面向可拓建筑设计的数据准备流程与方法

刘书宇,邹广天,周舟,肖俊龙

哈尔滨工业大学建筑学院;哈尔滨工业大学建筑计划与设计研究所,哈尔滨 150006

摘要　为面向可拓建筑设计进行可拓数据挖掘,以可拓建筑设计数据为对象,探讨将其转化为高质量的结构数据的流程与方法。根据跨行业数据挖掘标准流程(CRISP-DM),建立包括基元化表达、数据表设计、数据筛选、数据形式变换、变量标准化处理和变量维数约简6个步骤的数据准备流程,并根据建筑学专业特点,结合可拓学、几何学、统计学理论,设计各步骤的操作方法,构建出完整的数据准备流程与方法。案例检验结果表明,按照该流程及其操作方法,可有效地将可拓建筑设计数据转化为统一格式、高信度、量化且可运算的结构数据。

关键词　可拓建筑设计;数据准备;结构数据

中图分类号　TU18　　**文献标志码**　A　　**doi**　10.3981/j.issn.1000-7857.2014.36.005

Process and Methods of Data Preparation for Extension Architecture Design

LIU Shuyu, ZOU Guangtian, ZHOU Zhou, XIAO Junlong

School of Architecture, Harbin Institute of Technology; Architectural Planning and Design Institute, Harbin Institute of Technology, Harbin 150006, China

Abstract　The process and method of transforming extension architecture design data into structured data are investigated in order to execute extension data mining on the extension architecture design. According to the cross-industry standard process for data mining (CRISP-DM), a 6-stage data preparation process is built, which includes expression in basic-element form, data sheet design, data filtering, data form transform, standardization of variables, and variable dimension reduction. According to the characteristic of architecture, the methods for each stage are established by referring to extenics, geometry, statistics, etc. Finally, an integral data preparation system of the process and methods is constructed.

Keywords　extension architecture design; data preparation; structured data

在可拓建筑设计的过程中,当需要通过可拓变换解决矛盾、实现创新时,借助可拓分类挖掘对于正质变知识的发现[1],可以为变换提供科学的依据,进而保证设计的合理性。但由于可拓建筑设计行为本身较为复杂,需要考虑众多不同种类的要素,其记录数据的格式以及数据表达自身所蕴含信息的形式等各不相同,而现有的可拓数据挖掘方法主要针对量化程度较高的数据,两者无法直接衔接。因此,在进行可拓数据挖掘操作前,需要对可拓建筑设计相关的各方面要素进行数据准备,将其转化为可以进行可拓数据挖掘操作的数据。

就目前的数据挖掘技术而言,只能针对结构化和半结构化数据进行操作,无法对非结构化数据进行操作。其中,由

收稿日期:2014-10-10;修回日期:2014-11-30

基金项目:国家自然科学基金项目(51178132)

作者简介:刘书宇,博士研究生,研究方向为可拓建筑学,电子信箱:liushuyu0216@163.com;邹广天(通信作者),教授,研究方向为建筑计划学、可拓建筑学、建筑设计创新学、环境行为心理学等,电子信箱:zougt@hit.edu.cn

引用格式:刘书宇, 邹广天, 周舟, 等. 面向可拓建筑设计的数据准备流程与方法[J]. 科技导报, 2014, 32(36): 37-42.

于对半结构化数据的数据挖掘技术体系尚处于起步阶段,适用的广度与深度较低,目前主要针对化学、材料、通信等逻辑性极强的学科。因此,在进行面向可拓建筑设计的可拓数据挖掘时,应先通过数据准备将可拓建筑设计数据转化为结构数据。本文探讨一种面向可拓建筑设计的数据准备流程及其操作方法。

1 数据准备流程与方法

在一个完整的数据挖掘流程中,根据SPSS、NCR和Daimler Chrysler3家公司共同于1999年提出的跨行业数据挖掘标准流程(CRISP-DM),应该由业务理解、数据理解、数据准备、模型建立、模型评估、结果部署6个连续步骤组成[2]。数据准备是进行具体挖掘操作前最重要的步骤,也是整个数据挖掘流程中最关键的部分[3]。通过业务理解、数据理解确定挖掘目标及数据来源后,对各种来源的低质量非纯数据进行一系列的集成、清洗、变换、约简等操作,提高数据质量,并使其成为可供挖掘操作的形式[4]。

根据对数据准备及建筑学专业特点的分析,将数据准备过程分解为如图1所示的步骤:建筑设计数据基元化表达、建筑要素数据表设计、数据筛选、基于数据属性类别的数据形式变换、变量标准化处理和变量维数约简。

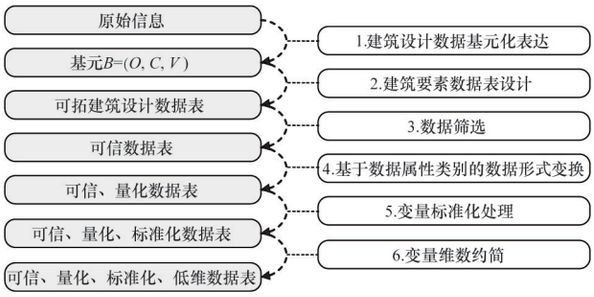

图1 数据准备流程

Fig. 1 Process of data preparation

1.1 建筑设计数据基元化表达

在建筑设计中,建筑数据通常以描述性的文本或图纸等记录格式出现。而在可拓建筑设计中,引入了基元这一概念作为建筑设计数据的基本记录格式。

1.1.1 单一元

对于各项基本的建筑要素,引入可拓学中单一元的概念,分别以物元(M)、事元(A)、关系元(R)3种单一元表示建筑设计中的物要素、事要素、关系要素。在将建筑数据记录为单一元格式的过程中,需要对对象和特征两者的命名原则进行约束,以保证之后数据筛选的顺利进行。

1)所有对象、特征的命名应依据《民用建筑设计术语标准(GB/T50504—2009)》中具有明确定义的术语进行。如大厅、大堂等应统一命名为门厅。

2)当需要进行命名的对象、特征在《民用建筑设计术语标准》中没有与之严格对应的标准术语时,应根据业务理解所确定的业务目标,依据《民用建筑分类标准(GB50352—2005)》、建筑设计通用标准和相应建筑设计专用标准中具有明确定义或规定的术语进行。如住宅区应严格按照户数与人口命名为居住区、住区、组团。

3)当需要命名的对象、特征无法依据上述两点进行命名时,应依据行业通用表述命名。

1.1.2 专业元

在单一元的基础上,引入可拓学中的复合元概念,结合建筑设计本身特点,定义4种特殊的复合元作为可拓建筑设计专业元[5],以从4个角度记录建筑设计中的各个系统。专业元可以独立表达,其对象与量值也可以由若干单一元组合形成。

从物质性出发,建筑包含实部和虚部两部分。其中,建筑本身可以看作是实部,而建筑的美感、精神则可以看作虚部。根据建筑设计三要素"坚固、实用、美观",可分别以建筑实体、由实体围合成的空间、由实体及空间共同产生的建筑美感和精神与其进行对应,其中前两者共同构成了建筑本身。由此构建建筑元(M_A)以表示建筑的实体、空间和精神,其中,空间特征包括建筑的功能、流线等,实体特征包括建筑的形态、结构、肌理等,精神特征包括建筑的审美感受、意义等。建筑元记作:

$$M_A=\begin{bmatrix}\text{建筑}, & \text{空间特征}, & v_1\\ & \text{实体特征}, & v_2\\ & \text{精神特征}, & v_3\end{bmatrix}$$

根据建筑设计中"形式追随功能"的原则,建筑的功能是其第一属性,而建筑空间则是这一属性的物质基础。从动态性出发,空间包含显部和潜部两部分。在建筑空间中,其针对性的空间形式特点对于使用目的的满足是显而易见的,可以看作是显部。而基于空间特点对使用者产生的空间感受、心理暗示则较难被意识到,可以看作是潜部。由此构建空间元(M_S)以表示空间的目的、感觉和心理,其中,目的属性指空间对使用功能的满足,感觉属性包括视觉效果、空间对行为主体生理需求的适应等,心理属性包括空间的环境行为特性、文化特性等。空间元记作:

$$M_S=\begin{bmatrix}\text{空间}, & \text{目的属性}, & v_1\\ & \text{感觉属性}, & v_2\\ & \text{心理属性}, & v_3\end{bmatrix}$$

在建筑设计中,每一个构件都可以看作是若干子构件以特定形式的组合。从系统性出发,建筑包含硬部与软部两部分,其中每一个组成更高一级构件的子构件可看作硬部,而其特定的组合形式可看作软部。由此构建形式元(R_F)以表示建筑各要素间的关系(如联系方式、从属关系等),其中,前项和后项指门窗、雨棚等具象造型元素或点、线、面、体等抽象造型元素。形式元记作:

$$R_F=\begin{bmatrix}\text{形式关系}, & \text{前项}, & v_1\\ & \text{后项}, & v_2\\ & \text{程度}, & v_3\\ & \text{联系方式}, & v_4\\ & \text{从属关系}, & v_5\\ & \vdots & \vdots\end{bmatrix}$$

建筑设计的最根本目的,是满足人的使用需求,人在其中进行的各种行为即是这种需求的体现。从对立性出发,建筑包含正部与负部两部分。建筑在使用中,当功能空间能够较好地满足人的行为时,可将其看作正部;若功能空间由于某原因无法满足人的行为需求或对人的行为带来了负面影响时,则可将其看作负部。由此构建行为元(A_B)以表示使用者在建筑中进行的行为。行为元记作:

$$A_B=\begin{bmatrix}\text{行为}, & \text{行为主体}, & v_1\\ & \text{行为空间}, & v_2\\ & \text{需求一}, & v_3\\ & \text{需求二}, & v_4\\ & \vdots & \vdots\end{bmatrix}$$

综上所述,通过将建筑设计相关要素转化为物元(M)、事元(A)、关系元(R)3种可拓建筑设计单一元和建筑元(M_A)、空间元(M_S)、形式元(R_F)、行为元(A_B)4种可拓建筑设计专业元,可以将建筑设计数据记录为统一的基本格式,实现与各类建筑设计源数据的对接。

1.2　建筑要素数据表设计

在结构化数据中,关系型数据库是其最广泛的存在形式,其逻辑结构主要为集合结构和线性结构,具体体现为二维数据表。根据上节所述,可将建筑数据转化为统一的基元格式:

$$B=(O,C,V)$$

由于在进行数据挖掘算法编写时,多维度(表头项目较多,在此体现为多特征)的宽表相对于若干个少维度(表头项目较少)的窄表的集合更加利于取数统计,因此结合建筑基元的格式特点,构造基础的多特征可拓建筑设计二维数据总表如表1所示。

表1 可拓建筑设计数据总表

Table 1　Primary table of extension architecture design data

楼梯	梯段数	踏步数	踏步高/mm	踏步宽/mm	…	C_n
O_1	2	30	150	300	…	V_{1n}
O_2	1	18	135	320	…	V_{2n}
O_3	3	48	160	280	…	V_{3n}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
O_m	V_{m1}	V_{m2}	V_{m3}	V_{m4}	…	V_{mn}

由于基元具有发散性,通过对基元的发散,可以在总表的框架下,根据“一对象多特征”、“一特征多对象”、“同对象同特征多值”等发散特点[6],构造可拓建筑设计数据子表。在子表中,通过O、C、$O\wedge C$作为通道与总表链接。

在可拓建筑设计中,“一对象多特征”体现为一个主体具有多方面属性,如“卧室”这一对象具有“面积”、“净高”、“朝向”等特征。由此,可构建I类子表如表2所示,且由于某些对象并不具有某些特征,如“办公室”这一对象不具有“地面排水找坡”这一特征,故此子表中的特征数y必然小于总表中大量对象的特征数n。

表2 I类子表

Table 2　Secondary table I

对象	面积/m²	净高/m	朝向	窗地比	…	C_y
卧室	15.12	2.8	正南	0.24	…	V_{1y}

注:$y<n$。

“一特征多对象”特点在可拓建筑设计中体现为若干主体都具有某一同样属性,如“起居室”、“餐厅”、“阅览室”等对象都具有“窗地比”这一特征。由此,可构建II类子表如表3所示,且由于某些特征并不被某些对象所具有,如“消防楼梯间数”这一特征不被“多层住宅”这一对象具有,故此子表中的对象数x必然小于总表中的对象数m。

表3 II类子表

Table 3　Secondary table II

特征	起居室	餐厅	厨房	卫生间	…	O_x
窗地比	0.28	0.22	0.18	0.14	…	V_{x1}

注:$x<m$。

“同对象同特征多值”特点体现为某一对象的某一属性具有多个可选择的数值,如剪力墙结构的厚度模数为200、300、400 mm;或某一主体的某一属性可以具有一个变化的区间值,如“某居住区”这一对象的特征“规划容积率”的值为[2.8,3.0]。由此,可构建III类子表,如表4所示。

表4 III类子表

Table 4　Secondary table III

对象	厚度/mm
剪力墙	200
	300
	400

通过以上可拓建筑设计数据总表和3种可拓建筑设计数据子表,可将经过基元化的建筑数据转化为二维表格式的结构数据,该数据为集合结构,从而使其成为符合可拓数据挖掘操作要求的格式。

1.3　数据筛选

在得到可拓建筑设计结构数据后,需对数据进行筛选以提高数据质量。

1.3.1　脏数据清洗

由于结构化的可拓建筑设计数据是由最初的多种格式

的异源数据转化而来,而各异源数据多是由人工进行总结和记录,不可避免地会出现残缺与错误。同时,由于数据来自于多源的结合,对于同一主体会有若干不完全重叠的数据来源,导致部分数据重复出现。因此,需要对这些数据分别加以处理。

1)残缺数据。由于可拓建筑设计数据表中的项与内容均来自于可拓建筑设计基元的转化,因此数据的残缺必然出现在对象O、特征C和量值V上,由源数据缺失和基元数据缺失两种原因造成。因此,需要与源数据和基元数据进行对照。若缺失发生在源数据中,则将由此转化的基元删除,并在数据表中删除与其对应的O、C、V。若缺失发生在基元数据中,则对其来源数据重新进行基元化处理,再将新基元数据导入数据表。

2)错误数据。在数据挖掘中,源数据的信度在数据理解阶段进行检验,数据准备阶段所处理的数据默认为是可信的。因此,数据的错误在此体现为O、C、V的不匹配。但由于建筑源数据多为描述性数据,在基元化处理时容易对其产生错误理解,由此可能产生O、C、V不匹配的错误。对此类错误的发现,需要较专业的建筑学知识,如发现某量值的单位无法与其特征对应,或某特征不应被某对象所具有等。通过与原始数据进行对照,可对此类错误进行修正。

3)重复数据。由于在源数据基元化的过程中,对单一元中对象、特征的命名原则进行了规定,对复合元中对象、特征进行了具体命名,因此在此阶段可以有效地发现来自于不同源数据的重复数据。对重复数据的处理应遵循以下原则:

若重复数据的特征与量值完全相同,则只保留一条数据,将其余删除;

若重复数据有一对"特征-量值"不同,由于无法考证其真实性,故将其全部删除;

若重复数据有两对及以上"特征-量值"不同,且来源数据的产生时间不同,则其有可能是由传导效应造成的,故将其全部保留。

1.3.2 数据孤立点处理

在一般的数据挖掘数据准备过程中,一般将数据孤立点提出以进行单独分析。但可拓建筑设计是一种解决建筑设计中矛盾、质量、创新问题的设计方法,而创新建筑往往与普通建筑有较大差别。因此,当满足以下两个条件之一时,应将孤立数据予以保留:

1)在业务理解阶段所确定的业务目标是解决创新问题;

2)在数据理解阶段所选取的源数据是有关创新建筑的。

若不能满足以上任一条件,则应将孤立点删除。

1.4 基于数据属性类别的数据形式变换

至此,数据表中的可拓建筑设计数据均为可信度较高的数据。然而,不同于传统的数据挖掘应用领域,这些数据表达自身所蕴含信息的形式复杂多样,除已经以量化形式表示的数据外,还有很多非量化数据,如建筑的形体、人对建筑的感受等,常用文字、代号、甚至图形表示,无法直接对其进行挖掘操作。因此,需要对此类形式复杂建筑数据进行量化,使其具有统一的表达所蕴含信息的形式。经分析,此类建筑数据可根据其形式分为三类具有较大类间差别和较小类内差别的类别,即形体数据、性能数据和感知数据,需分别进行量化。

1.4.1 形体数据

形体数据指建筑空间、实体的几何形状及其组合关系,是对建筑物质形体的真实记录,其包含体积、形状、关系三方面要素。由于建筑的体型通常以长方体为基础,因此,对于大多数建筑可以将其近似地看作若干个长方体的组合。在一个三维直角坐标系中,一个长方体的对角线即包含了其体积、形状、位置信息,各对角线所包含的位置信息即组成了空间、实体间的关系信息。因此,通过一个建筑的近似分解长方体的对角线集合,即可表示该建筑的基本形体数据。据此将形体数据的量化分为4个步骤:

1)将目标建筑附近的任意一点设定为三维直角坐标系原点。

2)将建筑形体近似地分解为n个长方体,根据设定的原点计算出每个长方体的任一对角线的端点坐标(x_{m1},y_{m1},z_{m1}),(x_{m2},y_{m2},z_{m2}),其中$m\in[1,n]\cap Z$,Z是整数集。

3)将单个长方体对角线写为式(1)形式:

$$\frac{x-x_1}{x_2-x_1}=\frac{y-y_1}{y_2-y_1}=\frac{z-z_1}{z_2-z_1}\quad(1)$$
$$x\in[x_1,x_2]\quad y\in[y_1,y_2]\quad z\in[z_1,z_2]$$

4)将各对角线函数结合为一函数集合作为形态数据的量化形式。

对于上述过程,当用于量化建筑整体形态时,其精度应精确到1 m;量化建筑内部空间形态时,应精确到0.1 m;量化建筑装饰等细部形态时,应精确到0.01 m。对于少量无法分解为长方体的形态数据,可按上述思路在圆柱坐标系或球坐标系中将其量化为函数形式,再代入坐标系转化系数转化为三维直角坐标系中的函数。

1.4.2 性能数据

性能数据指建筑物的性能指标,是完全由建筑的各方面物质因素决定的,如通透度、疏散效率等。对此类数据的量化,可以通过定义相关的已量化数据的乘积实现。如围合度,可以立面实体面积比立面面积表示,记作"围合度=实体面积×立面面积$^{-1}$",无单位;如采光效率,可以光通量比采光时间来表示,记作"采光效率=光通量×采光时间$^{-1}$",单位为$lm\cdot s^{-1}$。

1.4.3 感知数据

感知数据指人对建筑产生的感受,是由建筑的物质因素和人的精神因素共同决定的,如庄严感、亲切感等。对此类数据的量化,可以借助区间估计结合视觉类比实现[7]。其步骤为:

1)将要量化的感知数据以建筑行业习惯分为5个等级:极弱、弱、一般、强、极强。

2）请不少于10人在一定长的线段上分别对5个等级勾画范围，如图2所示。

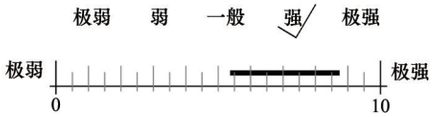

图2　区间估计视觉类比示意

Fig. 2　Schematic of interval estimation and vision analogy

3）将该线段20等分，每0.5分为一格，共0~10分。

4）对每个等级进行统计，将所有对该等级的勾画范围所覆盖点的值相加，除以所有对该等级的勾画范围所覆盖点的数量，得到每个等级的感知数据等距值。

5）将数据表中的描述性感知数据代入，转化为量化感知数据。

1.5　变量标准化处理

经过上述过程，得到的变量具有不同的量纲与值域，所体现的意义也不尽相同，若直接对这些变量进行挖掘操作，难以对所得结果的意义进行解释，进而影响利用。因此，需对其进行标准化处理，消除由量纲和值域的差异带来的影响。引入两种常用的标准化方法：

1）min-max标准化[8]。当所需量化的数据有明确、一致的值域时，按式(2)将每个要标准化的数据x减去值域的最小值，再除以极差，得到标准化数据x'，数据范围为[0,1]。

$$x'=\frac{x-x_{\min}}{x_{\max}-x_{\min}} \tag{2}$$

2）z-score标准化[9]。当所需量化的数据没有明确、一致的值域时，按式(3)将每个要标准化的数据x减去该特征下所有量值的均值，再除以标准差，得到标准化数据x'。得到的标准化数据围绕0上下浮动，大于0表示高于平均水平，小于0表示低于平均水平。

$$x'=\frac{x-\bar{x}}{s}\qquad s=\sqrt{\frac{\sum_{i=1}^{n}(x-\bar{x})^2}{n-1}} \tag{3}$$

经过标准化处理后，不同特征的量值数据可以共同运算，执行挖掘操作。

1.6　变量维数约简

在经过上述步骤之后，若要进行挖掘操作，还需确保运算量在可实现的范围内。为便于数据挖掘算法编写，将数据表设计为多维的宽表。但随变量的维数增加，对其挖掘需要进行的运算量会随之以指数增长，导致运算无法进行。因此，需要对变量维数进行约简，减少参与运算的特征数，以确保挖掘操作可以实现。维数的约简主要针对两类目标：

1）与业务目标关联较弱的特征。根据建筑学专业知识，可以发现并去除此类特征。如当业务目标是预测建筑疏散效率时，楼层净高、窗口面积等特征即可去除。

2）相互间关联较强的特征。由于特征间具有较强传导效应，对其中之一进行挖掘即可达到对业务目标的预测，不必全部代入运算。为发现并去除此类特征，可以引入Pearson相关系数公式[10]，按式(4)计算其相关系数r，即

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{4}$$

当$|r|\in(0.95,1]$时，两特征具有极强关联，去除其中之一；当$|r|\in(0.90,0.95]$时，两特征具有较强关联，视情况去除或保留某一特征；当$|r|\in[0,0.90]$时，两特征关联不强，均应保留。

2　案例检验

为对上述流程与方法的可行性、有效性进行检验，以图3所示的某经济适用住房小区建筑外部空间形态数据为例，应用上述流程与方法对其进行数据准备。该小区设计于2011年，各建筑的具体尺寸已知，其中7栋高层住宅建筑在外部空间上形成小区主体，其余配套公建及附属设施体量较小，且高度较为低矮，对整体外部空间影响较小，因此仅以7栋住宅建筑为检验对象。

图3　某经济适用住房小区

Fig. 3　Airscape of an affordable housing community

首先，在建筑设计中，性能数据是较能直观地体现建筑空间形态特点的数据类型，因此将部分性能数据以空间元的格式予以记录。以1#楼为例，其自身为对象，将其各项性能属性作为特征依次列出，并分别录入其量值，即性能数据，记作：

$$M_S=\begin{bmatrix}1^{\#}\text{楼}, & \text{高耸性}, & \text{较强}\\ & \text{实体性}, & \text{一般}\\ & \text{连续性}, & \text{较弱}\\ & \text{紧凑性}, & \text{较弱}\end{bmatrix}$$

与1#楼同法，将其他6栋住宅分别以上述形式进行基元化表达。在此基础上，将各个基元转化为统一的数据表，如表5所示。

在表5的基础上，使用各类经定义或自定义的已知的、量化的、基本属性的乘积，以体现表5中的各项特征，即

$$I_t=h/S_p=h(wl)^{-1} \tag{5}$$

式中，I_t为高耸性；h为高度；S_p为投影面积；w为面宽；l为进深。

$$I_s=S_w/S_f=S_w[2h(w+l)]^{-1} \tag{6}$$

式中,I_s为实体性;S_w为窗面积;S_f为立面面积;h为高度;w为面宽;l为进深。

$$I_c = h/w = hw^{-1} \quad (7)$$

式中,I_c为连续性;h为高度;w为面宽。

$$I_d = S_s/V = [2h(w+l)+wl](wlh)^{-1} \quad (8)$$

式中,I_d为紧凑性;S_s为表面积;V为体积;w为面宽;l为进深;h为高度。

表5 住宅建筑外部空间形态数据

Table 5 Outside space form data of housing

住宅	高耸性	实体性	连续性	紧凑性
1#楼	较强	一般	较弱	较弱
2#楼	一般	一般	较弱	较弱
3#楼	很强	较强	很弱	一般
4#楼	很强	较强	很弱	一般
5#楼	非常强	很强	极弱	一般
6#楼	较弱	较弱	一般	较强
7#楼	很弱	较弱	较强	很强

将各项特征转化为量化的性能数据,经转化后得到各建筑的空间形态数据如表6所示。

表6 住宅建筑外部空间形态量化数据

Table 6 Quantified outside space form data of housing

住宅	高耸性	实体性	连续性	紧凑性
1#楼	0.056	0.261	0.794	0.173
2#楼	0.040	0.259	0.795	0.178
3#楼	0.074	0.268	1.055	0.179
4#楼	0.074	0.268	1.055	0.179
5#楼	0.097	0.277	1.903	0.161
6#楼	0.037	0.252	0.517	0.181
7#楼	0.027	0.248	0.381	0.188

得出以上量化数据后,分别根据式(2)和式(3)对各性能数据进行标准化处理,其中建筑的高耸性、连续性和紧凑性没有明确、一致的值域,应用z-score公式进行标准化处理,而实体性具有明确、一致的值域,应用min-max公式进行标准化处理,得到标准化数据如表7所示。

表7 住宅建筑外部空间形态标准化数据

Table 7 Standardized outside space form data of housing

住宅	标准高耸性	标准实体性	标准连续性	标准紧凑性
1#楼	−0.042	0.448	−0.271	−0.500
2#楼	−0.708	0.379	−0.270	0.125
3#楼	0.708	0.670	0.253	0.250
4#楼	0.708	0.670	0.253	0.250
5#楼	1.417	1.000	1.956	−2.000
6#楼	−0.833	0.138	−0.827	0.500
7#楼	−1.250	0	−1.100	1.375

得到标准化的数据后,根据式(4)对各性能指标进行相关性分析。经计算,高耸性与实体性的相关系数为0.977,实体性与连续性的相关系数为0.956,其中实体性同时与高耸性、连续性具有极强相关,故将此项去除。其余项之间的相关系数均小于0.95,因此保留高耸性、连续型、紧凑性3项作为数据挖掘的对象,数据准备至此结束。

3 结论

以复杂的可拓建筑设计数据为对象,研究提出面向可拓建筑设计的数据准备流程与方法:建筑设计数据基元化表达、建筑要素数据表设计、数据筛选、基于数据属性类别的数据形式变换、变量标准化处理和变量维数约简。通过实施该流程及其操作方法,可以将可拓建筑设计数据转化为高质量的结构数据,为面向可拓建筑设计的可拓数据挖掘提供可行的操作对象。

参考文献(References)

[1] 杨春燕, 蔡文. 可拓数据挖掘研究进展[J]. 数学的实践与认识, 2009, 39(4): 134-141.
Yang Chunyan, Cai Wen. Research progress on extension data mining[J]. Mathematics in Practice and Theory, 2009, 39(4): 134-141.

[2] Sharma S, Osei-Bryson K M, Kasper G M. Evaluation of an integrated knowledge discovery and data mining process model[J]. Expert Systems with Applications, 2012, 39(13): 11335-11348.

[3] Lara J A, Lizcano D, Martinez A, et al. Data preparation for KDD through automatic reasoning based on description logic[J]. Information Systems, 2014, 44: 54-72.

[4] Zhang Shichao, Zhang Chengqi, Yang Qiang. Data preparation for data mining[J]. Applied Artificial Intelligence, 2003, 17(5-6): 375-381.

[5] 薛名辉. 可拓建筑设计的基本理论与应用方法研究[D]. 哈尔滨: 哈尔滨工业大学建筑学院, 2011.
Xue Minghui. The study on basic theory and applying methods of extension architecture design[D]. Harbin: School of Architecture of Harbin Institute of Technology, 2011.

[6] Yang Chunyan, Cai Wen. Extenics: Theory, method and application[M]. Beijing: Science Press, 2013: 27-33.

[7] 刘朝杰, 李秀宁, 任晓晖, 等. 问卷备择答案等距离量化合理性的验证方法初探[J]. 中国心理卫生杂志, 2004, 18(7): 468-479.
Liu Chaojie, Li Xiuning, Ren Xiaohui, et al. Testing the interval feature of alternative answers in questionnaire[J]. Chinese Mental Health Journal, 2004, 18(7): 468-479.

[8] 俞立平, 潘云涛, 武夷山. 基于极值法的学术期刊组合评价研究[J]. 图书与情报, 2009(4): 22-27.
Yu Liping, Pan Yuntao, Wu Yishan. Research on combined assessing for academic periodical based on extreme value method[J]. Library and Information, 2009(4): 22-27.

[9] Al Zaabi O S H. Potential for the application of emerging market Z-score in UAE Islamic banks[J]. International Journal of Islamic and Middle Eastern Finance and Management, 2011,4(2): 158-173.

[10] Adler J, Parmryd I. Quantifying colocalization by correlation: The pearson correlation coefficient is superior to the mander's overlap coefficient[J]. Cytometry Part A, 2010, 77(8): 733-742.

(责任编辑 韩星明)