基于张量的智慧校园服务体系构建

高熠徐 [1]
GAO Yixu

摘 要 随着各级院校信息化建设的高速发展，校园大数据治理与分析技术将推动各级院校从“数字校园”向“智慧校园”转变。然而在校园大数据治理与分析过程中仍存在诸多问题，包括如何将多源异构校园数据进行整和统一表示、如何对大数据进行高效降维存储与增量更新、如何在不破坏数据原有结构前提下对多源整体数据进行多维关联分析等。为有效解决以上问题，文章提出了一种基于张量的智慧校园服务体系，并在此基础上构建基于张量的大数据表示与融合模型、基于张量链分解的高维数据多维关联分析模型，全面提升学校大数据整合治理与多维分析能力。基于张量的智慧校园服务体系及其相关模型的构建，不仅有效解决了当前面临的多源异构数据处理难题，且在实际应用中展现出对教育质量提升与个性化学习促进的积极作用。

关键词 张量；多维关联分析；教育大数据；智慧校园；张量链分解

doi：10.3969/j.issn.1672-9528.2024.12.002

0 引言

全国校园的信息化发展历程显著，已逐渐由业务信息化和数据资源采集为主导的“数字校园”模式，转型为以数智化深度分析和一站式服务为基石的“智慧校园”架构 [1]。转型过程中，各类业务系统与智能化软硬件设施协同工作，产生了海量的校园数据。这些校园大数据具备容量大、产生速度快、数据类型多样、彼此分割互操性不强、数据价值密度低等特性。而目前各级院校缺少对其融合统一并进行多维关联分析挖掘的统一平台，难以实现“以海量保证信息的完备性，以关联分析挖掘隐含的规律性，以预测能力提升决策的科学性”[2]。

1 相关研究

在大数据技术的背景下，不少研究者就各级院校的智慧校园体系提出了详尽的数据治理理论框架和体系架构 [3]，并持续对信息业务系统进行优化升级 [4-5]，旨在打破数据孤岛，实现数据资产的统一管理和标准化。当前，部分院校已成功搭建起数据共享治理平台，有效促进了数据资源的整合与共享 [6-7]。然而，现有的理论框架和数字校园平台仅聚焦于对校园数据的采集与基础性维护，对更高层次的数据挖掘和智能研发等模块的构建仍然缺乏，未形成数智化、一站式的校园大数据治理与分析服务体系 [8]。并且在数据分析与应用层面，现有的研究大多需要依据给定的应用场景，再开展数据挖掘分析 [9-10]，并未总结出基于服务体系的通用数据智能研发、应用模式。

校园大数据治理与分析是依据海量的校园数据，通过关联分析挖掘出其中隐含的规律，从而做出科学的预测决策或总结的过程。其中，关联分析挖掘是治理与分析的关键所在，这要求大容量的多源异构校园数据在数据结构上具有统一表示，同时关联分析应是多维分析的，并且是在多源数据融合于统一空间的基础上进行的。传统的数据表示方法面对结构化、半结构化和非结构化数据在存储格式、编码方法、数据特征等方面的差异，难以有效地统一量化、整体表示 [11]。张量相较于传统的数据表示方式，可以统一表示多源异构数据，能将多源数据进行融合形成高维数据表示，并在此基础上进行分析与计算，从而更有效地分析和挖掘大数据中的复杂模式和关系，并应用于预测、推荐和聚类等应用中。

综上所述，本文提出的基于张量的智慧校园服务体系，通过构建基于张量的大数据表示与融合模型、基于张量链分解的高维数据多维关联分析模型，全面提升各级院校大数据整合治理与多维分析能力，推动校园治理现代化。

2 基于张量的智慧校园服务体系

基于张量的智慧校园服务体系，包括基础硬件平台、基

1. 华东师范大学软件工程学院 上海 200241

[基金项目] 2024 年浙江省职业技能教学研究所重点课题“在校行为数据驱动的技工院校学生学业预警体系构建”（2024-01-026）；2024 年浙江省人力资源和社会保障课题“在校行为数据驱动的技工院校学生学业预警体系构建研究”（2024215）

于张量链的数据治理平台、基于张量链的智能数据研发平台、智能应用服务平台。其中数据治理平台注重于对数据的管理，而智能数据研发平台则更注重于对数据的分析与对数据智能模型的研发。

2.1 基础硬件平台

基础硬件平台是保障整个智慧校园服务体系框架的基础硬件环境，采用校内专有云与商业公有云混合云模式为校园内各类核心信息化业务系统提供计算服务。其中，将私密数据存放在校内专有云中，而将不涉及私密数据的业务部署于商业公有云上，以此在保证数据安全的情况下，充分利用公有云可靠性、专业运维、快速资源扩容等优点。

2.2 基于张量链的数据治理平台

基于张量链的数据治理平台偏重于数据管理，主要分为三个步骤流程：

（1）数据采集、接入与同步。

（2）数据清洗、数据的张量统一表示、基于张量链的数据融合以及数据分析。

（3）面向服务应用分析需求统一提供数据、模型服务。

具体实现过程如下：

（1）由学校信息数据部门牵头，进行数据编码标准与程序代码编写标准的确立。

（2）从校园传感设备与服务采集、接入相关的结构化、半结构化、非结构化原始数据，作为原始数据层。

（3）通过张量对原始数据层多源异构数据进行统一表示，具体构建张量的方法可参考文献，提出了一个统一的张量模型，用于表示非结构化、半结构化和结构化数据，并通过张量扩展算子将这些数据表示为子张量，然后合并为统一的张量。

（4）在张量数据的基础上对这些多源子张量先进行张量链分解 [12] 得到张量链形式的数据，再在张量链的基础上实现张量融合操作 [13]，使其合并为围绕各个校园主体的融合张量，并以张量链形式进行分布式存储。

（5）在智能数据研发平台的数据分析与智能数据开发过程中，为其提供所需要的张量链形式的张量数据。

2.3 基于张量链的智能数据研发平台

基于张量链的智能数据研发平台注重于数据分析与数据智能模型的研发，包括数据同步、离线开发、实时开发、算法开发、数据 API 以及任务调度等模块。数据同步在基于张量链的增量式更新模型的基础上进行，基于张量链数据的增量式更新在处理流式数据时，可以有效避免在更新数据时对原始数据进行重复计算；离线开发使用相关基于张量链的张量计算方法，对张量链数据进行清洗、融合与分析；在实时开发模块中，针对具有高实时性需求的流式数据场景，利用张量链增量式更新模型，确保数据的实时清洗、融合与分析的高效执行；算法开发过程通过模块化的组件拖拽机制简化对张量链数据建模与训练的复杂性；数据 API 将平台产出数据组装成接口为上层智能应用服务平台提供服务。学校基于上述智能数据研发平台，可针对学生成绩预警系统、疫情精准防控系统以及校园数据可视化大屏等多个关键应用场景进行数据智能的研发，促进各项业务的稳定运行和持续优化。

2.4 智能应用服务平台

智能应用服务平台根据实际应用场景，基于智能数据研发平台提供的数据分析结果，为校园各部门用户提供高质量的应用服务，包括协助校园治理、辅助科学决策以及其他通过通用模块化组件进行建模的数据智能应用。

3 基于张量的校园大数据表示与融合模型

大数据包括结构化、半结构化、非结构化数据，校园大数据同样具有上述特性。而传统的智慧校园服务体系面对多源异构大数据在存储格式、编码方法、数据特征等方面的差异，难以有效地统一量化、整体表示。如图 1 所示，基于张量的校园大数据表示与处理框架中，来自不同数据源的异构数据，通过相关的校园传感设备与服务采集、接入到数据治理平台通过张量统一表示方法表示为不同阶和维度的子张量模型，并在此基础上对这些多源子张量先进行张量链分解得到张量链形式的数据，再在张量链数据的基础上实现张量融合操作，最终构造得到高阶张量空间对校园大数据进行统一表示。

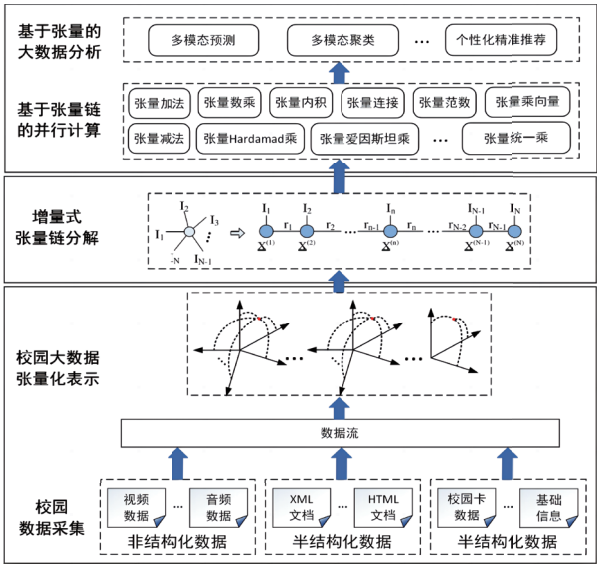

图 1 基于张量的校园大数据表示与处理框架

3.1 基于张量的校园大数据统一表示

3.1.1 张量定义

张量可以理解为一个多维数组。其中一阶张量即是向量，

可以用一个一维数组进行表示；二阶张量即是矩阵，可以使用一个二维数组进行表示。同理，一个 N 阶张量 $\underline{\boldsymbol{X}}$ 可以用一个 N 维数组进行表示：

$$\underline{\boldsymbol{X}} \in \mathbf{R}^{I_1 \times I_2 \times \cdots \times I_N} \tag{1}$$

式中：N 表示张量的阶数；$I_n (0 < n \leq N)$ 表示张量 $\underline{\boldsymbol{X}}$ 在第 N 阶上的维度。如图 2 所示，$\underline{\boldsymbol{T}} \in \mathbf{R}^{2\times3\times2}$ 表示一个三阶张量，其第 1 阶上的维度为 2；第 2 阶上的维度为 3；第 3 阶上的维度为 2。其中 $\underline{\boldsymbol{T}}(1,2,2)$ 用来表示张量 $\underline{\boldsymbol{T}}$ 中的元素，即图 2 中的 t_{122}。

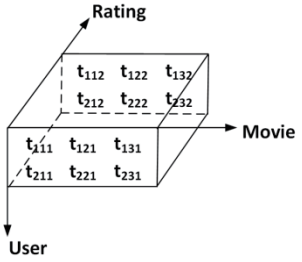

图 2 三阶张量 $\underline{\boldsymbol{T}} \in \mathbf{R}^{2\times3\times2}$

3.1.2 校园大数据张量化表示

在构建校园大数据的张量化表示时，首先借助特定的数据采集设备和服务，从校园不同部门获取非结构化数据（例如音视频文件）、半结构化数据（如 XML 文档）以及结构化数据（如校园卡交易记录、师生基础信息库等），如表 1 所示，以校园内行为数据表为例。这些数据在收集过程中以数据流的形式进行传输，并通过不改变原始数据格式的方式递交至上层处理系统。

表 1 校园行为数据表

数据类型	子数据	意义
静态数据	教师基本信息	工号、姓名、性别、年龄、专业、教研组、学部等
	学生基本信息	学号、姓名、性别、年龄、专业、学部、学制等
	课程基本信息	程代码、课程名称、课程学年等
	图书基本信息	书籍号、类型、名称、校区、书架号等
	宿舍基本信息	校区、楼号、寝室号、床号、所住人员学工号等
	教师授课数据	教学计划编号、教师工号、教师姓名、上课时间地点、上课课时等
	教学计划数据	教学计划编号、学年、学期、课程代码、上课时间地点、主讲教师工号、姓名等
动态数据	图书馆借阅记录	学工号、已借图书书籍号、借阅时间戳等
	校园卡消费记录	学工号、卡号、交易时间、交易地点、交易类型等
	进出宿舍记录	学工号、校区、宿舍楼、进出宿舍门、进出时间戳等
	进出校门记录	学工号、校区、进出校门、进出时间戳等
	网络行为记录	无线 AP 上下线时间地点、上网期间访问站点详情等

在上述基础上，采用张量分析技术，将收集到的多源异构数据转化为低阶子张量。并通过对张量拓展算子的应用，实现不同数据特征的融合，将各自独特的属性映射到张量空间的不同维度（或称为阶），这一过程确保了数据的完整性和多维性得以保留。最后，将融合后的低阶子张量嵌入到一个三阶基础张量空间中。三阶张量提供了一个足够复杂的框架，用于描述数据的多个方面及其相互关联。通过这种方法，构建出一个高阶张量空间，提供一个统一的表示模型用于有效地表示、分析和挖掘校园大数据中的复杂模式和关系。

3.2 基于张量链分解的张量融合

3.2.1 张量链分解定义

张量链分解（tensor train decomposition，TTD）是指张量通过连续对不同的展开形式进行 SVD 分解，得到由多个核心张量组成的张量表示形式的一种分解方式。具体定义如下，给定一个张量 $\underline{\boldsymbol{X}} \in \mathbf{R}^{I_1 \times I_2 \times \cdots \times I_N}$，TTD 将该张量分解为 N 个低阶张量。其分解形式为：

$$\underline{\boldsymbol{X}} = \underline{\boldsymbol{X}}^{(1)} \cdot \underline{\boldsymbol{X}}^{(2)} \cdot \ldots \cdot \underline{\boldsymbol{X}}^{(n)} \cdot \ldots \cdot \underline{\boldsymbol{X}}^{(N)} \tag{2}$$

式中：· 表示张量的缩并操作，这里指张量的单模乘。

核心张量为：

$$\underline{\boldsymbol{X}}^{(n)} \in \mathbf{R}^{r_{n-1} \times I_n \times r_n} \ (n \in [1, N], r_0 = r_N = 1) \tag{3}$$

式中：$\{r_0, r_1, \cdots, r_N\}$称为张量的 TT 秩。

3.2.2 基于张量链分解的张量融合

在基于张量的大数据统一表示框架中，对于多源异构的校园大数据，需构造高阶张量空间对校园大数据进行统一表示。但发现，对多个高阶张量进行张量融合操作时，由于张量高阶高维的特点，基于张量的融合操作会极大增加了算法的时间、空间复杂度。解决上述问题关键在于，在完成对多个张量融合的同时，避开融合过程中对高阶张量的处理操作。本文给出基于张量链分解的张量融合方案，多个高阶张量先分别进行张量链分解，以张量链形式（多组低阶的核心张量）进行分布式存储，并在张量链形式下，完成对张量数据的融合。

如图 3 所示，其中上半部分表示基于张量的张量融合，下半部分表示基于张量链分解的张量融合，两者都可以对多个张量数据进行融合，但其过程却存在差异。基于张量的张量融合过程可以解释为，给定 N 阶张量 $\underline{\boldsymbol{T}_1} \in \mathbf{R}^{I_1 \times I_2 \times \cdots \times I_n \times \cdots \times I_N}$，当另一个 Q-P+1 阶张量 $\underline{\boldsymbol{T}_n} \in \mathbf{R}^{I_P \times I_{P+1} \times \cdots \times I_q \times \cdots \times I_Q}$，沿 $I_n = I_q$ 阶融合到张量 $\boldsymbol{T}_1$ 中时，先将两者进行张量融合操作，得到融合后 Q-P+N 阶张量：

$$\underline{\boldsymbol{T}} \in \mathbf{R}^{I_1 \times I_2 \times \cdots \times I_{n-1} \times I_P \times I_{P+1} \times \cdots \times I_n \times \cdots \times I_N \times I_{q+1} \times \cdots \times I_Q} \tag{4}$$

再对张量 $\underline{\boldsymbol{T}}$ 进行张量链分解，得到最终结果：

$$\underline{\boldsymbol{T}}^{(1)}, \underline{\boldsymbol{T}}^{(2)}, \cdots, \underline{\boldsymbol{T}}^{(n-1)}, \underline{\boldsymbol{T}}^{(P)}, \underline{\boldsymbol{T}}^{(P+1)}, \cdots, \underline{\boldsymbol{T}}^{(n)}, \cdots, \underline{\boldsymbol{T}}^{(N)}, \underline{\boldsymbol{T}}^{(q+1)}, \cdots, \underline{\boldsymbol{T}}^{(Q)} \tag{5}$$

而基于张量链分解的张量融合过程则是先分别对两个张

量进行张量链分解，得到张量链形式的结果$\underline{\boldsymbol{T}}_1^{(1)},\underline{\boldsymbol{T}}_1^{(2)},\cdots,\underline{\boldsymbol{T}}_1^{(N)}$和$\boldsymbol{T}_n^{(P)},\boldsymbol{T}_n^{(P+1)},\cdots,\boldsymbol{T}_n^{(Q)}$，再基于张量链形式数据进行张量融合操作，从而得到最终结果为：

$$\underline{\boldsymbol{T}}^{(1)},\underline{\boldsymbol{T}}^{(2)},\cdots,\underline{\boldsymbol{T}}^{(n-1)},\underline{\boldsymbol{T}}^{(P)},\underline{\boldsymbol{T}}^{(P+1)},\cdots,\underline{\boldsymbol{T}}^{(n)},\cdots,\underline{\boldsymbol{T}}^{(N)},\underline{\boldsymbol{T}}^{(q+1)},\cdots,\underline{\boldsymbol{T}}^{(Q)} \quad (6)$$

两种方案返回的结果一致，但基于张量链分解的张量融合在时间、空间复杂度上要远低于直接基于张量进行张量融合。

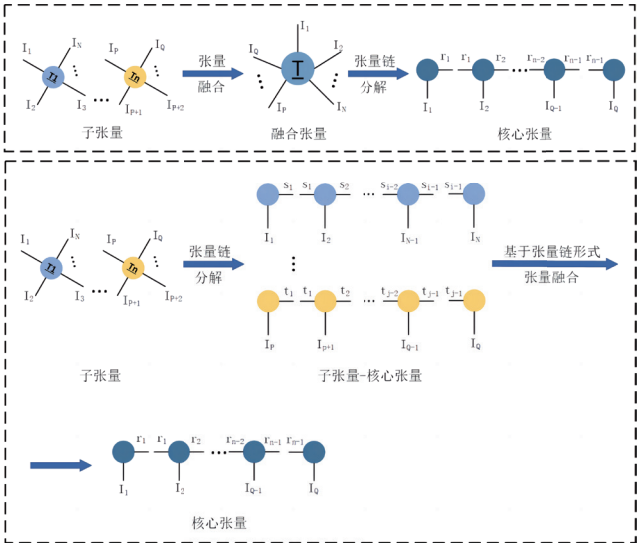

图 3 基于张量链分解的张量融合

4 基于张量链分解的高维数据多维关联分析模型（以学生学习资源精准推荐为例）

4.1 基于张量链分解的高维数据多维关联分析模型

传统的推荐算法一般基于二维矩阵，然而，当个性化资源推荐系统从矩阵模型（用户 - 资源 - 评分）发展为更高维模型（用户 - 用户特征 - 资源 - 资源特征 - 评分）时，在矩阵基础上的操作显然无法直接满足推荐的要求。通常，传统推荐算法将高维模型拆分为多个二维模型，再分别对其进行分析和推荐。虽然在处理后，也能有一定的推荐效果，但是对高维模型的拆分无疑会导致各影响因素之间特征结构和关联关系的缺失，故而推荐性能也会受到一定影响。

张量模型使高维空间中信息的完整性得以保留，在基于张量链分解的多维关联分析过程中，在高维空间从多个维度更整体地对数据进行关联分析，进而挖掘影响因素之间的潜在语义关联，达到更好的推荐效果。

如图 4 上半部分所示，基于张量链分解的多维关联分析的整体过程分为两步：第一步，对原始张量进行张量链分解，得到对应数量的核心张量；第二步，将上述各核心张量执行第一步的逆过程，进行张量重构，得到与原始张量同阶同维的近似张量。原始张量进行张量链分解，最后得到对应原始张量中每一阶的核心张量，可以看出，张量链分解的过程是对原始张量中每一阶（每一特征属性）都进行了一次奇异值的分解，从而能达到多维关联分析的效果。奇异值分解的过程如图 4 下半部分所示，对矩阵进行奇异值分解，可以得到左奇异值矩阵 $\boldsymbol{U}_n$、奇异值矩阵 $\boldsymbol{E}$ 与右奇异值矩阵 $\boldsymbol{V}_n$。其中，奇异值矩阵 E 对角线上的值为奇异值，其表示特征矩阵中各特征的重要程度，并据此从大到小排列。在实际应用中，如图 4 下半部分所示，可以根据实际需要截取排列在前的一部分 (r_n) 奇异值，将较小的奇异值及其对应的特征舍去，以此达到去除噪声数据的效果。因此，基于张量链分解的多维关联分析，是在高维空间从多维度整体对数据进行关联分析，并且分解过程中还去除了噪声数据，有利于在复杂情境下提高数据分析的效果，实现更精准的推荐。

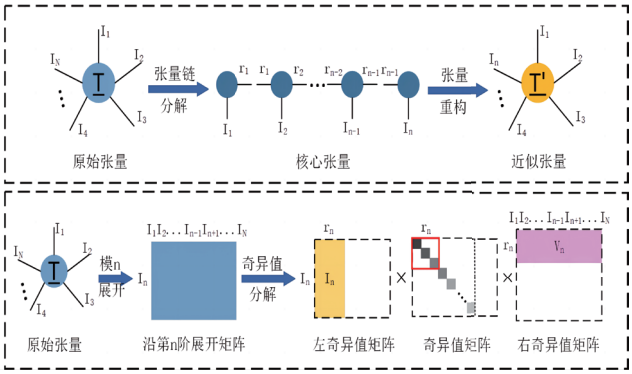

图 4 基于张量链分解的多维关联分析

4.2 基于张量链分解的学生学习资源精准推荐

基于张量链分解的学习资源推荐模型如图 5 所示，以职业院校为例，主要包含以下步骤：

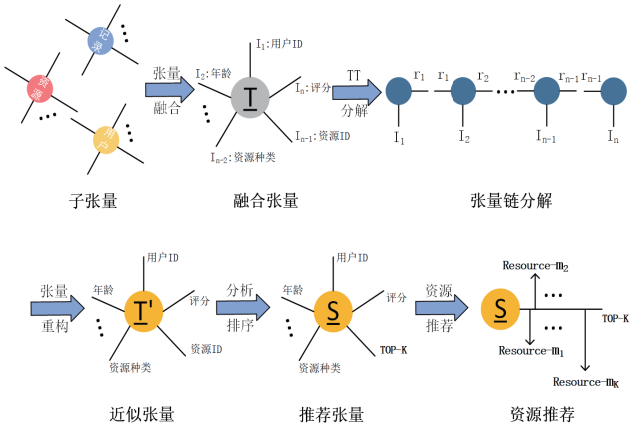

图 5 基于张量链分解的学习资源推荐流程

第一步：构建张量。首先根据学生、学习资源以及学习记录等数据，将其构建成相应的子张量，如图 6 所示；然后，通过张量融合，将各子张量融合成一个全局张量，由学生相关阶（例如年龄、性别、学习风格）、学习资源相关阶（例如难度、类型、发布时间）以及时间、地点、设备、对资源评分等组成，如表 2 所示。

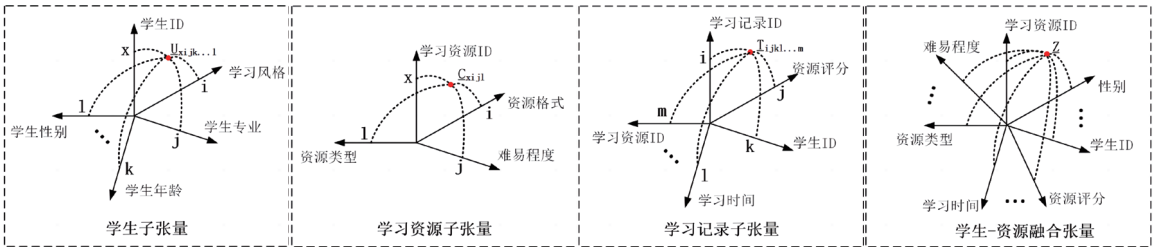

图 6 学生、学习资源、学习记录子张量及“学生 - 资源”融合张量

表 2 推荐模型子张量属性与取值

子张量	属性	取值范围	意义
学生	性别	1, 2	男、女
	年龄	1, 2, 3	一年级、二年级、三年级等
	学制	1, 2	三年制、五年制
	学习专业	1, 2, 3, 4, …	计算机、汽修、服装、美术等
	学习风格	1, 2, 3, 4	具体序列、具体随机、抽象序列、抽象随机
学习资源	资源类型	1, 2, 3, 4, …	课件、教案、教学视频、试题等
	媒体格式	1, 2, 3, 4, …	文本、图形图像、音频、视频等
	难易程度	1, 2, 3, 4, 5	难、较难、中、较易、易
学习记录	学习时间	1, 2, 3, 4, 5	6~18 点每 3 h 一个时段，18~24 点为一个时段
	学习设备	1, 2, 3	iOS、Android、Windows
	学习地点	1, 2, 3	学校、家庭、其他
	资源评分	1, 2, 3, 4, 5	好、较好、中、较差、差

第二步：关联分析。将张量融合后的全局张量进行张量链分解，再将分解后得到 N 个核心张量进行张量重构，得到近似张量。

第三步：资源推荐。得到近似张量后，规定张量 rating 阶中 4、5 两个维度所代表的评分表示学生对该学习资源为较高评价（受到学生的认同）。同时，根据所需推荐资源的学生相关信息，在近似张量中，保持学生相关信息所对应的张量阶为所给定的固定值，保持张量 rating 阶取值为 4、5 两个维度，截取近似张量的子张量。最后，将该子张量上对应 resource 阶的填充值进行排序，选取前 TOP-K 个学习资源为对应的学生进行推荐。

5 结语

各级院校校园治理由“数字校园”向“智慧校园”转型的过程中，关键在于对数据挖掘和智能研发部分内容的构建，其中对多源异构数据在高维张量空间中的统一的表示以及基于高维数据的多维关联分析将更有效地进行分析与挖掘校园大数据中的复杂模式和关系。由此，本文提出了基于张量的智慧校园服务体系，并通过构建基于张量的大数据表示与融合模型、基于张量链分解的高维数据多维关联分析模型，全面提升各级院校大数据整合治理与多维分析能力，推动校园治理现代化。

参考文献：

[1] 刘革平，钟剑，谢涛．基于流程驱动的高校智慧校园基础架构研究与实践 [J]. 中国电化教育，2019(4):23-28+35.

[2] 周炜．大数据视域下高校数据治理优化路径研究 [J]. 教育发展研究，2021, 41(9):78-84.

[3] 张辉，李健明，杨强．大数据视角下高校数据治理体系研究与实践 [J]. 中国高等教育，2022(Z2):16-18.

[4] 黄贤明，梁爱南，张汉君，等．教育信息化 2.0 背景下基于数据中台的高校数据治理方案研究 [J]. 现代信息科技，2022, 6(18):24-27.

[5] 魏建行，刘远志，罗超，等．基于数据中台的高校数据治理体系研究 [J]. 信息技术与信息化，2022(6):98-101.

[6] 毛文卉，吴驰，刘雅琴，等．数据治理背景下高校数据共享框架的研究与实践 [J]. 实验室研究与探索，2022, 41(8): 297-303.

[7] 李子昕，陈晋．基于数据中台的高校科学数据管理服务平台建设 [J]. 大学图书情报学刊，2021,39(2):56-62.

[8] 屠佳琪，王冬梅，高焕江，等．智慧校园背景下高校大数据服务体系的研究 [J]. 现代电子技术，2023, 46(20):76-80.

[9] 李有增，曾浩．基于学生行为分析模型的高校智慧校园教育大数据应用研究 [J]. 中国电化教育，2018(7):33-38.

[10] 邓嘉明．智慧校园学生数据画像生成方式研究 [J]. 现代电子技术，2019,42(21):58-62.

[11] 匡立伟．基于张量的大数据统一表示及降维方法研究 [D]. 武汉：华中科技大学，2017.

[12] OSELEDETS I V. Tensor-train decomposition[J]. SIAM journal on scientific computing, 2011, 33(5): 2295-2317.

[13] 刘华中．基于张量的大数据高效计算及多模态分析方法研究 [D]. 武汉：华中科技大学，2018.

【作者简介】

高熠徐（1999—），男，浙江绍兴人，硕士研究生，研究方向：教育大数据分析。

（收稿日期：2024-08-22）