中的数据故事化研究者。然而,数据可视化并非为数据故事化的唯一叙述方式。朝乐门提出了数据故事化叙述的 TSV 模型,认为数据故事的叙述方式有三种,即文本(Text)、语音(Speech)和视觉效果(Visuals)[15]。值得一提的是,目前已有一些公司通过自然语言生成技术将数据转化为故事,如 Arria NLG、Automated Insights、Narrativa、Narrative Science 以及 Yseop 等[25]。文本和语音作为数据故事叙述方式的方法基础和技术条件已走向成熟。因此,数据故事化中片面强调可视化效果并不妥当。视觉效果并非数据故事的必要组成部分,二者可以分离,相互独立存在,数据故事不一定要通过视觉效果呈现出来。第三,没有将故事的叙述和故事的生成分开,与本文提出的数据故事化的两个阶段理论不一致,不利于数据故事的自动生成和数据故事化的工程实现。

考虑到现有理论对数据故事要素的认识存在上述问题及不足,本文提出了数据故事的五个基本组成要素——故事人物、故事事件、故事情节、数据洞见和业务目的,如表 1 所示。在数据故事中,这五个要素缺一不可、相互制衡、相互映衬,共同组成一则完整的数据故事。

(1)故事人物。人物是数据故事所谈及的人物。根据人物之间的关系,数据故事中的人物可以分为主人公、正/反面人物和同/异类人物。其中,主人公通常为故事受众本身;正面人物和反面人物的区别在于是否违背大多数人的价值观、美好心愿和道德标准;同类人物和异类人物的区别在于其观察属性是否与故事主人公一致。需要注意的是,数据故事中的人物描述相对简单,一般不需要像文学故事那样刻画人物形象和性格特征。

(2)故事事件。事件是数据故事中所发生的事情,包括故事人物本身的属性变化和故事场景等影响因素的变化。对于数据故事而言,事件的具体表现形式为样本的属性值的变化或训练模型的参数/超参数变化。当故事事件发生时,数据故事需要根据新数据或(和)新模型进行重新计算。数据故事的事

表 1　数据故事的五个组成要素

序号	要素	含　义	类　型
1	故事人物	数据故事所谈及的人物。	主人公、正/反面人物、同/异类人物
2	故事事件	数据故事中发生的事件,包括故事人物的属性变化和数据模型的参数的变化。	样本事件(如局部抖动、上钻/下钻、平移、放大/缩小、旋转和映射等)和模型事件(如回归、分类、聚类、密度估计等)
3	故事情节	数据故事中事件之间的前后顺序和因果关系,代表的是数据故事的结构和逻辑。	金字塔模型、英雄之旅模型和男孩追到女孩模型等
4	数据洞见	洞见是数据故事中激活数据和分析的重要机制,为数据故事的冲突选择、细节描写和价值趋向奠定基础。	冲突类洞见、细节类洞见和价值类洞见
5	业务目的	数据故事必须服务于某种具体的业务目的。	解释类目的、体验类目的和启发类目的

件必须是可计算的事件。通常,数据故事的事件可以分为样本事件和模型事件。样本事件是特定人物身上发生的属性变化,常见的样本事件有局部抖动、上钻/下钻、平移、放大/缩小、旋转和映射等;模型事件是分析模型的模型选择、超参数设置和参数训练结果的变化,代表的是样本数据集的更新及其背后的业务环境和业务逻辑的变化,常见的模型事件有回归、分类、聚类和密度估计等。

(3)故事情节。情节是故事事件之间的前后顺序和因果关系,代表的是数据故事的结构与逻辑,通常称之为故事线。对于数据故事而言,设计故事情节的关键在于识别事件之间的因果关系。通常,故事情节是由一个或多个事件组成。数据故事的情节描述及发展应围绕特定冲突展开,冲突是故事情节的灵魂。数据故事冲突的选择取决于数据故事化过程中的洞察结果。在数据故事中常用的情节模型有 Freytag 提出的金字塔模型、Campbell 提出的英雄之旅模型和 Vonnegut 提出的男孩追到女孩模型等[15]。与文学故事不同的是,数据故事的模型必须遵循 KISS 原则,确保故事情节的通俗易懂,避免受众的理解障碍和认知误区。此外,数据故事的自动生成与工程实现需要故事情节的自动切换能力,即可以在上述常见数据故事模型之间自动切换,进而实现对同一个数据故事采用不同的故事情节进行建模的目的,提高数据故事的体验效果。

(4)数据洞见。洞见是数据故事中激活数据来源和分析活动的重要机制,为数据故事的冲突选择、细节描写和价值倾向提供依据。Feigenbaum 和 Alamalhodaei 提出,抓住冲突以及描写好细节是写好数据故事的两个要点[22]40。然而,数据故事的冲突和细节需要由良好的数据洞见决定。通常,冲突是故事情节发展的主要推动力,而关键细节的描述可提高数据故事的生动性和吸引力。对于数据故事而言,细节描述的重要性体现在两个方面,一是提高故事的生动性和吸引力,二是具体细节描写可以提升故事本身的可信度。数据故事中的细节描述必须依赖数据洞见,而不能虚构。此外,数据故事的特征还要求数据洞察必须符合业务需求,并具备较高的业务价值。因此,数据洞见可分为冲突类洞见、细节类洞见和价值类洞见,分别代表的是数据故事中的主要冲突、关键细节以及故事叙述的业务价值。

(5)业务目的。业务性是数据故事的主要特征之一,脱离于具体业务的数据故事是无效的数据故事。数据故事必须服务于某种具体的业务目的。因此,业务目的是数据故事的必要组成部分,而业务目的的缺失会导致数据故事的业务性特征的缺失。从数据故事的作用看,数据故事中常见的业务目的可以分为解释类、体验类和启发类三大类。

数据故事的建模活动不仅需要识别上述五个基本要素,而且还应明确数据要素之间的内在联系,进而实现故事要素之间的相互关联和相互映衬。在数据故事的五个元素中,故事人物是数据故事中出现的角色及被叙述主体;故事事件是数据故事的数据分析中发生的属性和(或)模型变化;故事情节是故事事件的演化趋势和叙述结构;数据洞见是数据故事的冲突、细节和价值的来源;业务目的是故事讲述的目的。

2 数据故事的生成过程

关于如何将数据转化为数据故事,即如何做好数据故事化,已有观点大致可归为三类。第一类是与传统产品设计相同的分阶段理论。此类观点认为没有必要重新讨论数据故事化的过程,数据故事化过程与一般产品设计过程相同,数据故事化的流程并没有特殊性。因此,此类观点一般不讨论数据故事生成过程的阶段划分问题,而有选择性地提及数据故事化中相对重要的活动,比较有代表性的是 Dolan 提出的"数据+故事"模型,认为有效结合数据的客观性和故事的主观性可以达到吸引与启发受众的目的[26]。第二类是以数据为中心的分阶段理论。此类观点认为数据故事化过程应采用不同于传统产品设计过程,以数据的分析洞察为中心进行设计。因此,持此类观点的学者认为数据故事化具有自己独特的阶段划分方法,比较有代表性的是 Ryan 提出的四阶段论,即数据故事化的过程可以分为数据、分析、洞见和故事四个阶段[27]。第三类是以故事为中心的分阶段理论。此类观点也承认数据故事化过程与传统产品设计的差异性,但关注点在于故事,而不是数据。持此类观点的学者主要从故事角度划分数据故事化流程,比较有代表性的是 Andrea 提出的六阶段论,认为数据故事化过程包括设定故事化的目的、选择故事场景、设计故事情节结构、场景润色、情节打磨、呈现给受众六个阶段[28]。

从数据故事的自动生成和数据故事化的工程实现角度看,以 Ryan 为代表的四阶段论较好地解释了数据故事化与一般数据产品设计的差异性,对于理论研究和实践操作更具有指导价值。因此,本文在四阶段论的基础上,进一步明确数据、分析、洞见和故事的具体含义,强调数据故事化中的作者意图和受众决策的重要地位,并提出了一种从作者意图到受众决策的迭代模型——DAIS 模型,如图 6 所示。其中,明确作者意图是数据故事化的前提,影响受众决策是数据故事化的最终目的。

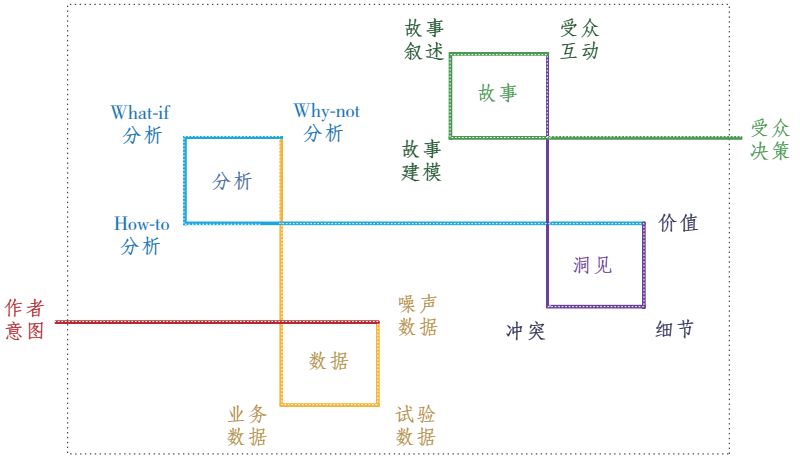

图 6 数据故事化的 DAIS 模型

2.1 数据

数据是数据故事的原材料,脱离于数据的故事并非数据故事。当然,并不是所有的数据利用都需要故事化,只有难以正确理解的数据集、容易产生曲解的数据集以及复杂模型或算法的输出数据才需要进行数据故事化。同时,通过数据可视化就可以实现数据认知和有效利用目的的任务,也不需要进行数据故事化。因此,数据故事化中的数据为复杂性较高且难以有效认知的数据对象。

数据故事化中的数据与通常所说的大数据分析任务中的数据不同。在大数据分析任务中,被分析的数据是历史痕迹类数据,也就是真实的业务数据,一般不对试验类数据进行分析。数据故事化中的数据有三种。①业务数据,即业务系统生成的真实的历史痕迹数据。②试验数据,当真实数据的特征及分布并不满足故事化需求时,为了提高数据故事的体验性,需要引入试验数据。试验数据是根据特定业务目的及故事受众给出的假设进行定义与生成。数据故事化中的试验数据必须基于业务系统的模型及其代理模型生成,不得破坏或背离对应业务逻辑。③噪声数据,在数据故事化任务中,还需要加入一定的噪声数据。噪声数据指的是那些不属于原始数据集但有助于故事叙述的额外数据。加入噪声数据是数据故事化中 What-if 分析、Why-not 分析和 How-to 分析的常用方法。

数据是数据故事的数据性特征的主要保障。数据采集的实时性和数据记录的准确性是影响数据故事质量的主要因素。在数据故事化中,数据的采集、生成和分析活动应与故事作者的意图一致,但也需要避免数据故事化“幸存者偏差”的出现——避免数据故事化中只使用容易获得的数据,而忽略难度较大但对故事化更有意义的数据。

2.2 分析

在数据故事化中,分析是指以故事化叙述为目的、对数据集进行的必要的分析活动。通常,数据故事化并非仅仅停留于将数据直接转写为故事的目的,而需要通过分析活动挖掘和发现隐藏在数据背后的更有价值的信息及逻辑,进而实现受众对数据的深层信息认知和因果关系解释的目的。因此,将数据背后的“数据的故事”讲述给受众才是数据故事的主要挑战所在。分析是讲好数据背后的“数据的故事”的前提条件和工作基础。

面向数据故事化的分析活动与一般意义上的数据分析活动不同,主要区别在于分析工作的目的和动因。数据故事化中的数据分析的目的和动因是解决数据故事化的三类基本科学问题。①What-if 分析是根据某个样本的特征取值预测其目标值。What-if 分析可以基于业务模型进行,也可以采用业务模型的代理模型来实现。What-if 分析中常用的方法是损失函数的定义和残差分析,即对数据故事化过程中预测值与观测之间的误差进行分析。②Why-not 分析的计算方向与 What-if 分析相反,即根据目标值推算特征值,进而分析样本目标值中出现故事受众所期待值时需要关注的属性集及应做出的属性变化量。Why-not 分析的关键在于特征分布和特征敏感性分析,需要从候选特征中找到最小正/负相关特征集并进行优化。③How-to 分析为目标受众提供优化方案。当数据集确定的情况下,故事受众对目标值并不满意时,需要进行 How-to 分析。数据故事化的一个重要任务是为受众提供行动建议或解决方案,而这种行动建议或解决方案必须建立在 How-to 分析及其所提供的最优解决方案上。How-to 分析的关键在于指导性分析,为受众提供有效的决策支持和自动辅助决策。

分析是数据故事的科学性和业务性两个特征的主要保障。在数据故事化中,数据分析的引入可以避免数据故事脱离真实数据与具体业务。从数据故事化过程看,分析是实现数据洞察的主要手段。如果没有科学有效的数据分析,数据故事的作者和叙述者很难从数据中发现有趣且有价值的洞见,导

致数据故事化项目的失败。

2.3 洞见

在数据故事化过程中,洞察是指通过洞察活动为数据故事找到潜在的、有意思的冲突、细节和价值。从海量多变的数据集中快速发现有价值的信息,并将其转换为故事要素是数据故事的作者和叙述者的核心竞争力之一。在数据故事化的过程中,数据洞见的选择需要同时兼顾故事作者所关注的业务目的和目标受众的利益诉求。因此,对于数据故事化而言,理解业务需求和了解目标受众是做好数据洞察的前提条件。

数据故事化过程中的洞察有三个目的。①为数据故事找到冲突。冲突是数据故事叙述的一个抓手,数据故事中的故事人物、故事事件、故事情节、数据洞见和业务目的等组成要素需要围绕故事冲突展开讨论,没有冲突,故事人物就没有成长,故事事件难以区分主次,故事情节发展容易缺乏逻辑。②为数据故事识别重要细节。数据故事的生动性和吸引力主要源自细节描述。数据故事需要提供必要的细节描述,而这种细节描述程度需要做到恰到好处,不宜过多过于细小,也不能过少过于笼统。因此,数据故事的细节描述需要以数据洞见作为依据进行过滤和推荐。③确定数据故事的价值。数据故事必须有价值,不仅对故事叙述者有价值,而且也需要对故事受众有用。通常故事叙述者和故事受众所关注的价值不同,甚至可能出现相互冲突的现象。在数据故事化中,数据洞察的另一个重要作用是不仅可以为数据故事赋予特定的业务价值,包括解释、推荐、宣传、启发、体验、教育以及改变受众的认知,而且还可以权衡数据故事的叙述者和受众之间价值期待的冲突。

洞见是数据故事的业务性和故事性两个特征的主要保障。通过洞察,对分析结果进行过滤和排序,将与业务目的无关或与目标受众关注点无关的分析结论过滤掉,并对故事化提供高质量的冲突、细节和价值。

2.4 故事

在数据故事化过程中,故事是成果呈现形式,通常基于数据洞察中发现的有趣和有意义的冲突、细节和价值,并按照数据故事五个要素模型进行生成。根据前文所述数据故事的内涵,数据故事需要具备数据性、故事性、业务性和科学性四个基本特征。同时,故事的生成和叙述需要与目标受众的认知特点和实际需求相结合,确保数据故事具有较高的可体验性,最终按照数据故事化业务需求影响受众的决策和行为。

对于数据故事化中的故事叙述而言,需要重视三项活动。①故事建模。识别数据故事中的人物、事件、情节、数据洞见和业务目的等五个基本要素,并根据数据洞见中确定的冲突、细节和价值,定义故事要素之间的内在联系。②故事叙述。将故事模型转化为故事叙述,并采用口述、视觉效果和文本等叙述方式向目标受众叙述故事。故事叙述的策略选择与故事模型相互独立,故事叙述方式的选择主要取决于目标受众,目标受众的认知习惯和思维模式是故事叙述策略选择的重要依据。③与受众互动。数据故事较高的体验性要求故事叙述环节应与目标受众进行必要的互动。当然,与受众的互动并不仅限于语言和行为上的表层互动,还可以是情感共鸣等深层互动。

故事是数据故事的故事性和业务性两个特征的又一保障因素。其中,故事性主要通过数据故事的建模实现,业务性通过故事叙述和与受众互动方式体现。当然,数据故事化中的故事并非一次性直接生成的,而需要多次迭代式的评估与改进。因此,评估和改进对故事阶段是必不可少的。数据故事

的评估内容包括故事要素及其之间联系的校验、故事模型与业务需求之间的一致性评价,以及故事叙述的效果。故事的改进是一个不断迭代和推进的过程,需要根据故事评估中发现的问题和不足,从数据故事中的数据、分析、洞见和故事四个阶段的某个环节开始调整故事化策略及活动。

3 数据故事的生成方法

从方法论角度看,数据故事的生成方法可以分为四种:模型相关的全局故事化、模型相关的局部故事化、模型无关的全局故事化、模型无关的局部故事化,如图 7 所示。

图 7 数据故事化的方法体系

3.1 全局故事化与局部故事化

全局和局部的区别在于数据故事化所对应的数据,即针对整个训练集进行数据故事化还是针对特定样本进行数据故事化。从方法论角度看,两者的区别如表 2 所示。

表 2 全局故事化与局部故事化的区别

	全局故事化	局部故事化
故事化对象	总体(样本集)	特定样本
故事化目的	叙述总体的统计特征及其变化	叙述特定样本的统计特征及其变化
故事化方法	概率分布与密度分析、规则提取、模型蒸馏和激活最大化	个体条件期望、反事实解释、对比解释法、敏感性分析、局部近似解释、梯度反向传播解释、特征反转解释、类激活映射解释

(1)全局故事化。全局故事化的分析对象为所有样本——总体,对应数据分析的目的是分析总体的整体特征,其数据洞察的目的是发现总体的规律或模式。在全局故事化中可使用的方法有四类:解释总体分布的概率分布与密度分析、以解释规则的提取为核心的“规则提取”[29]、以降低待解释模型的复杂度为特征的“模型蒸馏”[30]、以在特定层上找到神经元的首选输入最大化神经元激活为特征的“激活最大化”[31]。

(2)局部故事化。局部故事化的分析对象为特定样本,对应数据分析的目的是进行对比分析和趋势分析。在局部故事化中,样本通常作为故事主人公,整个故事的情节和事件描述围绕样本进行展开。在局部故事化中可采用的方法有很多,例如,以显示某个样本的预测值如何随着特征发生变化为目的的"个体条件期望"[32]、以因果分析为基础的"反事实解释"[33]、以相关正特征和相关负特征的洞察为基础的"对比解释法"[34]、以自变量对因变量的影响分析为基础的"敏感性分析"、以利用简单可解释的解释模型来模拟和拟合待解释模型的结果为中心的"局部近似解释"[35]、以深度神经网络的反向传播机制为基础的"梯度反向传播解释"[36]、以神经网络中间特征表征技术为基础的"特征反转解释"[37]和利用全局均值池化技术的"类激活映射解释"[38]等。

3.2 模型相关故事化与模型无关故事化

从理论上看,数据故事化过程的两种不同模型——分析模型和故事模型,分别代表的是基于数据建立的分析模型和基于洞见构建的故事模型。在此,模型是特指数据故事化任务中对应的分析模型。模型相关故事化与模型无关故事化的区别在于数据故事化的方法是否只针对某一特定的、已知的、可解释的模型,如表3所示。

表3 模型相关故事化与模型无关故事化的区别

	模型相关故事化	模型无关故事化
使用前提	模型已知且模型的可解释性较好	模型未知或模型的可解释性差
故事化特征	针对特定模型进行的专用故事化方法	模型未知或适用于多种不同模型
故事化方法	针对决策树、简单线性回归、随机森林等模型的故事化	部分依赖图、特征交互、特征重要性分析、替代模型、Shapley值解释、LIME算法和Anchors算法
主要应用场景	数据分析服务提供方自己新增故事化功能	第三方实现数据故事化

(1)模型相关故事化是针对特定可解释模型进行的专用故事化方法。在模型相关的数据故事中,数据分析模型与数据故事化方法的耦合度高,数据故事的作者对分析模型的原理及其参数具有较高的知情权和操作能力。因此,模型相关的数据故事化方法主要应用于模型已知且可解释的场景,通常由数据服务的提供方自己提供。模型相关的数据故事化方法只能用来故事化特定模型,例如针对决策树、简单线性回归、随机森林等模型的故事化。

(2)模型无关故事化指不限于特定可解释模型进行故事化,故事化方法可以应用于任何数据分析模型。故事化方法与模型之间的耦合度低,不要求数据故事的作者了解分析模型本身的原理和参数。因此,模型无关的数据故事化方法的主要应用场景为模型未知或模型可解释性差。与模型相关的数据故事化不同的是,模型无关的数据故事化方法可以由第三方提供。常用的模型无关故事化方法包括部分依赖图[39]、特征交互[40]、特征重要性[41]、替代模型[42]、Shapley值解释[43]、LIME(Local Interpretable Model Explanations)算法[44]和Anchors算法[45]。

4 数据故事的应用场景

数据故事化工作不能只停留于以故事形式呈现数据。数据故事化工作需要通过数据故事化方

式,改变目标受众的认知,进而影响其决策,最终实现特定业务目的。数据故事化的目的和主要应用场景可以总结为 EEE 模型——Experience(体验)、Explain(解释)和 Enlighten(启发),如图 8 所示。

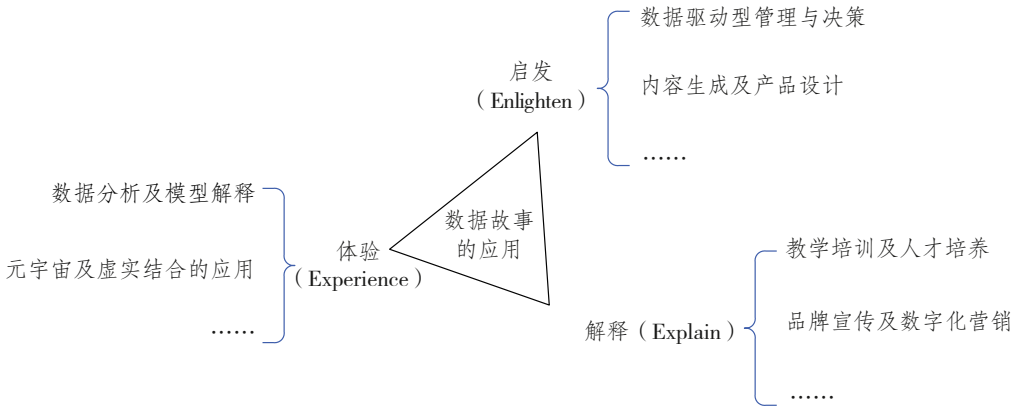

图 8　数据故事化的目的及应用场景

4.1　体验

增强体验是数据故事的基本设计目的和应用场景。相对于数据可视化等其他呈现方式,数据故事化具有更好的体验感和参与度。通过数据故事,尤其是交互式故事叙述方法,可以提高数据分析结果和决策过程的沉浸式体验,进而提升数据分析和决策可体验性。目前,数据故事化在增强体验方面的应用较为广泛,具有代表性的有以下两类应用场景。

(1)数据分析及模型解释。数据故事化是数据分析和模型解释的重要手段。目前,Tableau 和 Power BI 等数据分析软件中已提供数据故事化功能。对于这些数据分析软件而言,数据故事化主要作为数据可视化的一种扩展功能,用于对数据分析过程和结果进行故事化叙述。从未来发展趋势看,数据故事化与增强分析的融合是相关研究和实践的一个重要趋势。Yellowfin Team 曾提出数据故事化和增强分析的结合正在改变商业智能的未来[46]。2019 年,*Nature* 为纪念创刊 150 周年,将自 1869 年以来发表于该期刊的涉及 22 个研究领域和 157 个子领域的 14.6 万项研究成果,按标题、摘要、主题、关键字、作者性别、作者国别和国际合作等方面进行分析[47],得出了“小研究有大影响”等有意思的洞见[48],并用交互式叙述法提供了故事产品,具有较好的用户体验,产生了良好的社会影响,成为具有代表性的“数据—分析—洞见—故事”数据故事化过程的成功应用。

(2)元宇宙及虚实结合的应用。数据故事是建设元宇宙的基础要素,增强现实、虚拟现实、混合现实、扩展现实、平行宇宙、数字孪生、沉浸式体验、增强式再现均需要以数据故事模型为基础。Pennington 提出元宇宙是技术和故事的 1:1 组合,故事对于元宇宙很重要[49];BasuMallick 提出元宇宙的出现是从“故事讲述”到“故事生活”的转变[50],沉浸式 VR/AR“故事讲述”将会成为“故事生活”。2022 年,迪士尼公司宣称将其元宇宙定位于“下一代故事叙述”。

4.2　解释

解释是数据故事的中级设计目的和应用场景,主要用于解释因果关系,进而达到教育、说服、劝说的目的。由于数据故事采用的是故事化叙述方式,数据故事较好地继承了人类古老的教育、说服、劝说模式,对于多数受众而言不存在理解障碍,其解释效果往往高于信息图和统计指标。因此,数据故事化将应用于因果分析、可解释性机器学习和负责任人工智能等领域。目前,数据故事在解释方面的

应用场景主要有两个。

(1)教学培训及人才培养。在人类的发展史中,故事一直是教育儿童和他人的主要工具之一。通常说服力强的叙述需要同时具备证据来源可信、逻辑推理严谨和与受众情感共鸣三个特点,而数据故事正好具备这三个特点。对 TED 最受欢迎的 500 个演讲数据进行分析发现,故事占演讲者演讲时间的 65%[51];Echeverria 等探讨了面向教与学的数据故事化,并提出当复杂的可视化中有多个可能的故事时,数据故事的引入可以提高解释的清晰度[52];Matei 等提出数据故事讲述是一种教学和学习的形式,因为它要求讲述者用新的、意想不到的解释代替被认为是事实的公认解释[53]。因此,数据故事化将成为课堂教学、实验实训的一种新手段。

(2)品牌宣传及数字化营销。近年来,数据故事化开始应用于品牌宣传和数字化营销,提升了品牌宣传的解释能力和数字化营销的成功率。比较典型的成功实践是基于数据的签名故事。签名故事是指一种有趣、真实的故事,涉及具有战略信息的叙事,通过澄清或增强其品牌、客户关系、组织,使公司得以发展的业务策略[54]。签名故事是数据故事化在品牌宣传和数字营销领域的一个重要应用。著名户外用品品牌 L. L. Bean 的签名故事[55]是一个典型案例,该故事主要讲述了品牌创始人 Leon Leonwood Bean 为何设计皮靴并如何成功的故事,广泛应用于该公司的品牌宣传和数字营销之中,并取得了良好的效果。

4.3 启发

启发是数据故事的最高级设计目的和应用场景。相对于体验和解释,启发类目的的难度最大,但其价值更为重要。数据故事的启发类应用是指通过向目标受众讲述数据故事的方式,改变其认知,进而做出故事叙述者所期待的决策和行为。数据故事的启发功能的难度在于把握“度”,超过或达不到这个“度”的故事都将失去启发能力。就目前而言,数据故事的启发类应用主要体现在两个方面。

(1)数据驱动型管理与决策。数据驱动型管理与决策是区别于传统的目标驱动或模型驱动型管理与决策的一种新范式。相对于传统的决策和服务模式,数据驱动型管理与服务更加重视以数据为中心的设计模式,强调数据的驱动作用,具有更高的敏捷性。数据故事在管理与决策中的引入可以提高其敏捷性和动态性。Scott 明确提出为实现数据故事化的应用,应从数据分析和用数据讲故事拓展至数据管理[56]。数据故事将成为包括数据管理在内的多种管理工作的主要手段之一,进而提升管理活动的动态性和决策行为的敏捷性。

(2)内容生成及产品设计。数据故事本身具有较固定的故事内容要素、较高吸引力的情节结构和较强说服力的叙述体验。产品设计师可以使用故事化叙述方式洞察用户,与他们建立情感连接,更好地理解用户对产品设计的需求[57]。因此,数据故事受到内容生成和产品设计领域的广泛关注和应用。其中,数据新闻是现阶段数据故事最成功的应用领域之一。对 2013—2016 年“全球编辑网络数据新闻奖”的 44 部获奖作品分析发现,数据故事化的类型有七种:反驳声明、揭示意外后果、揭示个人感兴趣的信息、使对现象有更深入的了解、揭示系统中的异常和缺陷、跟踪系统中的变化、以越来越详细的程度揭示有关实体的信息[58]。目前,数据故事化已成为以数据新闻为代表的数据内容生成和数据产品设计的一个重要辅助工具。

5 结语

正确理解数据故事及数据故事化的内涵与特征是从事相关研究的必要条件,也是推动数据故事

化领域理论创新、产业发展、人才培养的重要前提。但是,目前对数据故事及数据故事化两个核心术语的解读方式并不系统,多数流于表面,缺乏深刻性,严重限制了数据故事相关研究的深度、信度和效度。为此,本文改变了现有的研究视角,加深研究层次,提出了数据故事领域的一些创新性观点,包括数据故事的公式化定义、数据故事化的三类基本科学问题、数据故事的四个特征和五个组成要素。

数据故事化的目的不能仅仅停留于数据的故事化描述,还需要服务于具体业务和目标受众。数据故事之所以得到产业和学术界的共同关注,是因为数据故事具有包括数据可视化在内的其他数据呈现形式不具备的体验、解释和启发能力,并能够转化为实际生产力和经济价值。通常,数据故事是围绕一项具体业务生成和叙述,进而改变受众的认知与决策,并将数据洞见转换为目标受众的具体行动。因此,相对于数据可视化,数据故事化的作用力更大,应用价值更高。目前,数据故事化主要作为数据可视化软件的一个新功能的形式存在。随着自然语言理解和生成技术的发展及文本和语音转换能力的提高,文本和语音也会成为故事叙述模式。未来的数据故事化软件并不仅限于数据可视化,文本、语音、动画、视频、多媒体和富媒体也将成为数据故事化的重要叙述方式。

目前,数据故事化研究的主要瓶颈在于如何完成数据故事的自动生成及数据故事化的工程化实现。数据故事化的研究不仅需要强调叙述层次的科学问题,更需要重视另外两个新问题:一是基于机器学习和人工智能的数据故事自动化生成技术,包括数据故事的自动化建模和形式化描述;二是引入工程化方法完成数据故事化的技术实现,进而提升数据故事化过程的高效性、标准化和规模化,降低数据故事化过程的成本,加强与数据故事化软件研发团队的合作与分工。为此,本文提出了数据故事的形式化定义、数据故事化的两阶段理论和数据故事化的 DAIS 模型。

近年来,以元宇宙为代表的虚实结合型应用成为社会新关注点。数据故事处于现实世界和虚拟世界的交叉之处,是二者之间的桥梁。数据故事可以打通现实世界和虚拟世界,不仅可以成为从现实世界转入虚拟世界必经的建模过程,也是从虚拟世界转变为现实世界的行动指南。虚实结合的数据故事的特点是以数据为基础,以增强现实和虚拟现实为表现手段,通过在虚拟世界和现实世界之间建立映射关系,抽象出二者之间共享的故事模型。未来,数据故事将成为以元宇宙为代表的虚实结合型应用领域的新研究课题。

致谢:本文系国家自然科学基金项目“预测性分析结果的数据故事化描述方法及关键技术”(项目编号:72074214)的研究成果。

参考文献

[1] RYCHKOVA I. Storytelling in Ted Talks[D]. Oxford,USA:The University of Mississippi,2020:1-72.

[2] SATRIANI I. Storytelling in teaching literacy:benefits and challenges[J]. English Review:Journal of English Education,2019,8(1):113-120.

[3] DOLAN G,NAIDU Y. Hooked:how leaders connect,engage and inspire with storytelling[M]. New Yok:John Wiley & Sons,2013:1-208.

[4] KRZYWINSKI M,CAIRO A. Storytelling[J]. Nature Methods,2013,10:687.

[5] Gartner. Top 4 data & analytics trends in finance[R/OL]. (2021-09-30)[2022-12-26]. https://emt.gartnerweb.com/ngw/globalassets/en/finance/documents/trends/top-4-data-and-analytics-trends-finance.pdf.

[6] KOSARA R, MACKINLAY J. Storytelling: the next step for visualization[J]. Computer, 2013, 46(5): 44-50.

[7] 朝乐门. 数据科学理论与实践[M]. 第3版. 北京: 清华大学出版社, 2022: 138. (CHAO L M. Data science theory and practice[M]. 3rd ed. Beijing: Tsinghua University Press, 2022: 138.)

[8] DYKES B. Data storytelling: the essential data science skill everyone needs[EB/OL]. (2016-03-31)[2022-12-26]. https://www.forbes.com/sites/brentdykes/2016/03/31/data-storytelling-the-essential-data-science-skill-everyone-needs/?sh=695cd34152ad.

[9] 朝乐门. 数据故事的自动生成与工程化研发[J]. 情报资料工作, 2021, 42(2): 53-62. (CHAO L M. Automatic generation and engineering research & development of data stories[J]. Information and Documentation Services, 2021, 42(2): 53-62.)

[10] DUARTE N. Data story: explain data and inspire action through story[M]. Oakton: Ideapress Publishing, 2019: 1-224.

[11] DYKES B. Effective data storytelling: how to drive change with data, narrative and visuals[M]. New York: John Wiley & Sons, 2019: 37.

[12] Microsoft. What is data storytelling[EB/OL]. [2022-12-26]. https://powerbi.microsoft.com/en-us/data-storytelling/.

[13] RICHARDSON J, SCHLEGEL K, SALLAM R, et al. Magic quadrant for analytics and business intelligence platforms [R/OL]. (2021-02-15)[2022-12-26]. https://b2bsalescafe.files.wordpress.com/2020/04/gartner-magic-quadrant-for-analytics-and-business-intelligence-platforms-feb-2020.pdf.

[14] 朝乐门, 张晨. 数据故事化: 从数据感知到数据认知[J]. 中国图书馆学报, 2019, 45(5): 61-78. (CHAO L M, ZHANG C. Data storytelling: from data perception to data cognition[J]. Journal of Library Science in China, 2019, 45(5): 61-78.)

[15] 朝乐门. 数据故事化[M]. 北京: 电子工业出版社, 2022: 65. (CHAO L M. Data storytelling[M]. Beijing: Publishing House of Electronics Industry, 2022: 65.)

[16] STRACHNYI K. ColorWise: a data storyteller's guide to the intentional use of color[M]. Sebastopol: O'Reilly Media, 2022: 57-69.

[17] SEGEL E, HEER J. Narrative visualization: telling stories with data[J]. IEEE Transactions on Visualization and Computer Graphics, 2010, 16(6): 1139-1148.

[18] WEBER W, ENGEBRETSEN M, KENNEDY H. Data stories: rethinking journalistic storytelling in the context of data journalism[J]. Studies in Communication Sciences, 2018, 18(1): 191-206.

[19] SLEEPER R. Practical tableau: 100 tips, tutorials, and strategies from a Tableau Zen Master[M]. Sebastopol: O'Reilly Media, 2018: 533-534.

[20] SCHRAMM J. A refresher on storytelling 101[J]. Harvard Business Review, 2014(10): 1-3.

[21] HEATH C, HEATH D. Made to stick: why some ideas survive, and others die[M]. New York: Random House, 2007: 25-238.

[22] FEIGENBAUM A, ALAMALHODAEI A. The data storytelling workbook[M]. London: Routledge, 2020.

[23] DYKES B. Data storytelling[J]. Research Wold, 2015(9): 11-13.

[24] COTE C. Data storytelling: how to tell a story with data[EB/OL]. (2021-11-23)[2022-12-26]. https://online.hbs.edu/blog/post/data-storytelling.

[25] VEEL K. Make data sing: the automation of storytelling[J]. Big Data & Society, 2018, 5(1): 1-8.

[26] DOLAN G. Stories for work: the essential guide to business storytelling[M]. New York: John Wiley & Sons, 2017: 10.

[27] RYAN L. Visual data storytelling with Tableau:story points,telling compelling data narratives[M]. London:Addison-Wesley Professional,2018:10.

[28] DE ANDREA M. Data analytics made easy[M]. Mumbai:Packt Publishing,2021:308.

[29] MASHAYEKHI M,GRAS R. Rule extraction from decision trees ensembles:new algorithms based on heuristic search and sparse group Lasso methods[J]. International Journal of Information Technology & Decision Making,2017,16(6):1707-1727.

[30] FROSST N,HINTON G. Distilling a neural network into a soft decision tree[J/OL]. arXiv,2017[2022-12-26]. https://arxiv.org/abs/1711.09784.

[31] POERNER N,ROTH B,SCHÜTZE H. Interpretable textual neuron representations for NLP[J/OL]. arXiv,2018[2022-12-26]. https://arxiv.org/abs/1809.07291.

[32] GOLDSTEIN A,KAPELNER A,BLEICH J,et al. Peeking inside the black box:visualizing statistical learning with plots of individual conditional expectation[J]. Journal of Computational and Graphical Statistics,2015,24(1):44-65.

[33] WACHTER S,MITTELSTADT B,RUSSELL C. Counterfactual explanations without opening the black box:automated decisions and the GDPR[J]. Harvard Journal of Law & Technology,2017,31:841-887.

[34] DHURANDHAR A,CHEN P,LUSS R,et al. Explanations based on the missing:towards contrastive explanations with pertinent negatives[C]//Proceedings of the 23rd International Conference on Neural Information Processing Systems. New York:Curran Associates Inc.,2018:590-601.

[35] GUO W,MU D,XU J,et al. LEMNA:explaining deep learning based security applications[C]//Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security. New York:Association for Computing Machinery,2018:364-379.

[36] SMILKOV D,THORAT N,KIM B,et al. SmoothGrad:removing noise by adding noise[J/OL]. arXiv,2017[2022-12-26]. https://arxiv.org/abs/1706.03825.

[37] DU M,LIU N,SONG Q,et al. Towards explanation of DNN-based prediction with guided feature inversion[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York:Association for Computing Machinery,2018:1358-1367.

[38] TAGARIS T,SDRAKA M,STAFYLOPATIS A. High-resolution class activation mapping[C/OL]//2019 IEEE International Conference on Image Processing. New York:IEEE,2019:4514-4518.

[39] FRIEDMAN J H. Greedy function approximation:a gradient boosting machine[J]. Annals of Statistics,2001,29(5):1189-1232.

[40] MOLNAR C,CASALICCHIO G,BISCHL B. Quantifying interpretability of arbitrary machine learning models through functional decomposition[J/OL]. arXiv,2019[2022-12-22]. https://arxiv.org/abs/1904.03867v2.

[41] DU M,LIU N,HU X. Techniques for interpretable machine learning[J]. Communications of the ACM,2019,63(1):68-77.

[42] NÓBREGA C,MARINHO L. Towards explaining recommendations through local surrogate models[C]//Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing. New York:Association for Computing Machinery,2019:1671-1678.

[43] ANCONA M,ÖZTIRELI C,GROSS M. Explaining deep neural networks with a polynomial time algorithm for shapley values approximation[J/OL]. arXiv,2019[2022-12-26]. https://arxiv.org/abs/1903.10992.

[44] RIBEIRO M T,SINGH S,GUESTRIN C. Why should I trust you explaining the predictions of any classifier[C]//Pro-

ceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: Association for Computing Machinery, 2016: 1135-1144.

[45] RIBEIRO M T, SINGH S, GUESTRIN C. Anchors: high-precision model-agnostic explanations[C]//Proceedings of the AAAI conference on artificial intelligence. New Orleans: AAAI Press, 2018: 1527-1535.

[46] Yellowfin Team. How data storytelling and augmented analytics are shaping the future of BI together[EB/OL]. [2022-11-01]. https://www.yellowfinbi.com/blog/how-data-storytelling-and-augmented-analytics-are-re-defining-bi-together.

[47] MONASTERSKY R, VAN NOORDEN R. 150 years of *Nature*: a data graphic charts our evolution[J]. Nature, 2019, 575(7783): 22-24.

[48] GATES A J, KE Q, VAROL O, et al. *Nature*'s reach: narrow work has broad impact[J]. Nature, 2019, 11(575): 32-34.

[49] PENNINGTON A. The metaverse is a 50/50 mix of technology and storytelling. The storytelling part counts[EB/OL]. (2022-05-01)[2022-12-26]. https://amplify.nabshow.com/articles/ic-the-metaverse-is-a-50-50-mix-of-technology-and-storytelling-the-storytelling-part-counts/.

[50] BASUMALLICK C. What is the metaverse? Meaning, features, and importance[EB/OL]. (2022-10-10)[2022-12-26]. https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-metaverse/.

[51] GALLO C. Talk like TED: the 9 public speaking secrets of the world's top minds[M]. London: Pan Macmillan, 2014: 37.

[52] ECHEVERRIA V, MARTINEZ-MALDONADO R, BUCKINGHAM S S. Towards data storytelling to support teaching and learning[C]//Proceedings of the 29th Australian Conference on Computer-Human Interaction. Queensland, Australia, 2017: 347-351.

[53] MATEI S A, HUNTER L. Data storytelling is not storytelling with data: a framework for storytelling in science communication and data journalism[J]. The Information Society, 2021, 37(5): 312-322.

[54] AAKER D, AAKER J L. What are your signature stories[J]. California Management Review, 2016, 58(3): 49-65.

[55] AAKER D. Creating signature stories: strategic messaging that persuades, energizes and inspires[M]. New York: Morgan James Publishing, 2018: 6.

[56] SCOTT T. Telling your data story: data storytelling for data management[M]. Sedona, USA: Technics Publications, 2020: 1-196.

[57] GUIMARAES F, AELA E. Storytelling: how to tell the right story[EB/OL]. (2022-09-05)[2022-12-26]. https://aelaschool.com/en/userexperience/storytelling-tell-right-story/.

[58] OJO A, BAHAREH R H. Patterns in award winning data storytelling: story types, enabling tools and competences[J]. Digital Journalism, 2018, 6(6): 693-718.

朝乐门 数据工程与知识工程教育部重点实验室(中国人民大学)研究员,中国人民大学信息资源管理学院教授,信息管理与分析系主任,博士生导师。北京 100872。

(收稿日期:2023-01-03)