

doi: 10.12052/gdutxb.210045

知识图谱的拓展及其智能拓展库

汪培庄¹, 曾繁慧^{1,2}, 孙 慧^{1,2}, 李兴森³, 郭建威¹, 孟祥福¹, 何 静⁴

(1. 辽宁工程技术大学 智能工程与数学研究院, 辽宁 阜新 123000; 2. 辽宁工程技术大学 理学院, 辽宁 阜新 123000;
3. 广东工业大学 可拓学及创新方法研究所, 广东 广州 510006; 4. 国家高性能计算中心 成都分中心, 四川 成都 610094)

摘要: 知识图谱发展的终极目标是要把数据库构建成为人脑知识生长的同构体, 它不是被动的僵化储存所, 而是主动生长的拓展体。本文提出了智能拓展库的构想, 它是知识图谱的升级版, 也是一个多层次的嵌入式图数据库系统。每个概念节点都是一个窗口, 打开窗口即生成一个子库。所有概念都按层次编码, 可快捷查询。一个概念节点称为花苞, 如果该概念的多个因素都以它为定义域, 则打开花苞的窗口, 就是一张因果分析表, 用因素空间理论可以回答智能判断、推理、评价、决策、规划等一系列问题, 使数据库成为智能孵化生长的活体。本文建立的智能拓展库语言有可能用于机制主义通用人工智能理论, 并开展智能孵化的全民工程。

关键词: 知识图谱; 机制主义人工智能; 因素空间; 可拓学; 智能拓展库

中图分类号: TP391; TP182 **文献标志码:** A **文章编号:** 1007-7162(2021)04-0009-08

Extension of Knowledge Graph and its Intelligent Extension Library

Wang Pei-zhuang¹, Zeng Fan-hui^{1,2}, Sun Hui^{1,2}, Li Xing-sen³,
Guo Jian-wei¹, Meng Xiang-fu¹, He Jing⁴

(1. Institute of Intelligence Engineering and Mathematics, Liaoning Technical University, Fuxin 123000, China; 2. College of Science, Liaoning Technical University, Fuxin 123000, China; 3. Institute of Extenics and Innovation Methods, Guangdong University of Technology, Guangzhou 510006, China; 4. Chengdu Sub Center, National High Performance Computing Center, Chengdu 610094, China)

Abstract: The ultimate goal of constructing the database of the knowledge graph should be the isomorphism of human brain knowledge growth, and it should not be a passive solidified storage body but an actively growing hatching body. The idea of intelligent reproduction library is proposed, which is an updated version of the Knowledge graph, a multi-level embedded graph database system. Each concept node is a window. Open the window, and there is a sub-library produced. All concepts are coded hierarchically for fast query. A concept node is called a bud if the concept has multiple factors taking it as domain. Open the window of the bud, and a cause and effect analysis table is prepared, on which, factor space theory can perform algorithms to make intelligent judgment, reasoning, evaluation, decision-making and planning, by which the database becomes a living intelligent incubator. The Factorial Query Language (FQL) is also established, which may be used in the preparation of the intelligent incubation civil intelligent engineering, with guidance of the mechanistic general artificial intelligence theory.

Key words: knowledge graph; mechanical artificial intelligence; factor space; Extenics; intelligent extension library

人工智能所面临的一个迫切问题是要有统一的
人工智能理论^[1-3]; 知识图谱是当前人工智能发展的

热点^[4-5], 在肯定其重大意义的同时, 要看到它在理论
上的不完善, 以及发展方向上的不明确。知识图谱发

收稿日期: 2021-03-10
基金项目: 辽宁省教育厅科学技术研究经费资助项目(LJ2019JL019); 辽宁工程技术大学学科创新团队资助项目(LNTU20TD-35)
作者简介: 汪培庄(1936-), 男, 教授, 博士生导师, 主要研究方向为模糊数学、因素空间理论等
通信作者: 曾繁慧(1970-), 女, 教授, 博士, 主要研究方向为基于因素空间理论的数据挖掘与知识表示技术、模糊结构元等, E-mail: 597873883@qq.com

展的最终目标是要把数据库构建成为人脑知识生长的同构体,它不是被动的僵化储存所,而是主动生长的拓展体。本文提出了智能拓展库的构想,这是知识图谱的升级版,是多层次的嵌入式图数据库系统。因素牵引的知识增长表达式是知识谱系多支图的基元,由此建成一个概念谱系图网络,勾画出有关领域的知识本体结构。基于因素表达事实作基元所生成的另外一个图网络则是进行智能生成与工作的系统,利用因素空间的理论和方法来辅助智能的开拓。本文的结构:第1节对知识图谱进行简评;第2节介绍因素空间对知识图谱的拓展问题的帮助;第3节对智能拓展库进行构想,并介绍因素查询语言;第4节说明问题和展望。

1 知识图谱简评

人工智能有2个截然不同的任务:(1)把人类已有的知识从图书、文件、信函中转移给机器;(2)让机器模拟人脑认识和改造世界。前者似乎是容易且该先做的事情,但事实恰好相反,它比后者更难而被长期拖延。在2012年前后,“知识图谱”的名称在谷歌叫响,掌握互联网资源的几家巨头公司竞相用此技术来开发新的搜索引擎。互联网是传递信函的渠道,知识图谱必须理解自然语言,因而也必然会加速自然语言理解的研究进程,它的历史贡献首先是实现图书、文件、信函的数字化。现在,它已使自然语言理解从数据驱动的字频统计方法转向知识连同数据驱动的研究途径。图数据网络模型已经在跨越同义字、反义字等歧义鸿沟方面取得了明显成效。知识图谱必会在机器模拟人脑认识和改造世界方面取得卓越成效,它为智能化提供了网络传输的翅膀,其前景不可限量。

在肯定其意义的同时,也要看到其在基础理论方面的不足。知识图谱与传统的知识表示的分界线在于它们所采用语言的分离:关系数据库的查询靠SQL语言(Structured Query Language);知识图谱的查询要靠SPARQL语言(Simple Protocol and RDF Query Language)。SQL叫做表库语言,SPARQL叫做图库语言,图数据库的名称便由此而来。

知识表示的每一种方法都要表达事实,无论差异如何,都要在语言上符合主、谓、宾的SPO表达式,把主语和宾语视为节点,把谓语视为有向边,它们都是图的基元。知识图谱的数学定义是一个有向图,有向图的定义是在一组前节点 H 、一组有向边 G 和一组后节点 T 之间所建立的特定关系,也就是笛

卡尔乘积空间 $H \times G \times T$ 中的一个特定子集。一张关系数据库表是由对象、属性名、属性值所建立的一个三元关系,设三者所构成集合分别是 H, G, T ,表中每个格子点所对应的三元组就是 $H \times G \times T$ 中的一个点,一个用属性名作边的图基元;整个表就是 $H \times G \times T$ 中的一个子集,就是一个知识图谱。所以,表数据库与图数据库并没有本质的差别;数学上的“图”并非生活当中的图,生活当中的图是看得见的绘制品,称SPARQL为图数据库语言,是因为Web是生活中的图,有节点和边。但Web图其实并不符合图的数学定义,SPARQL图中的节点与Web节点的域名毫无关系。所以,知识图谱从基础上说理论尚欠严谨。现在人们误以为图数据的特征是直观可见,这适用于小数据,当节点稍多时,复杂一点的图谱就不再直观,这使可视化研究变得格外重要。

SPARQL语言并不是理想的语言,它的程序在阅读和编写上极其繁琐,它的查询功能不像SQL语言那样可进行是非判断和推理,只是回答给定的一个基元在不在库中,进而回答给定的一组基元是否与某个子库同构。SPARQL语言的推理功能更是无法与SQL语言相比,只能发现“叔侄”是“兄弟”和“父子”关系的叠加这一类规则,尽管这种推理有其独特作用,但都需要进行提升和改进。

SPARQL语言打破了人们的惯性思维,给编程带来新的研究前景,开源的知识图谱可以在短短几周的时间内就发展到节点过亿的规模。各种限制都可以考虑取消,新的设想和构思不断涌现,但也要保持头脑清醒,在大潮中求实求稳。图谱要大而不乱,活而不杂。现在已经发现有错,也有纠错的算法,但是在庞大的图谱中要搜寻迥路,可能是一个NP(Nondeterminism Polynomial)难问题。图谱要计算代价、讲求效率、防止浪费,更要防止对环境可能带来的污染。

在知识图谱的发展方面,实践一直发展在理论的前头,现在需要加强基础理论,尤其是数学理论。

2 因素空间的知识谱系理论

2.1 因素是智能数学的元词

人脑面对事物的第一反应就是要回答“这是什么?”的问题。神经中枢把对象信息传递到记忆单元,查找该对象的存储位置,或建立新档,或用旧档进行对比判断,迅速做出应答。这是最基本的思维活动环节,其数学形式可表达为

$$e \text{ is } p$$

这里 e 是对象或实体(包括物和事), p 是一个概念(包括属性,属性是简单概念)。这种表达虽很自然但却不能采用。难道人脑的反映如照镜子般,一个对象就只有同一个映相?如果如此,思维的目的性怎样体现?为了选拔举重运动员,是否需要注意对象的面貌?为了选美,能否不注意对象的面貌?同是一个对象,不同的目的有不同的注意视角,得到不同的表相。注意的视角是认知过程中最重要的元素,叫做因素。

因素在数学上被定义成一个映射 $f: D(f) \rightarrow I(f)$,这里, $D(f)$ 叫做 f 的定义域或论域, $I(f)$ 叫做 f 的相域或信息域。

例如,因素 f =颜色, $D(f)$ =楼前停的5辆车 $\{d_1, d_2, d_3, d_4, d_5\}$, $I(f)=\{\text{红, 白, 黑}\}$ 。颜色是把车变为车色的映射,如 $f(d_1)=\text{红}$, $f(d_2)=\text{白}$, $f(d_3)=\text{红}$, $f(d_4)=\text{黑}$, $f(d_5)=\text{白}$ 。

知元表达式:

$$f(e) = p \quad (1)$$

例如,“这辆车的颜色是红的”。

从对象 e 变为对象在某个因素 f 之下的相 $f(e)$,把注意引入思维描述,式(1)之所以是思维的元表达式,是因为它引进了因素,因素是智能描述的元词。智能科学于与物质科学的根本区别就在于是否把因素引入思维的元描述。

可拓学^[6]早就睿智地提出了“物元”和“事元”的概念,并在这方面做了很好的研究^[7]。“知元”是对“物元”和“事元”的一种承袭,只是特别强调了因素的地位和作用。

记因素 ζ =“自身”,这个因素把任何对象映射成自身: $\zeta(e) = e$,于是,“ $f(e) = p$ ”可以概括“ e 是 p ”。

关系也可以用因素来表示。例如,“张三是李四的妻子”表达的是李四和张三之间有夫妻关系。取因素 f =“妻”,取 $D(f)$ 为某人群中的已婚男子, $I(f)$ 为该人群中的已婚女子,因素“妻”是从 $D(f)$ 到 $I(f)$ 的映射。这句话可表为“ $f(\text{李四})=\text{张三}$ ”。表示关系的因素叫做关系因素。

因素的定义域 $D(f)$ 往往被取为一个概念 α 的外延,这样的因素叫做谱系因素。概念 α 叫做因素 f 的被分概念;因素 f 叫做 α 的被定义因素; $I(f)$ 中的概念叫做因素 f 的生出概念;因素 f 叫做它们的导出因素。

单一对象也可以被视为概念,例如,“北京”只有一个,也可以当作概念。所以,关系因素也是谱系因素。

概念定义因素,因素导出概念,因素与概念是否

一样多?的确,同一个因素在对象的层次结构中会产生众多的后裔因素,例如, f =“形态”这一因素,当 $D(f)=[\text{人体}]$ 时,相是头、身和四肢的尺寸比例;当 $D(f)=[\text{面貌}]$ 时,相是眼、眉、鼻、口等五官搭配。同是一个字根“形态”,以被分概念作前缀,就有许多派生因素。但是,字根因素并不多,从注意方向来分有3类:目标、形式和效用;从人的感官来分有视、听、嗅、味、触等。

因素的相是表示属性、情感和意向的词汇,也可以是表示形态或程度的数。相域 $I(f)$ 不是相的随机凑合,而是由因素 f 所统帅的整齐阵列。颜色统帅红、黄、蓝等色而不能混入“大”“高”“忙”等词汇。因素是变项,它在自己的相域中取值变化。

人的智能活动不靠条件反射,而是靠人脑所具有的因果分析的能力。因素是因果分析的要素。因素非因,乃因之素。“雨量充沛”是取得“好收成”的原因,但却不是因素,这里的因素是降雨量。它是一个变量,其变化可以使农作物丰收,也可以使之颗粒无收,显示了它对收成有重要影响,这才使人断定“雨量充沛”是取得“好收成”的原因。因果分析的核心思想不是从属性或状态层面孤立静止地去找原因,而是要先从更深层面上去寻找对结果最有影响的因素,只有找到了这组因素,才能找到最佳的原因。从找原因到找因素是人脑认识的一种升华,也是因果性科学的思想核心。

因素是定性的变项,是广义的变量。它可以把定性的相域嵌入欧氏空间的定量相域中去,转化为普通的变量。前提是要把相域按一定目标有序化。例如职业相域 $=\{\text{工人, 农民, 士兵, 企业主, 雇员, 教师, 医生, 律师, 官员, } \dots\}$,这些职业之间没有次序。但是在高考生报考志愿时就要对未来的职业排序。工资待遇、社会需要、兴趣爱好、综合加权各有不同排法。当 $I(f)$ 变成了全序或者偏序集合以后,定性相域就可以嵌入到一个实数区间或多维超矩形里。这个实空间可以选择为 $[0, 1]$ 或者 $[0, 1]^n$,这时,所有相域都是对目标的某种满足度。而满足度又可化为某种逻辑真值。嵌入实空间的相域是离散的。可取二值相域 $I(f) = \{0, 1\}$,或三值相域 $I(f) = \{1, 2, 3\}$ 或 $\{-1, 0, +1\}$ 。离散值相域称为子架或托架。若把每个因素视为一个坐标轴,多个因素就立定出一个坐标架,形成的空间叫做因素空间,所有事物都可成为因素空间中的一个点,为人工智能提供了普适性的事物描述框架。

因素有几种特殊的叫法:(1) 两极叫法,如“美丑”; (2) 后面加问号,如“美丽?”; (3) 前面加“有无”

或“是否”,如“是否美丽”;(4)后面加“性”字,如“美丽性”。

因素与属性不能混淆。属性能问是非:“这花是紫的吗?”因素不能问是非:“这花是颜色吗?”。属性是被动描述的静态词;因素是主动牵引思维的动态词。“Attribute”在形式概念分析(Formal Concept Analysis, FCA)^[8]中代表属性值,如“红”“黄”等;但在粗糙集(Rough Sets, RS)^[9]中代表属性名,如“颜色”等。在术语上出现了混乱。FCA和RS都是在1982年与因素空间(Factor Space, FS)^[10]同年创立,协调的约定是:FCA中的“Attribute”=FS中的“属性”;RS中的“Attribute”=FS中的“因素”。

因素比属性高一个层次,因素统帅属性。人脑是高效率的信息处理器,按因素来组织感觉神经元。孟德尔在遗传学研究中苦于生物属性的杂乱,在1865年提出了基因的概念,他所使用的英文名字就是Factor(约翰森在1909年才改名为Gene),基因就是生物属性的质根,因素是广义的基因。基因打开了生命科学的大门,因素是从数学上帮助打开信息科学大门的一把钥匙。

2.2 因素/概念谱系

概念是一个二元组 $\alpha = (a, [a])$,其中, a 是对概念 α 的描述语句,称为 α 的内涵, $[a]$ 是由满足内涵描述的全体对象的集合,称为 α 的外延。婴儿出生的时候只有零概念,内涵是零描述,外延是整个宇宙混沌一团。人类知识是从零概念开始,经过一步一步的概念团粒分裂进化而来的。每次分裂,外延缩小,内涵描述的语句增加。

每个内涵描述句都是由因素所表达的知元表示句,见式(1)。它们被外延中全体对象所满足,也就是说,团粒中的所有对象在有关因素下的相值均相同。当目标需求不满足时,知识就需要更新了。这时,人的注意力就会根据目标需求而确定一个新的因素,它在团粒中能取不同的相值,按照它的不同相值来分类,概念团粒就分裂了。因素就是概念团粒细化的分化器,这是知识增长的关键环节,需要设立一种统一的表达形式。

知增表达式:

$$U = D(f) \rightarrow I(f) = \{\alpha_1, \dots, \alpha_k\} \quad (2)$$

式(2)中, U 是上位概念的外延,它被取为谱系因素 f 的定义域; $\alpha_1, \dots, \alpha_k$ 是上位概念所分化出来的一组下位概念名称。它们的外延之并等于 $U: [a_1] + \dots + [a_k] = U$ 。这里, $[a_j]$ 表示概念 α_j 的外延。

例1 宇宙 = $D(\text{虚实}) \rightarrow I(\text{虚实}) = \{\text{精神}, \text{物质}\}$ 。

这是一个知增表达式,简称概念分化式。被分化的是零概念,它的概念团粒是宇宙。“虚实”是定义在宇宙万事万物上的一个因素。除了虚实之外是否能找到以宇宙为定义域的因素?“身高”是因素,但只对能直立起来的动物有意义,对石头则没有意义。“重量”是因素,但只对物质有意义,对精神界则没有意义。因此能找到最普遍、最抽象的因素寥寥无几。在某种意义上,甚至是唯一的。例如“阴阳”,它也是最具普遍意义的一个因素,但和虚实可以相互转化。虚实这个因素把宇宙划分成两大类,零概念被分化为物质与精神2个概念。

物质 = $D(\text{生命性}) \rightarrow I(\text{生命性}) = \{\text{生物}, \text{非生物}\}$ 。这又是一个分化式,被分化的是概念“物质”。“生命性”是定义在物质上的一个因素。除了生命性之外,还有“体积”“质量”“物态”等因素对所有物质都有意义,也可以对物质进行分类。生命性这个因素把物质外延分成2个团粒,生出生物和非生物2个下位概念。

精神 = $D(\text{文理}) \rightarrow I(\text{文理}) = \{\text{文科}, \text{理科}\}$; 生物 = $D(\text{动植物}) \rightarrow I(\text{动植物}) = \{\text{动物}, \text{植物}\}$; 非生物 = $D(\text{金属?}) \rightarrow I(\text{金属?}) = \{\text{金属}, \text{非金属}\}$; 动物 = $D(\text{脊椎?}) \rightarrow I(\text{脊椎?}) = \{\text{脊椎}, \text{非脊椎}\}$; 植物 = $D(\text{高度}) \rightarrow I(\text{高度}) = \{\text{乔木}, \text{灌木}, \text{草}, \text{苔}\}$; 脊椎动物 = $D(\text{哺乳?}) \rightarrow I(\text{哺乳?}) = \{\text{哺乳动物}, \text{非哺乳脊椎动物}\}$ 。

例1说明了因素是怎样牵引着概念分化和知识增长的。因素在划分概念的过程中,也塑造出自己发展的谱系。因素被定义域所制约,在定义域之外,因素会失去意义。于是,在因素与因素之间出现了生与被生的关系:没有虚实因素的划分,就没有物质的概念,没有物质的外延,就没有因素生命性的定义域,生命性就失去了生存的土壤。所以,因素虚实生出了因素生命性。

定义1^[3] 如果因素 f 的定义域 $D(f)$ 真包含(不相等)因素 g 的定义域 $D(g)$,即 $D(f) \supset D(g)$,则称 f 是 g 的祖先而称 g 是 f 的后裔。

对谱系因素来说,一个因素 f 必对应着一个概念 α ,其外延 $[a]$ 是 f 的定义域。显然,因素的祖裔关系等价于概念的上下位关系,在因素之间称上下位,在概念之间称祖裔。也可以对因素和概念列出共享的祖裔关系表,叫做因素/概念谱系表。它是因素集或概念集上的关系,可写成一个0-1方阵 $C, c_{ij} = 1$ 当且仅当因素 f_i 是因素 f_j 的祖因素;否则为0。

显然,祖裔关系具有传递性,构建 C 的传递闭包

[C]: 记 $C^2=C\times C$, 这里, $C\times C$ 是矩阵的模糊乘积, 它仿照 C 与 C 的普通矩阵乘法, 把数的加法改成取大, 乘法改成取小。如果 $C^2=C$, 则说明是传递闭包, 得到 $[C]=C$; 否则, 计算 $C^4=C^2\times C^2$, 若 $C^4=C^2$, 则得到 $[C]=C^2$ 。如此下去, 可得到传递闭包。

为了能画出祖裔关系图, 需要先把祖裔表变为父子关系表, 具体做法是: 对 C 中的每个非零元素进行甄别: 设 $c_{ij}=1$, 若存在 k , 使有 $c_{ik}=1$ 且 $c_{kj}=1$, 则改令 $c_{ij}=0$ 。记改后的矩阵为 C_0 , 它就是父子关系表。有父子关系表就能直接画成一个图, 称为因素/概念谱系图, 见图1, 也称为知识谱系图。由于图1的基元都是从一个前节点分为多个后节点, 故是多支图。知识谱系图反映了相关知识的本体结构。

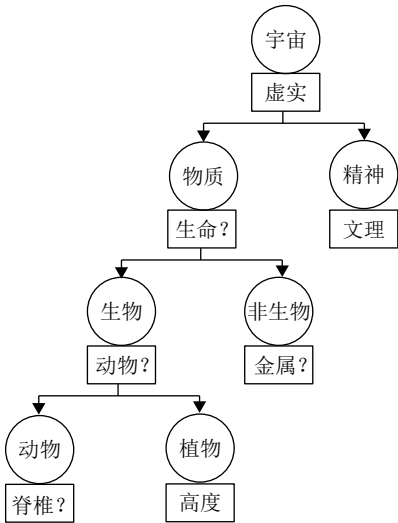


图1 因素/概念谱系
Fig.1 Factor/concept spectrum

用记号“ $\geq$ ”来表示因素谱系的序关系, $f\geq g$ 当且仅当因素 f 是 g 的祖因素或 f 与 g 是同一个因素。“ $\geq$ ”满足反身性和传递性, 要想使它成为偏序, 还必须满足反对称性: 若 $f\geq g$ 且 $g\geq f$, 则 $f=g$ 。但是, 同一个定义域上可以有多个不同的因素, 要想使反对称性成立, 唯有把定义域相同的因素都归为一类。在任意一个因素集 S 中, 如果 $D(f)=D(g)$, 则记 $f\sim g$ 。显然“ $\sim$ ”是 S 中的一个等价关系, 取商空间 $(S*,\geq)=S/\sim$, 序“ $\geq$ ”在类之间保持不变, $(S*,\geq)$ 便构成一个偏序集。这个偏序集的图与因素谱系图是一样的, 只不过要把因素改为因素类。如果一个类只包含一个因素, 它就是原来图中所标注的因素; 如果一个类包含多个因素, 那就在该类贴上一个花苞的标签, 表示这里要开花, 所开的花就是一个因素空间。

定义2 称 $I_F=(E,F)$ 为论域 E 上的一个因素空

间, 如果 $F=\{f_1,\cdots,f_n\}$ 是一组因素的集合, 它们有共同的定义域: $D(f_1)=\cdots=D(f_n)=E$ 。

记号 I_F 表示一个以因素为足码的信息空间。由于信息域经常要定量化, 因素空间也习惯地记为 X_F 。

因素空间是智能活动的平台, 它在同定义域的因素之间定义了逻辑、数学和注意力权重等多方面的运算, 可以快捷地进行识别、判断、归纳、推理、预测、控制、评价、决策等智能活动^[10-15]。

如果说粗糙集为关系数据库提供了样本理论, 那么, 因素空间就为粗糙集提供了母体理论。如前所述, 粗糙集的属性名就是因素, 信息系统表就是因素分析表, 表的每一行就是因素空间中的一个样本点。信息决策表就是因果分析表, 粗糙集的决策分析和分类学习都在因果分析与决策方法之中。因素空间能完成粗糙集能做的事, 而且做得更快, 粗糙集不能做而因素空间能做的事情有6类。

(1) 对因素(包括属性名)的逻辑运算给出了明确的数学定义。

(2) 提出了背景关系和背景分布的概念。给定因素空间 $I_F=(U,F=\{f_1,\cdots,f_n\})$, 这 n 个因素的联合相域是 n 个因素相域的笛卡尔乘积空间, 即 $I(F)=I(f_1)\times\cdots\times I(f_n)$ 。假定所有因素的相域都有3个相, 分成3个格子, 那么联合相空间就有 3^n 个格子。这么多的格子不是都有样本点出现。例如“气温”和“降雨量”是2个因素, 由于它们彼此并非独立而是互成正变, 气温高且降雨量低和气温低且降雨量高这两个格子就不可能出现样本点, 把这两个格子去掉, 剩下的格子点所形成的集合有特别重要的意义。

定义3 给定因素空间 $I_F=(E,F):F=\{f_1,\cdots,f_n\}$, 记

$$R=F(E)=\{f_1(e),\cdots,f_n(e)|e\in E\} \tag{3}$$

R 叫做 $f_1,\cdots,f_n$ 之间的背景关系。

背景关系反映了因素之间的相互影响。当背景关系没有去掉任何格子点时, 因素就是独立的。粗糙集没有引进背景关系, 就只能表现独立因素, 而独立因素之间不存在因果联系, 这就先天制约了它进行因果分析的能力。

(3) 背景关系决定了因素之间的一切归纳与推理, 它一般是凸集, 可以由少数个顶点生成。这一组顶点的集合叫做背景基 B 。 B 中点的凸组合可以把 R 的点全部复原出来, 故可以用 B 取代 R , 这可以实现数据的大幅压缩, 把大数据控制在一个小数水平, 背景基是大数据的降龙罩。

(4) 将概率逻辑引入因果分析。

定义4 给定联合相空间中的一组样本点 $\{f_1(e_1),\cdots,$

$f_n(e_m)\}(e_1, \dots, e_m \in E)$, 记 $r_{i(1) \dots i(n)}$ 为落入格子 $i(1) \dots i(n)$ 中的样本点频率, 与此频率所对应的概率分布叫做因素空间中的背景分布。

这个定义把概率逻辑引入了因素空间, 贝叶斯的逆向推理可以从目标倒逼条件, 为问题求解提供了便利。

(5) 将模糊逻辑引入识别与推理。

定义5 设背景分布的最大值为 L , 记

$$A_{i(1) \dots i(n)} = (1/L)r_{i(1) \dots i(n)}$$

(4)

$A_{i(1) \dots i(n)}$ 称为模糊背景分布在格子 $i(1) \dots i(n)$ 中的样本隶属度。

因素空间是模糊数学的升级版, 它所导出的模糊落影理论, 把模糊度定义为随机集的覆盖概率, 把模糊现象转化为幂上的随机现象。为主观性测度建立了坚实的理论基础和可行的实践方法。

(6) 通过定性坐标的定量化、线性规划等优化理论可以进入智能孵化的过程。

2.3 用因素对子库名进行编码

对概念名词进行编码是自然语言理解最重要的工程。这一工程进展不易的原因是没有聘用因素。内涵都是靠因素来叙述的, 只有因素才能确定内涵, 靠因素来编码就能看准同义字、近义字和反义字, 并确立国际语言的范本。

任何表名或图名都是一个概念名词, 这些名词必定组成一个概念谱系。下面涉及8个因素和它们所生出的19个概念, 按祖裔的世代排列见图2。

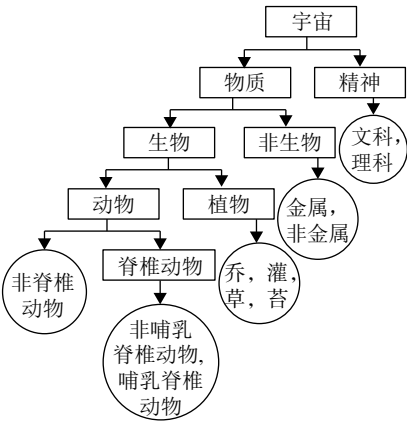


图2 祖裔世代排列谱系

Fig.2 Pedigree of ancestral generation arrangement

在始祖概念“宇宙”之下, 18个概念按世代排成5行。

预备步骤为将所有因素的相分别编码, 加到括号中。

- $I(\text{虚实}) = \{\text{精神}(0), \text{物质}(1)\};$
 $I(\text{生命性}) = \{\text{非生物}(0), \text{生物}(1)\};$
 $I(\text{文理}) = \{\text{文科}(0), \text{理科}(1)\};$
 $I(\text{动静}) = \{\text{植物}(0), \text{动物}(1)\};$
 $I(\text{高度}) = \{\text{乔}(1), \text{灌}(2), \text{草}(3), \text{苔}(4)\};$
 $I(\text{金属?}) = \{\text{非金属}(0), \text{金属}(1)\};$
 $I(\text{脊椎?}) = \{\text{非脊椎动物}(0), \text{脊椎动物}(1)\};$
 $I(\text{哺乳?}) = \{\text{非哺乳脊椎动物}(0), \text{哺乳动物}(1)\}。$
- 它们的编码为
- 物质: 1; 精神: 0;
- 生物: 1I; 非生物: 10; 文科: 00; 理科: 01;
- 动物: 111; 植物: 110; 金属: 101; 非金属: 100;
- 脊椎: 1111; 非脊椎: 1110; 乔: 1101; 灌: 1102;
- 草: 1103; 苔: 1104;
- 哺乳动物: 11111; 非哺乳的脊椎动物: 11110。

宇宙是始祖, 不编码。其他概念的编码原则是: 以概念的世代来定码位, 再按各因素相域的编码来填入码值。码的位数越少, 码字越短, 概念的辈分越高, 概念团粒越大。子概念必须继承父概念的编码, 在父码之后加上自己对其生出因素的相值码。

例如, 排在第一行的精神由宇宙的被定义因素“虚实”所分出, 在虚实相域中的编号为0。因它是宇宙的第一代子概念, 就把0排在第一位。又如, 理科是因素“文理”所分, 它在相域中的编号为0, 因属于第2代, 将0填入第2位。因它是精神的子概念, 第一位照写0, 故其编码为00。又如, 苔是因素“植物高度”所分, 它在相域中的编号为4, 因属于第4代, 将4填入第4位。因它是植物的子概念, 前3位照写110, 故其编码为1104。

编码根本意义在于把内涵数字化, 这蕴含着难以想见的应用。最显眼的应用是查询。在包含大量概念的图中要找出其中的一个, 先按第一足码缩小搜索范围, 再查第二足码缩小范围, 直到该概念被查出, 这是最快的方式。值得指出的是, 例1的概念谱系图是一颗倒树, 由于目标的不同, 会得到不同的倒树, 这样所得到的林, 会打乱世代次序, 两树的交点会有不同的编码。但就像图书目录有多种查号一样, 能有多种编码, 但不能一个编码对应着多个不同义的概念, 如果是那样, 就需要寻找新的因素增加码位。

编码的意义除了快速查询之外, 另一个重要应用是快速进行演绎推理。

命题1 若概念乙的编码是概念甲编码的前一部分, 则推理句“甲 $\rightarrow$ 乙”必真。

证明 若概念乙的编码是概念甲编码的前一部分,则概念乙必是概念甲的祖先,概念甲的外延必被概念乙的外延所包含,按集合论中著名的Stone表现定理知,逻辑的蕴含就是外延的被包含,故推理句“甲→乙”必真。

树状图编码有明确的世代,多树成林,会出现乱代现象,编码不唯一,但像人的家谱一样,没有害处。

3 智能拓展库的构建

3.1 库的结构

智能拓展库是由一系列知识谱系图所构成的系统。每张知识谱系图的基元就是一个多支图。图与图之间存在着层次关系,形成嵌套结构。

(1) 子库的嵌入和编码。以图1为例,“理科”位于图的末端,称为足概念。它的主库中有码01,若不满足这个概念,还想知道有多少不同的理科,这就可以把理科当做按键,一按就开出一个窗口,里面亮出一个以理科为始祖的子库。由数、理、化、代数、几何、分析等6个概念组成。编码原则是层层负责,子库要对它所属的全部6个节点进行编码:数1、理2、化3、代数11、几何12、分析13。

但是,子表不能给自己的表名编码。因为子表名就是上级足概念节点之名,都叫做“理科”,所以子表表名的编码已经被上级定为01,子表中各节点全名等于子表名下接表内名:数 011、理012、化013、代数0111、几何0112、分析0113。这就叫做图的嵌入和编码的层层负责制。

(2) 概念查询。假定有一概念,已知其编码,要在总库中查出来,则在总表中寻找具有该概念码的第一位码的节点,直到无法下移,点击按键进入下层子表。继续查询,直到查获为止。

(3) 因素空间藤。知识谱系图中有一些带花的节点,在那些概念上同时定义多个因素形成因素空间。带有因素空间的谱系就叫做因素空间藤。对每个花苞,都必须在窗口下存放一张因果分析表。因果分析图的职责是实现因素空间所能做的一切工作:识别、归纳、推理、预测、评价、决策、规划等智能活动。

3.2 因素查询语言FQL(Factorial Query Language)

FQL和SQL语言一样,具有4种功能:数据定义、数据操纵、数据控制和数据查询。下文分述不同之处。

1) 数据定义语言

数据定义的任务是根据目标从主库中建立子库

的知识谱系图。

生成知识谱系图的前提是要靠专家先给程序员一张因素词表,表上用自然语言写出所有要用到的因素名称,并在每一个因素 f 下面写出它的定义域 $D(f)$ 的名称,并把该名称所属的要用的对象填入括号。例如,若在主库中有知元式:

字号(周瑜)=公瑾,字号(关羽)=云长,……

则必须由专家输入:因素“字号”, D =中国古人={周瑜,关羽,……}

D 不能再大,不能扩大为中国人,因为现代中国人没有字号,也不宜缩小为三国时代人。

在这份表中还要注明定义域之间的包含关系,例如:[人] $\supseteq$ [中国古人],[脊椎动物] $\supseteq$ [哺乳动物]等。

如果输入的名单足够充分,程序员就可以操作如下:(1) 在因素之间定义祖裔关系: f 是 g 的先祖, g 是 f 的后裔,记作 $f > g$,如果 $D(f) \supset D(g)$ 。(2) 建立从因素到因素的祖裔关系矩阵 M , $r_{ij} = 1$,如果 $f_i > f_j$;否则 $r_{ij} = 0$ 。(3) 计算矩阵乘法: $M^2, M^4, \dots$ 直到 $M^k = M^{2k} = M^*$, M^* 是传递闭包,它使祖裔成为一种偏序关系,这时用图来表示,节点祖通过因素祖裔指向后裔节点,这就是所要的因素谱系图。

写出因果分析表的过程与关系数据库表的一样。

2) 数据的查询

规定1 对于给定的子库,数据查询的基本任务是:给定一个域名,要写出子库的因素谱系图;对任意一个输入的因素,要回答它在不在主库中。若在,则找到所在的因果分析表进行查询,其查询方式同于关系数据库的查询。

给定因果分析表 D ,若 $F(D)$ 是个单因素 f ,称 $(x_i, [x_i])$ 是一个单因素概念,这里, $[x_i]$ 是 x_i 在论域中的原像,即 $[x_i] = \{u \in U | f(u) = x_i\}$;若 $F(D) = f_1 \wedge \dots \wedge f_n$,设 $x = (x_1, \dots, x_n)$ 在 $I(F)$ 中出现的频数不接近于零,则称 $(x, [x])$ 是一个原子概念。 $[x]$ 是 x 在论域 U 中的原像,即 $[x] = \{u \in U | F(u) = x\}$ 。

FQL概念生成器要实现的功能是:(1) 给定多因素的因果分析表,生成原子概念集;(2) 按因素的分辨率从高到低,对论域进行划分,得到概念格。

框架中的槽,就是一组因素,一个侧面就是一个因素,到了侧面就没法再往下走了。因素谱系把框架思想发展了。

因素库语言可以描写状态空间,可以与图数据库的知识图谱相互转化,只不过把传统的知识图谱划分为非关系因素表和关系图谱,这样才能提高数据库的效能;盲目的混合开源扩张不宜渲染。FQL能

融合SQL和知识表示的各种语言。

4 问题和展望

知识图谱构建的目标是为系统性、全面性认识事物、提高决策水平和解决问题等服务。从被动的凝固储存体到智能拓展、主动生长是知识图谱发展的内在要求。本文提出了知识图谱智能孵化库的基本概念和构想,并用因素空间理论探讨实现智能判断、推理、评价、决策、规划等问题的基本策略,使知识图谱数据库变为智能孵化、智能拓展的活体。

后续的研究将进一步与可拓学理论结合,进行知识图谱智能拓展的交叉研究,研究大数据环境下知识图谱拓展库的的拓展、变换算法,研究知识图谱的基元-因素化表达及其运算,进一步提升知识图谱在问题处理等领域的智能化程度。本文提到的拓展还很成熟,离可拓学的拓展还有一定距离,需继续努力。

展望未来,希望智能拓展库可以进一步实现智能孵化器的作用,自上而下与自下而上相结合地开展智能孵化的全民工程。实现分布式多智能体想实现的系列功能。

参考文献:

- [1] 钟义信. 机制主义人工智能理论——一种通用人工智能理论[M]. 北京: 北京邮电大学出版社, 2021.
- [2] 何华灿. 泛逻辑学理论——机制主义人工智能理论的逻辑基础[M]. 北京: 北京邮电大学出版社, 2021.
- [3] 汪培庄. 因素空间理论——机制主义人工智能理论的数学基础[M]. 北京: 北京邮电大学出版社, 2021.
- [4] 刘峤, 李杨, 段宏, 等. 知识图谱构建技术综述[J]. 计算机研究与发展, 2016, 53(3): 582-600.
LIU Q, LI Y, DUAN H, *et al.* Knowledge graph construction techniques [J]. *Journal of Computer Research and Development*, 2016, 53(3): 582-600.
- [5] HIBA A, SHADY E. Effective searching of RDF knowledge graphs [J]. *Web Semantics: Science, Services and Agents on the World Wide Web*, 2018(48): 66-84.
- [6] 蔡文, 杨春燕. 可拓学的基础理论和方法体系[J]. 科学通

报, 2013, 58(13): 1190-1199.

CAI W, YANG C Y. Basic theory and method system of topology [J]. *Chinese Science Bulletin*, 2013, 58(13): 1190-1199.

- [7] 李兴森, 许立波, 刘海涛. 面向问题智能处理的基元—因素空间模型研究[J]. 广东工业大学学报, 2019, 36(1): 1-9.
LI X S, XU L B, LIU H T. A research on problem oriented intelligent processing model by basic-element and factor space [J]. *Journal of Guangdong University of Technology*, 2019, 36(1): 1-9.
- [8] WILLE R. Restructuring lattice theory: an approach based on hierarchies of concepts [J]. *Ordered Set*, 1982: 445-470.
- [9] PAWLAK Z. Rough sets [J]. *International Journal of Computer and Information Sciences*, 1982, 11(5): 341-356.
- [10] 汪培庄, 李洪兴. 知识表示的数学理论[M]. 天津: 天津科技出版社, 1994.
- [11] WANG P Z, SUGENO M. The factors field and background structure for fuzzy subsets [J]. *Fuzzy Math*, 1982, 2(2): 45-54.
- [12] 汪培庄. 模糊集与随机集落影[M]. 北京: 北京师范大学出版社, 1985.
- [13] 任思行, 曾繁慧, 郭嗣琮. 结构元理论下的模糊Markov决策过程[J]. 辽宁工程技术大学学报(自然科学版), 2020, 39(2): 180-183.
REN S H, ZENG F H, GUO S Z. Fuzzy markov decision process based on structural element theory [J]. *Journal of Liaoning Technical University (Natural Science)*, 2020, 39(2): 180-183.
- [14] 蒲凌杰, 曾繁慧, 汪培庄. 因素空间理论下基点分类算法研究[J]. 智能系统学报, 2020, 15(3): 528-536.
PU L J, ZENG F H, WANG P Z. Base point classification algorithm based on factor space theory [J]. *CAAI Transactions on Intelligent Systems*, 2020, 15(3): 528-536.
- [15] 孙慧, 曾繁慧, 蒲凌杰. 因素空间理论下多目标因果分析的降维算法研究[J]. 辽宁工程技术大学学报(自然科学版), 2021, 40(5): 466-472.
SUN H, ZENG F H, PU L J. Research on dimension reduction algorithm of multiple targets causal analysis based on factor space theory [J]. *Journal of Liaoning Technical University (Natural Science)*, 2021, 40(5): 466-472.

(责任编辑: 张玮欣)