大数据科学的认知和数学基础引论

WANG Yingxu[1,2],彭军[3]

1. 清华大学大数据系统软件国家工程重点实验室,清华大学北京信息科学与技术国家研究中心,北京 100084
2. 国际认知信息学与认知计算学会;卡尔加里大学电气与计算机工程系,Schulich工程学院,Hotchkiss脑科学研究所,加拿大卡尔加里 T2N 1N4
3. 重庆科技学院智能技术与工程学院,重庆 401331

摘要　大数据不仅在科学、工程与计算智能中有着广泛的应用,而且在人类感知、估计、量化、记忆和推理的认知机制中发挥着基础性作用。通过对大数据科学理论的基础研究,提出一组大数据系统的一般原理和分析方法。为了从形式上解释大数据的起源和本质,探讨大数据的认知基础及其数学模型,严格地引出了根植于科学、工程和社会各个领域的大数据的一般模式。研究发现大数据不再是传统实域上的纯数,而是一个前所未有的新型数学结构,称为递归类型化超结构(RTHS)。这一大数据系统的基本拓扑特性揭示了大数据工程的复杂性及其操作与处理的全新认知、理论挑战,以及可选解决方案。

关键词　大数据科学;大数据工程;大数据数学模型;递归超结构;认知计算;计算智能

大数据是人类社会信息化时代的代表现象之一[1-7]。由于大脑中所表示的认知对象可以根据其抽象层次从下至上分为数据、信息、知识和智能4种形式[1,6,8-20],人类活动的几乎所有领域都会产生不断增长的数据。数据在人类认知机制的基本层面如感觉、估计、量化、推理和与现实世界互动等方面均发挥着不可或缺的作用。

数论和数域曾经不断扩展,经历了从无类型数(𝕆)、二进制数(𝔹)、自然数(ℕ)、整数(ℤ)、实数(ℝ)、复数(ℂ)、模糊数(𝔽),到超结构数(ℍ[19,21],简称超数)的演化,如图1所示[4-5,13-16,22-24]。图1展示了

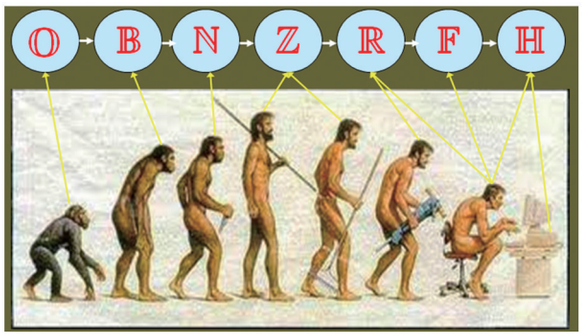

图1　数域与人类数据量化方法的扩展

收稿日期:2019-11-09; 修回日期:2020-01-19

基金项目:国家重点研发计划项目(2016YFB0501504);国家自然科学基金项目(U1509213)

作者简介:WANG Yingxu,教授,研究方向为认知信息学、软件科学、大数据代数和指称数学,电子信箱:yingxu@ucalgary.ca

引用格式:WANG Yingxu, 彭军. 大数据科学的认知和数学基础引论[J]. 科技导报, 2020, 38(3): 35-46; doi: 10.3981/j.issn.1000-7857.2020.03.002

人类为了处理现实世界中的实体及其在大脑中的严格表示而对抽象和量化方法的需求的进展过程。大数据科学中的重要发现之一[19]是数域的需求已经远远超出了传统的$\mathbb{R}$域中的纯数。譬如,近代的模糊数域$\mathbb{F}$的特征是一个二维结构[25-26],即$\mathbb{F}=\mathbb{R}\times\mathbb{I}=\{(\mathbb{R},\mathbb{I})\}$,其中每个成员被表示为在$[-\infty,+\infty]$中的一个变量和在(0,1]中的一个关联的隶属度。然而,超数$\mathbb{H}$是数域最新和最一般化的扩展,其可被形式表示为一个类型化n元组[4-5],如定义13所示。

定义1 数据(data)是现实实体或认知对象的量在特定量化尺度下的抽象表示,因而其是有物理量纲和数域类型的数字(digit)。

数据的抽象表达和经典关系模型由Ullman[17]和Codd[27]提出,其在计算机科学和传统数据库理论中得到了广泛的研究[3-4,16,22,28-31]。由于数据集的庞大性和复杂性已经超出了传统理论和技术的处理能力,自20世纪90年代以来,科学和工程界就面临着大数据的挑战。"大数据"一词最早是由Mashey在1998年提出[32]。McKinsey和Gartner从工业应用的角度探讨了大数据的高价值,而不仅仅是它的规模[32]。大数据在科学、工程和社会中无处不在的普遍性在随后的文献中也得到重视[1-2,32-33]。

定义2 大数据(big data)是在量和质、多样性、存储、检索、提取、计算、语义认知、维护和处理诸方面有别于无类型离散数字的超大规模异构类型化量所构成的一个超结构(hyperstructure)。

大数据的基本特征是非结构化、异构、单调增长、非描述性、混合/模糊语义,且一致性随时间衰减或熵随时间增加[23,34]。由于这些固有的复杂性和极大规模的多维超结构对象,大数据科学的理论和大数据工程的方法均面临前所未有的问题。

定义3 大数据工程是用于处理大数据中固有的复杂性和应用需求的系统方法,涵盖大数据系统表示、获取、存储、组织、操作、搜索、分发、标准化、一致性和安全性。

由于大数据呈指数级增长的复杂性和需求,在大数据工程的各个方面和阶段都遇到了前所未有的挑战。这就要求对大数据工程进行理论和形式化研究,从而建立大数据科学。

定义4 大数据科学是关于大数据系统的原理、性质、理论、数学模型和方法论,以及大数据工程的有效组织技术的一门抽象科学。

根据定义1~定义4,基于对大数据本质的研究,大数据科学已经像系统科学、数学科学、软件科学、知识科学、智能科学和认知科学一样,成为当代抽象科学的一门新兴学科[35]。

本文探讨大数据作为人类推理和工程应用的基本认知对象的哲学、认知、计算和数学基础,以便揭示大数据的认知模型及其计算本质,诸如其类型属性、超结构模型和递归层次操作。

1 大数据科学的认知基础

数据是大脑中最基本的认知对象,因为它通过感觉和量化将现实世界的实体及其属性与抽象认知联系起来[1,15-16,19,23-24,31,36-42]。数据起源于对实体、属性及其关系的量化和抽象。复杂数据由算术、函数、分析、统计、数值和计算操作生成。数据的演绎分析和归纳综合导致了信息、知识和智能,体现了大数据系统的认知价值。

1.1 数据的认知模型

根据定义1,数据是抽象世界中无形的对象。它们是现实实体的属性或关系及其符号表示的抽象量化。数据的实质曾在概念代数[19]中研究如下。

定义5 数据的认知模型C_d,可被形式化地描述为知识K背景下的属性(A)、对象(O)及其与外部概念相对的内部(R^e)、输入(R^i)、输出(R^o)关系的一个超结构,即

$$C_d(\text{数据}: A,\ O,\ R^e,\ R^i,\ R^o)=\begin{cases}A=\{cognitive\_object^*,\ abstraction,\ quantity,\ fact,\ figure,\ unit\}\\ O=\{sensory_d,\ observed_d,\ exprimental_d,\ measured_d,\ typed_d,\ derived_d,\ statistical_d\}\\ R^e=O\times A\\ R^i\subseteq\ \times C_d\\ R^o\subseteq C_d\times\end{cases}\tag{1}$$

在上述数据的形式概念模型中,主属性*cognitive_objects*表示C_d(数据)所属的概括概念。A中的其他属性表示该概念的蕴含。O表示该概念的一组外延或对象。形式概念C_d(数据)的认知单位$\left|\ (C_d)\right|$由其内部关系的尺度决定,即$\left|\ (C_d)\right|=\left|R^c\right|=|O|\times|A|=7\times6=42$bir,其中形式概念的认知单位可由二元关系(binary relation,bir)度量[5]。

二元关系简称为*bir*(比尔),其作为知识科学的基本理论[5]和知识的基本单位被Wang在2018年发现[5]。值得注意的是,它与数据和信息的单位bit(比特)[34]有着根本的不同。

1.2　数据作为现实实体和抽象属性的量化

数据的普遍性建立在人类认知过程、形式推理、系统量化、数学运算、计算操作和信息处理的普遍性之上。数据如表1所示可以分为观察类、推断类、工程类和社会类[19-20]。

表1　数据的分类和起源

序号	类别/来源	类型	属性/特性
1.1	观察数据	事实	特征、序数、基数、计数、数量
1.2		状态	存在形式、构造、状态变化
1.3		行为	互动、规范、周期、分布、频率
2.1	推断数据	类比	相似、比较、等价类
2.2		关系	关联、因果、映射、序列、并发
2.3		测量	量化、限定、缩放、标准化、加权、分类
2.4		语义	物理、经验、抽象、数学、规则、性质
2.5		数学	线性、非线性、多项式、解析、微分、积分、细化、函数系统
2.6		统计	概率、范数、偏差、分布、条件、随机过程
2.7		复合	插值、外推、指数、幂函数、阶乘、笛卡尔积、搜索、排序、组合、排列、系统融合
3.1	工程数据	信息系统	信息技术、通信、计算、搜索引擎、多媒体、视频、图像采集
3.2		数字化	事实、统计、信号、媒体、电视、照片、音频、序列、比特流
3.3		数据库	数据中心、公共/工业/个人数据库、云服务器
3.4		知识库	一般、学科、学术、教育、道路、地球、宇宙
3.5		应用领域	工业、搜索引擎、智能系统、媒体、图像、视频、语音、文本、语言
4.1	社会数据	社交网络	商业、个人、学术、兴趣团体、意见收集
4.2		政府	行政、人口普查、投票、法律、图书馆
4.3		数据中心	全球、区域和城市中心、多对多分布中心

大数据的根本来源是人类的集体智慧。产生大数据的典型人类活动有多对多通信、海量数据复制下载、数字图像采集和网络舆情形成等。产生大数据的典型计算是笛卡尔积($O(n^2)$)、排序($O(n\lg n)$)、搜索(穷举,$O(n^2)$)、知识库更新($O(n^2)$),以及复杂度为$O(2^n)$、$O(n!)$或更高的诸如排列和NP问题[22,43-44]。

大数据的数学性质是对现实世界实体或认知对象的数量的抽象表示,通过量化映射到一定的尺度上。大数据也可能通过极简数学运算产生,比如实域内最大的数据是1/0=∞。创建大数据的典型数学运算有排列、组合、幂函数、阶乘、发散结构(如编码器和逻辑门等)。例如:(1) 欧拉数是10!=26525285981219105863630848000000;(2) 2016年已知的最大素数[45]为$2^{74207281}-1$;(3) 有10000个输入的与门的状态空间为2^{10000}比特,而其语义空间高达$2^{\lceil I_\kappa\rceil}(I_\kappa-1)=2^{\lceil 2^{10000}\rceil}(2^{10000}-1)\approx 2^{2^{10000}+10000}$ bir[35]。

定义6　数据的数学模型D是现实实体或认知对象的数量Q在度量尺度S上的抽象映射,其导出量是一个实数$D\in\mathbb{R}$,但其所载有的量纲是S,即

$$D\triangleq f_s:Q\to S \tag{2}$$

定义7　量化$\Gamma_S(X)$是将未知量X映射到给定度量尺度$\mathbf{S}$上的一个度量函数及其过程,即

$$\Gamma_S(X)\triangleq f_q:X\to\mathbf{S},\mathbf{S}\in\mathbb{N}$$

$$= \frac{Q(X)}{} [S] \quad (3)$$

式中,[S]表示量化单位或量纲。

引理1 相对于测量标度S,数量X所对应的数据$D(X, S)$是通过量化$\Gamma_S(X)$产生的实数$I_X \cdot R_X$。在单位[S]中,$I_X \cdot R_X$由用小数点分隔的两部分组成,其中I_X为整数,R_X为小于[S]的余数。

$$D(X,S) \triangleq \Gamma_S(X) = \frac{Q(X)}{}, S \in \mathbb{N}$$

$$= I_X \cdot R_X [S], I_X \in \mathbb{Z}, 0 \leqslant R_X \in \mathbb{R} < S \quad (4)$$

$$= \begin{cases} I_X = mod\ (X) \\ R_X = rem\ (X) \end{cases}$$

式中,数据对象由称之为(整数,余数)的量化对偶(I_X, R_X)表示,分别由模(*mod*)和余数(*rem*)操作生成。

例1 给定一实体近似长度为$L \approx 1286$ mm。根据引理1和式(4),与这一实体相应的数据可以在给定的测量尺度S上生成如下:

$$S_1 = 1\ \text{m}: \quad D_1 = \frac{Q(L)}{}_1 = \frac{Q(L)}{1\ [\text{m}]} = \frac{1286}{1000} = 1.286\ \text{m}$$

$$S_2 = 1\ \text{mm}: \quad D_2 = \frac{Q(L)}{}_2 = \frac{Q(L)}{1\ [\text{mm}]} = \frac{1286}{1} = 1286\ \text{mm}$$

$$S_3 = 1\ \mu\text{m}: \quad D_3 = \frac{Q(L)}{}_3 = \frac{Q(L)}{1\ [\mu\text{m}]} = \frac{1286}{0.001} = 1286000\ \mu\text{m}$$

式中,D_2表示测量标度S_2和量化标度S^0相同的特殊情况,其使量化数据D_2等于被测对象L。

1.3 数据作为认知和推理的基本对象

在人类认知的层次结构中,数据、信息、知识和智能之间的关系可以形式化地描述如下。

定义8 人类大脑中的认知对象H_{co}的层次结构是一个四元组,从下至上分为数据(𝔻)、信息(𝕀)、知识(𝕂)和智能($\dot{\mathbb{I}}$),即

$$H_{co} \triangleq (\mathbb{D}, \mathbb{I}, \mathbb{K}, \dot{\mathbb{I}})$$

$$= \begin{cases} \mathbb{D} = f_d : O \to Q \\ \mathbb{I} = f_i : D \to S \\ \mathbb{K} = f_k : I \to C \\ \dot{\mathbb{I}} = f_i : I \to B \end{cases} \quad (5)$$

式中符号分别表示对象(O)、数据(D)、数量(Q)、信息(I)、语义(S)、概念(C)和行为(B)。

基于定义8,数据作为人类最基本的认知对象之一具有一系列如文献[24]和[34]中所确定的20个基本认知属性,例如抽象性、概括性、累积性、对认知的依赖性、类型化的元语义、可共享性、无空间尺度、无重量、数据到信息/知识/智能之间的可转换性、多种表示形式、多种承载介质、多种传输形式、对介质的依赖性、对能量的依赖性、无磨损性、时间依赖性、熵守恒、不同于物理属性的信息质量属性、易失真和稀缺性。

2 大数据科学的计算基础

从计算的角度来看,数据的一个重要抽象属性是它们的类(type)或形态。在类型理论中[5,16],类决定了某一形态数据的域、单位和被允许的操作。一个类型系统指定了所有数据对象类的建模和操作规则[3,16,29,45]。在一给定完备类型系统中,任何数据对象都可以被指定到一个类型、一个有限类型集,或它们的组合。每种类都是一个集合,其中所有数据对象共享一组公共属性、域约束和预定义操作。类可以分为基本和复杂类型。前者是一组最简类型,其不能被进一步约减;而后者是由基本类型按照一定的类规则组合而成的混合类型。

定义9 数据对象O是一个有指定类$\mathbb{T}$的变量$v, v \in V \subset \overline{P}V \sqsubset U$,其由类型约束$\mu_v(\mathbb{T})$限制以便将此一般类$\mathbb{T}$通过子域裁剪$\mathbb{T}''$转化为一给定问题的特定类$\mathbb{T}'$,即

$$O \triangleq <v: \mathbb{T} \mid \mu_v(\mathbb{T}) = \mathbb{T} \setminus \mathbb{T}'' = \mathbb{T}'> \quad (6)$$

式中,$\mu_v(\mathbb{T}) = \mathbb{T} \setminus \mathbb{T}'' = \mathbb{T}'$将给定数域$D$中的$\mathbb{T}$裁剪成问题域$D_p$中的子集$\mathbb{T}'$,即$v|\mathbb{T}' = v|\mathbb{T} \setminus v|\mathbb{T}''$。大数据的论域$U$将在定义14中详述。

定义10 数据对象的基本类型T_p是一组完备的且在不丢失逻辑或语义属性的限制下不可再分解的数据类,如下所示:

$$T_p \triangleq \{\text{N, Z, R, S, L, B, H, P}\} \quad (7)$$

式中的符号分别表示诸如自然数(N)、整数(Z)、实数(R)、字符串(S)、逻辑变量(L)、字节(B)、十六进制数(Hx)和指针(P)的类型。

表2不但给定了数据的基本类型,而且确定了它们的表示法、域和属性。表2中的每一个类域都可以用其数学域(𝕋)和问题域(𝕋′)表示,其中问题域𝕋′受给定问题或机器内存(M_{max})空间限制。

表2　数据对象的基本类和域

序号	类型	语法	域	等效性
1	自然数	N	N∈[0, 65535], N′∈[0, M_{max}−1]	默认的算术运算
2	整数	Z	Z∈[−32768, +32767] Z′∈[−(M_{max}/2), +(M_{max}/2)−1]	
3	实数	R	R∈[−2147483648, 2147483647] R′∈[−(M_{max}/2), +(M_{max}/2)−1]	
4	字符串	S	S∈[0, 255] S′∈[0, M_{max}−1]	默认的字符和字符串操作
5	逻辑变量	L	L∈[T, F]	布尔常量 [T|L, F|L]
6	字节	B	B∈[0, 255], B′∈[0, 1023]	默认的二进制操作
7	16进制	H	H∈[0, 65535], H′∈[0, M_{max}−1]	

虽然十进制系统广泛用于面向人的数据表示,但最精细的数据类是比特(bit)[13],这是计算机与信息科学融合的基础。复杂的数据类型和系统可以基于比特位进行组合。然而最新发现表明,知识科学与智能科学的基本单位是二进制关系比尔(bir)[5],这种重要的类为知识科学和智能科学提供了理论基础。

定义11　数据对象的复杂类型T_c包含一组由基本类型派生的组合类型,即

$$T_c \triangleq \{\mathbb{T}, SM, TX, A, F, V, T, D, TM, Bir\} \quad (8)$$

式中的符号分别表示诸如任意(𝕋,在运行时确定)、结构模型(SM)、文本(TX)、音频(A)、照片(F)、视频(V)、时间(T)、日期(D)、日期-时间(TM),和知识(Bir)的复杂类型。其中Bir是基于单位bir的知识类型。

复杂数据对象的组合类型在表3中给出,用以指定了它们的类、符号、域和属性。在表3中有一些复杂类型是问题相关的,比如{TX, A, F, V}。为方便数据表示和定性,在本文数据对象x的组合形式中采用了一种类型后缀约定,其由竖线分隔,即$x|\mathbb{T}$,其中$|\mathbb{T}$可以由表2、表3中的任何类型或基于它们的用户定义类型来表示。

表3　数据对象的复杂类和域

序号	类型	符号	域	等效性
1	任意类型	𝕋	–	运行时确定的任何虚拟类型
2	结构	SM	SM∈$\tau^n\|T$(Eq. 11)|	默认字段引用:x|SM.y|T
3	文本	TS=S′	S′∈[0, M_{max}−1]	问题相关
4	音频	A=B′×T	B′∈[0, 1023], TI∈[hh:mm:ss:ms]	
5	照片	F=B′×B′	B′∈[0, 1023]	
6	视频	V=B′×B′×T	B′∈[0, 1023], TI∈[hh:mm:ss:ms]	
7	时间	T=hh:mm:ss:ms	hh∈[0, 23], mm∈[0, 59], ss∈[0, 59], ms∈[0, 999]	默认的时间操作
8	日期	D=YY:MM:DD	YY∈[0, 99], MM∈[1, 12], DD∈[1, 31]	
9	日期/时间	TM=YYYY:MM:DD:hh:mm:ss:ms	YYYY∈[0, 9999], MM∈[1, 12], DD∈[1, 31], hh∈[0, 23], mm∈[0, 59], ss∈[0, 59], ms∈[0, 999]	
10	知识	Bir	[0, ∞)	数据的语义

定义12　大数据的类型系统 T 是表2、表3所示的17种基本类型和复杂类型的集合,即

$$\begin{aligned} T &\triangleq T_p \cup T_c \\ &= \{N, Z, R, S, L, B, H, P\} \\ &\cup \{\mathbb{T}, SM, TX, A, F, V, T, D, TM, Bir\} \end{aligned} \quad (9)$$

例2　大数据应用中任何用户定义的类型都可以通过基本类型、复杂类型及其组合在类型环境T中进行形式化描述。例如,多媒体社交网络中的典型大数据对象类型,如文本(电子信件、消息)、语音、照片、视频、知识等均可严格定义如下:

$$\begin{aligned} &Text|TX: && TX \triangleq S \\ &Voice|A: && A \triangleq B \times T \\ &Photo|F: && F \triangleq B \times B \end{aligned} \quad (10)$$

Video|V: V ≜ B×B×T

Knowledge|Bir: K ≜ C×C

式中,|TX, |A, |F, |V 和 |Bir是复杂类型或用户定义类型的后缀, 然而C是一个如定义5所示的认知概念。

大数据科学的另一发现揭示了大数据的形式载体是一种称为类型化元组的超结构[19]。

定义13 数据和数量的超结构类型是一个通用的类型化的n元组,$\tau^n|T$,其在满足一定约束条件$\mathbb{T}_i'$下,将n个数据对象封装在异构类$\mathbb{T}_i(1\leqslant i\leqslant n)$的超结构中,即

$$\tau^n|T\triangleq\left(\mathop{R}_{i=1}^{n}O_i|\mathrm{SM}\right)=\left(\mathop{R}_{i=1}^{n}<v_i:\mathbb{T}_i'\mid\mathbb{T}_i'=\mathbb{T}_i\backslash\mathbb{T}_i''>\right)\tag{11}$$

式中,$\mathop{R}_{i=1}^{n}S_i$表示一个循环结构或迭代行为[16,46];$\mathbb{T}_i'$是对标准类型$\mathbb{T}_i$的一个类型约束。

数域的类型约束广泛用于指定用户定义域或特定于问题的类型,其是大数据建模中用于构造或细化数据类型的一般方法。

3 大数据科学的数学基础

数据是将实体的属性映射到一个度量尺度的量化的结果。大数据是在科学、工程和社会学科中产生的大规模、异构数据的超结构复杂系统。本节将在上述概念模型的基础上, 建立一组大数据的严格数学模型。

3.1 大数据系统的论域

大数据科学的数学结构的整体性质和公理可以通过对大数据系统的抽象模型的归纳而得出。该形式化方法从研究抽象数据和大数据系统的论域开始。

定义14 大数据与复杂抽象数量的论域U是一个六元组:

$$U\triangleq(E,\ T,\ Q,\ R,\ V,\ H)\tag{12}$$

式中,E是一实体和/或其可测量属性集;T是E的类型或性质集;Q是E上的一量化标度集;R是一量化关系($Q\times E$)或限定关系($E\times T$)集;V是一类型化的值集$Q\times E\times T\rightarrow V|T$;$H$是一超结构集$E\times V\times T$。

论域U的定义揭示一般大数据比定义2中给出的简单数据更加复杂。在U的基础上,将形式化引出大数据系统的数学模型,分为基本和一般大数据模型。

3.2 大数据系统的数学模型

大数据系统的数学模型可以由引理1和定义14导出。在简单数据单维结构的基础上,基本大数据系统可被描述为二维类型化超结构;而一般大数据系统将被定义为多维类型化超结构。

定义15 基本大数据系统(BDS)的数学模型$\Theta^2=\mathop{R}_{i=0}^{n}\mathop{R}_{j=0}^{m}d_{ij}|\mathbb{T}_{0j}$是论域$U$中的一个二维$n\times m$类型化超结构,其中数据元素$d_{ij}|\mathbb{T}_{0j}$由类型后缀$|\mathbb{T}_{0j}$指定,可以是任意一种定义在$T$中的初级或复杂类型,即

$$\Theta^2\triangleq\mathop{R}_{i=0}^{n}\mathop{R}_{j=0}^{m}d_{ij}|\mathbb{T}_{0j}=\begin{bmatrix}\theta_0 & e_1|\mathbb{T}_{01} & e_2|\mathbb{T}_{02} & \cdots & e_m|\mathbb{T}_{0m}\\ \kappa_1 & d_{11}|\mathbb{T}_{01} & d_{12}|\mathbb{T}_{02} & \cdots & d_{1m}|\mathbb{T}_{0m}\\ \kappa_2 & d_{21}|\mathbb{T}_{01} & d_{22}|\mathbb{T}_{02} & \cdots & d_{2m}|\mathbb{T}_{0m}\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ \kappa_n & d_{n1}|\mathbb{T}_{01} & d_{n2}|\mathbb{T}_{02} & \cdots & d_{nm}|\mathbb{T}_{0m}\end{bmatrix}\tag{13}$$

式中,$\mathop{R}_{i=0}^{n}\kappa_i|SM$中的每一行表示一个称作为结构模型$(|SM)$的类型化元组,而$\mathop{R}_{j=0}^{m}e_j|\mathbb{T}_{0j}$中的每一列代表一个具有特定类型$(|\mathbb{T}_{0j})$的域[47],其中大$R$算符$\mathop{R}_{i=0}^{n}X_i$用以表示循环结构或重复行为[46]。

如定义15所示,大数据Θ^2数学模型中的第0行θ_0是一个特殊的类型化元组,称为BDS的模式。模式θ_0指定BDS各数据域的结构和约束类。值得注意的是,BDS数学模型中采用的类后缀$|\mathbb{T}$可用于定义任意类型的大数据,以便满足对于容纳广泛异构数据的需求。

例3 给定一个有100万用户和7个数据域的社交网络大数据系统BDS_1。其二维大数据模型的

形式结构 $\Theta^2(BDS_1)$ 可以根据式(13)严格描述如下:

$$\theta_0(BDS_1)=(ID|\mathrm{N},\ UName|\mathrm{S},\ GName|\mathrm{S},\ Text|\mathrm{T},\ Voice|\mathrm{A},\ Photo|\mathrm{F},\ Video|\mathrm{V})$$

式中,大数据系统模式 $\Theta^2(BDS_1)$ 的7个域分别代表识别号(lN)、用户名(lS)、组名(lS)、文本数据(lT)、语音数据(lA)、照片数据(lF)和视频数据(lV)。$\Theta^2(BDS_1)$ 大数据中如式(14)所示的每一记录都受公共模式 $\theta_0(BDS_1)$ 的约束。

$$\Theta^2(BDS_1)=\mathop{R}_{i=0}^{1000000}\mathop{R}_{j=0}^{7}d_{ij}|\mathbb{T}_j$$

$$=\begin{bmatrix}\theta_0(BDS_1) & ID|\mathrm{N} & UName|\mathrm{S} & GName|\mathrm{S} & Text|\mathrm{T} & Voice|\mathrm{A} & Photo|\mathrm{F} & Video|\mathrm{V}\\ \kappa_1 & 0000001 & John & G_{0001} & R_t0000001|\mathrm{T} & R_a0000001|\mathrm{A} & R_p0000001|\mathrm{F} & R_v0000001|\mathrm{V}\\ \kappa_2 & 0000002 & Judy & G_{0301} & R_t0000002|\mathrm{T} & R_a0000002|\mathrm{A} & R_p0000002|\mathrm{F} & R_v0000002|\mathrm{V}\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\ \kappa_{1000000} & 1000000 & Mike & G_{1806} & R_t1000000|\mathrm{T} & R_a1000000|\mathrm{A} & R_p1000000|\mathrm{F} & R_v1000000|\mathrm{V}\end{bmatrix}$$

(14)

式(13)所定义的大数据系统基本模型 Θ^2 可扩展到一般的 n 维类型化超结构大数据系统 Θ^n。

定义16　大数据系统(BDS)的一般数学模型为 $\Theta^q=\mathop{R}_{i_1=0}^{n_1}\mathop{R}_{i_2=0}^{n_2}\cdots\mathop{R}_{i_q=0}^{n_q}d_{i_1i_2\cdots i_q}|\mathbb{T}_{i_1i_2\cdots i_q}$。该模型在U中任何 q 维异构数据可被形式化描述成一个递归的类型化超结构(RTHS),即

$$\Theta^q\triangleq\mathop{R}_{k=q}^{1}\Theta^k(\Theta^{k-1})$$

$$=\mathop{R}_{i_1=0}^{n_1}\mathop{R}_{i_2=0}^{n_2}\cdots\mathop{R}_{i_q=0}^{n_q}d_{i_1i_2\cdots i_q}|\mathbb{T}_{i_1i_2\cdots i_q}$$

$$=\mathop{R}_{k=q}^{1}\begin{bmatrix}\theta_0^k & e_1^k|\mathbb{T}_{01}^k & e_2^k|\mathbb{T}_{02}^k e_2 & \cdots & e_m^k|\mathbb{T}_{0m}^k & \theta_0^{k-1}|\mathrm{P}\\ \kappa_1^k & \tau_{11}^k|\mathbb{T}_{01}^k & \tau_{12}^k|\mathbb{T}_{02}^k & \cdots & \tau_{1m}^k|\mathbb{T}_{0m}^k & \kappa_1^{k-1}|\mathrm{P}\\ \kappa_2^k & \tau_{21}^k|\mathbb{T}_{01}^k & \tau_{22}^k|\mathbb{T}_{02}^k & \cdots & \tau_{2m}^k|\mathbb{T}_{0m}^k & \kappa_2^{k-1}|\mathrm{P}\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\ \kappa_n^k & \tau_{m1}^k|\mathbb{T}_{01}^k & \tau_{m2}^k|\mathbb{T}_{02}^k & \cdots & \tau_{mn}^k|\mathbb{T}_{0m}^k & \kappa_m^{k-1}|\mathrm{P}\end{bmatrix}$$

(15)

式中,$\mathop{R}_{j=1}^{m}\kappa_j^k|\mathrm{SM}=(\tau_1^k|\mathbb{T}_{01}^k,\tau_2^k|\mathbb{T}_{01}^k,\ldots,\tau_m^k|\mathbb{T}_{01}^k)$ 表示通过指针 $\theta_0^{k-1}|\mathrm{P}$ 递归链接到大数据系统 $\mathop{R}_{i=1}^{n}\kappa_i^{k-1}|\mathrm{SM}$ 中低层结构的一个类型化元组。

通过对比定义15和定义16,可见二维基本大数据模型 Θ^2 是一般大数据模型 Θ^n 的一个特例。一般大数据模型 $\Theta^q(BDS)$ 中的元素除了终端层外都是一个类型元组,而基本大数据模型 $\Theta^2(BDS)$ 中的元素是一个终端数据对象。因此,根据定义16,任何前者可以通过有限步转变为后者,如果 $\Theta^0(BDS)$ 已给定。

例4　在例3的基础上,建立一个给定社交网络 $\Theta^3(BDS_2)$ 的一般大数据模型,其中 $\Theta^0=\mathop{R}_{i=0}^{1000000}\mathop{R}_{j=0}^{4}d_{ij}|\mathbb{T}_{0j}$。按照定义16,该问题的解是将 $\Theta^2(BDS_1)$ 分层细化为一个3层一般大数据模型,如式(16)所示。

$$\Theta^3(BDS_2)=\mathop{R}_{k=3}^{1}\Theta^k(\Theta^{k-1}),\ \Theta^0=\mathop{R}_{i=0}^{1000000}\mathop{R}_{j=0}^{4}d_{ij}|\mathbb{T}_{0J}$$

$$=\mathop{R}_{i_0=0}^{n_0}\mathop{R}_{i_1=0}^{n_1}\mathop{R}_{i_2=0}^{n_2}d_{i_0i_1i_2}|\mathbb{T}_{i_0i_1i_2}$$

$$=\Theta^3(BDS_2):\begin{bmatrix}\theta_0^3(BDS_2) & ID|\mathrm{N} & UName|\mathrm{S} & GName|\mathrm{S} & \theta_0^2|\mathrm{P}\\ \kappa_1^3 & 0000001 & John & G_{0001} & \kappa_1^2|\mathrm{P}\\ \kappa_2^3 & 0000002 & Judy & G_{0301} & \kappa_2^2|\mathrm{P}\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ \kappa_{1000000}^3 & 1000000 & Mike & G_{1806} & \kappa_{1000000}^2|\mathrm{P}\end{bmatrix}$$

$$\Downarrow\Theta^2(BDS_2):\begin{bmatrix}\theta_0^2(BDS_2) & Rec\#|\mathrm{N} & TStamp|\mathrm{TM} & Size|\mathrm{B} & \theta_0^1|\mathrm{P}\\ \kappa_1^2 & 0000001 & TM_1 & S_1 & \kappa_1^1|\mathrm{P}\\ \kappa_2^2 & 0000002 & TM_2 & S_2 & \kappa_2^1|\mathrm{P}\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ \kappa_{1000000}^2 & 1000000 & TM_{1000000} & S_{1000000} & \kappa_{1000000}^1|\mathrm{P}\end{bmatrix}$$

$$\Downarrow\Theta^1(BDS_2):\begin{bmatrix}\theta_0^1(BDS_2) & Text|\mathrm{T} & Voice|\mathrm{A} & Photo|\mathrm{F} & Video|V\\ \kappa_1^1 & R_t0000001|\mathrm{T} & R_a0000001|\mathrm{A} & R_p0000001|\mathrm{F} & R_v0000001|V\\ \kappa_2^1 & R_t0000002|\mathrm{T} & R_a0000002|\mathrm{A} & R_p0000002|\mathrm{F} & R_v0000002|V\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ \kappa_{1000000}^1 & R_t1000000|\mathrm{T} & R_a1000000|\mathrm{A} & R_p1000000|\mathrm{F} & R_v1000000|V\end{bmatrix} \tag{16}$$

在 $\Theta^3(BDS_2)$ 中，每一层自顶向下通过指针 $\theta_0^{k-1}|\mathrm{P}$ 链接到下一层，从而任一 k 层都将被递归地细化（⇓）为一个基本二维结构。$\Theta^3(BDS_2)$ 的顶层设定为 $\theta_0^3(BDS_2)$，并通过 $\theta_0^2|\mathrm{P}$ 将其链接到下一层的二维结构 $\theta_0^2(BDS_2)$。第二层对 $\theta_0^2(BDS_2)$ 进行细化，并通过 $\theta_0^1|\mathrm{P}$ 将其链接到下一层的二维结构 $\theta_0^1(BDS_2)$。然后，第一层将 $\Theta^1(BDS_2)$ 表示为在第0层给出的一组终端数据对象。

上述定义所给大数据系统的一般数学模型（Θ^q）可以用递归层次结构清晰严格地表达任意 q 维大数据结构，其中每一层都是一个统一的二维基本子结构（Θ^2）。例4显示大数据系统 Θ^q 不再是简单的纯数字一维或二维结构，而是一个递归类型化的超结构（RTHS），其可分解为一组层次化链接的二维超结构 Θ^2。

大数据系统的一般数学模型（Θ^q），不但揭示了大数据的递归超结构的数学实质，而且为现实世界中任何复杂大数据系统的建模和分析提供了一种严格、通用、灵活和高效的层次细化和综合递归方法。该理论也为大数据在超R域中的数学分析（简称大数据代数[19]）奠定了基础。

4 大数据系统的性质

4.1 大数据系统的递归超结构特性

定理1 大数据系统的一般递归超结构性质可由一个 q 维层次结构 Θ^q 形式化地表述。当最底层终端结构 Θ^0 已给定为 m 个数据对象的类型化元组，任一高层结构可以由低层结构递归得到，

$$\Theta^q=\mathop{R}_{k=1}^{q}\Theta^k\left(\Theta^{k-1}\right),\qquad \Theta^0=\mathop{R}_{j=0}^{m}d_j|\mathbb{T}_j \tag{17}$$

证明 定理1可以通过定义15、定义16归纳证明如下：

$$\begin{aligned}&\forall\,\Theta^0=\mathop{R}_{j=0}^{m}d_j|\mathbb{T}_j,\\ &\Theta^q=\begin{cases}\Theta^1=\Theta^0=\left(\mathop{R}_{j=0}^{m}d_j|\mathbb{T}_j\right)\\ \Theta^2=\Theta^1\left(\Theta^0\right)=\Theta^1\left(\mathop{R}_{j=0}^{m}d_j|\mathbb{T}_j\right)\\ \cdots\\ \Theta^q\left(\Theta^{q-1}\left(\cdots\left(\Theta^1\left(\mathop{R}_{j=0}^{m}d_j|\mathbb{T}_j\right)\right)\cdots\right)\right)\end{cases}\\ &=\Theta^q\left(\Theta^{q-1}\left(\cdots\left(\Theta^1\left(\Theta^0\right)\right)\cdots\right)\right)\\ &=\mathop{R}_{k=1}^{q}\Theta^k\left(\Theta^{k-1}\right)\end{aligned} \tag{18}$$

上述定理是大数据科学基于最新系统科学原理[35]所得出的另一发现。在定理1的基础上，可以推导出如下BDS层次细化和层次抽象的原则。

推论1 在U中任何一个BDS都可以通过自上而下的低层结构以不断增加的细节来演绎和分析，即

$$\Theta^q=\mathop{\Downarrow}_{k=q}^{0}\Theta^k=\mathop{R}_{k=q}^{1}\Theta^k\left(\Theta^{k-1}\right),\qquad \Theta^0=\mathop{R}_{i=0}^{n}\mathop{R}_{j=0}^{m}d_{ij}|\mathbb{T}_j \tag{19}$$

逆对称地，推论1可以表示为如下层次抽象原则。

推论2 在U中任何一个BDS都可以通过自下而上的结构归纳和综合而合成，即

$$\Theta^q=\mathop{\Uparrow}_{k=0}^{q}\Theta^k=\mathop{R}_{k=1}^{q}\Theta^k\left(\Theta^{k-1}\right),\quad \Theta^0=\mathop{R}_{i=0}^{n}\mathop{R}_{j=0}^{m}d_{ij}\left|\mathbb{T}_j\right. \tag{20}$$

该推论可以看作是对被称为“信息隐蔽”的系统建模经验原理的形式化表达。

证明　推论1和2可引用定理1证明。

推论3　一般大数据库(BDB)的数学模型 $\overset{a}{\uplus}$ 是根据定理1建立的递归超结构类型化数据库,其二维和n维模型如下:

$$\overset{a}{\uplus}\triangleq\begin{cases}\Theta^2=\mathop{R}_{i_1=0}^{m_1}\mathop{R}_{i_2=0}^{m_2}d_{i_1i_2}\left|\mathbb{T}_{i_1i_2}\right.\\ \Theta^n=\mathop{R}_{i_1=0}^{m_1}\mathop{R}_{i_2=0}^{m_2}\cdots\mathop{R}_{i_n=0}^{m_n}d_{i_1i_2\cdots i_n}\left|\mathbb{T}_{i_1i_2\cdots i_n}\right.\end{cases} \tag{21}$$

根据定理1和推论1~推论3,任何一般的n维大数据系统都可以由低阶的2维系统递归而成。反之,任何基本的二维大数据系统都可以由一般的n维系统演绎而成。

4.2　数据、信息、知识和智能之间的关系

虽然大数据的基本属性由递归超结构和类型理论主导,但信息、知识和智能的基本属性分别由组合机制、形式概念和超函数构成。因此,在人类认知的层次结构中,数据、信息、知识和智能之间的关系可以通过以下原理来形式化描述。

定理2　在认知系统S_{co}的层次结构中,认知对象之间的可转换性形成了数据($\mathbb{D}$)、信息($\mathbb{I}$)、知识($\mathbb{K}$)、智能($\dot{\mathbb{I}}$)之间的递归结构如下:

$$\begin{aligned}S_{co}&=\mathop{R}_{k=4}^{1}O^k\left(O^{k-1}\right),\ O^0=\mathop{R}_{i=0}^{n}d_i\left|\mathbb{T}_i\right.\\&=O^4\left(O^3\left(O^2\left(O^1\left(O^0\right)\right)\right)\right)\\&=\dot{\mathbb{I}}\left(\mathbb{K}\left(\mathbb{I}\left(\mathbb{D}\left(\mathop{R}_{i=0}^{n}d_i\left|\mathbb{T}_i\right.\right)\right)\right)\right)\end{aligned} \tag{22}$$

其中层次框架的终端层由具有特定类型的n维数据对象 $\mathop{R}_{i=0}^{n}d_i\left|\mathbb{T}_i\right.$ 给定。

证明　根据定理1,大数据系统的认知转换结构S_{co}可被递归归纳地证明如下:

$$\begin{aligned}&\forall O^0=\mathop{R}_{i=0}^{n}d_i\left|\mathbb{T}_i\right.\\&O^1=O^1\left(O^0\right)=O^1\left(\mathop{R}_{i=0}^{n}d_i\left|\mathbb{T}_i\right.\right)=f_d\left(\mathop{R}_{i=0}^{n}d_i\left|\mathbb{T}_i\right.\right)\\&\quad=\mathbb{D}\left(\mathop{R}_{i=0}^{n}d_i\left|\mathbb{T}_i\right.\right)\end{aligned}$$

$$\begin{aligned}&O^2=O^2\left(O^1\right)=O^2\left(\mathbb{D}\right)=f_i\left(f_d\left(\mathop{R}_{i=0}^{n}d_i\left|\mathbb{T}_i\right.\right)\right)\\&\quad=\mathbb{I}\left(\mathbb{D}\left(\mathop{R}_{i=0}^{n}d_i\left|\mathbb{T}_i\right.\right)\right)\\&O^3=O^3\left(O^2\right)=O^3\left(\mathbb{I}\right)=f_k\left(f_i\left(f_d\left(\mathop{R}_{i=0}^{n}d_i\left|\mathbb{T}_i\right.\right)\right)\right)\\&\quad=\mathbb{K}\left(\mathbb{I}\left(\mathbb{D}\left(\mathop{R}_{i=0}^{n}d_i\left|\mathbb{T}_i\right.\right)\right)\right)\\&O^4=O^4\left(O^3\right)=O^4\left(\mathbb{K}\right)=f_l\left(f_k\left(f_i\left(\mathop{R}_{i=0}^{n}d_i\left|\mathbb{T}_i\right.\right)\right)\right)\\&\quad=\dot{\mathbb{I}}\left(\mathbb{K}\left(\mathbb{I}\left(\mathbb{D}\left(\mathop{R}_{i=0}^{n}d_i\left|\mathbb{T}_i\right.\right)\right)\right)\right)\\&\Rightarrow S_{co}=O^4\left(O^3\left(O^2\left(O^1\left(O^0\right)\right)\right)\right)\\&\quad=\dot{\mathbb{I}}\left(\mathbb{K}\left(\mathbb{I}\left(\mathbb{D}\left(\mathop{R}_{i=0}^{n}d_i\left|\mathbb{T}_i\right.\right)\right)\right)\right)\end{aligned} \tag{23}$$

其中每一给定层次对象都是其直接下层对象的合成。

上述定理2揭示了认知对象之间的普遍可转换性和对人类归纳能力的高度依赖,该定理不但涵盖数据的抽象/量化、信息获取、知识生成和大脑智能/智慧创造,而且也表明大数据的语义推导和归纳是复杂代数运算,而不是通常认为的逻辑运算。定理2不但解释了低层认知对象是高层认知对象的一个不可或缺的基础和范例,而且解释了高层认知对象是低层认知对象的一个升华和融合。因此,对于任一给定层次的认知对象,其不仅是低层认知对象的语义,也是高层认知对象的实例。

例5　一个具有1000个输入、n=1000的逻辑与门的状态空间所能产生的大数据,$O^0=\mathop{R}_{i=0}^{n-1}d_i\left|\mathbb{T}_i\right.=\mathop{R}_{i=0}^{999}d_i\left|\mathrm{B}\right.$,可根据定理2确定为

$$\begin{aligned}&O^1=\mathbb{D}\left(O^0\right)=\mathbb{D}\left(\mathop{R}_{i=0}^{999}d_i\left|\mathrm{B}\right.\right)=2^{1000}\left|\mathrm{B}\right.\\&O^2=\mathbb{I}\left(\mathbb{D}\right)=\mathbb{I}\left(\mathbb{D}\left(\mathop{R}_{i=0}^{999}d_i\left|\mathrm{B}\right.\right)\right)=1000\left|\mathrm{B}\right.\\&O^3=\mathbb{K}\left(\mathbb{I}\right)=\mathbb{K}\left(\mathbb{I}\left(\mathbb{D}\right)\right)=\{\mathrm{rule}_1,\ \mathrm{rule}_2\}\\&\begin{cases}\mathrm{rule}_1\left|\mathrm{Bir}:\forall\mathop{R}_{i=0}^{999}d_i\right|\mathrm{B}=1\Rightarrow AND\left(\mathop{R}_{i=0}^{999}d_i\left|\mathrm{B}\right.\right)=1\\ \mathrm{rule}_2\left|\mathrm{Bir}:\exists\mathop{R}_{i=0}^{999}d_i\right|\mathrm{B}=0\Rightarrow AND\left(\mathop{R}_{i=0}^{999}d_i\left|\mathrm{B}\right.\right)=0\end{cases}\end{aligned} \tag{24}$$

$$O^4=\dot{\mathbb{I}}(\mathbb{K})=\dot{\mathbb{I}}\left(\mathbb{K}\left(\mathbb{I}(\mathbb{D})\right)\right):\forall n, AND\left(\overset{n}{\underset{i=1}{R}}d_i|\mathrm{B}\right)=\overset{n}{\underset{i=1}{\wedge}}d_i|\mathrm{B}$$

例5表明,在对一个认知对象在不同归纳层次上的量化描述,在数据科学的意义上会有极大的不同。在底部的数据层(O^1),所给逻辑与门表示一个具有2^{1000}比特状态空间的超大数据系统。然而,在信息层(O^2),它只表示1000比特的信息,即它的输入尺度。在知识层(O^3),它被归纳为两个Bir的规则,即当所有输入为1时,输出为1;而当任一输入为0时,输出为0。最终,在智能层(O^4),无论输入的尺度(n)是多大,此逻辑与门已升华为与n无关的一个通用逻辑与函数。

定理2和例5揭示了人类智能是一种非常高效、快速收敛的知识生成和智慧获取的归纳机制,其中数据仅仅是现实世界几乎无限状态空间中的事实材料和个体实例。归纳的认知过程是大数据转化为知识和智慧的一种典型的、人类智能中最强大的推理能力。

6 结论

介绍了一个从认知、数学和计算3个领域对大数据科学的基础研究。发现了大数据系统是一种新型的超越R域的递归类型化超结构(RTHS),创建了大数据系统(BDS)的基本数学模型($\Theta^n(BDS)$)和大数据科学的数学方法。研究揭示了大数据系统数据的抽象/量化、信息获取、知识生成和智能/智慧创生,均高度依赖于归纳推理机制,为现实世界中任何复杂大数据系统的建模和分析提供了一种通用和高效的层次细化方法。本文对大数据原理与性质的研究,不仅为解释BDS的本质提供了一个新颖的数学结构,也为大数据科学在大数据工程、机器学习和知识提取等新兴领域的广泛应用提供了一种严格的理论基础。

致谢:本文工作获得加拿大自然科学与工程研究委员会(NSERC)发现基金、清华大学大数据系统软件国家工程重点实验室及重庆科技学院科研基金的部分资助。陆建华院士和孙家广院士对本文撰写提供了精心指导,王建民和刘琳对专题文章的组织给予了帮助,魏子麒对文本编辑提供了帮助。

参考文献(References)

[1] Hassanien A E, Azar A T, Snasel V, et al. Big data in complex systems: Challenges and opportunities[M]. Berlin: Springer, 2015.

[2] Jacobs A. The pathologies of big data[J]. Queue, 2009, 7 (6): 10.

[3] Mitchell J C. Type systems for programming languages [M]//van Leeuwen J. Handbook of Theoretical Computer Science. Amsterdam: Elsevier, 1990: 365-458.

[4] Wang Y. Software science: On general mathematical models and formal properties of software[J]. Journal of Advanced Mathematics and Applications, 2014, 3(2): 130-147.

[5] Wang Y. On cognitive foundations and mathematical theories of knowledge science[J]. International Journal of Cognitive Informatics and Natural Intelligence, 2016, 10(2): 1-24.

[6] Wang Y. Keynote: On the emergence of abstract sciences and breakthroughs in machine knowledge learning[C]// 18th IEEE International Conference on Cognitive Informatics and Cognitive Computing (ICCI*CC 2019). Piscataway N J: IEEE Press, 2009: 5.

[7] Wang Y. Keynote: The cognitive and mathematical foundations of big data science and blockchain engineering [C]//International Conference on Big Data and Blockchain (ICBDB'19). Piscataway N J: IEEE Press, 2009: 4.

[8] Bender E A. Mathematical methods in artificial intelligence[M]. Los Alamitos: IEEE CS Press, 1996.

[9] Berkeley B. Principles of human knowledge[M]. London: Berkeley, 1954.

[10] Debenham J K. Knowledge systems design[M]. New York: Prentice Hall, 1989.

[11] McCarthy J, Minsky M L, Rochester N, et al. Proposal for the 1956 dartmouth summer research project on artificial intelligence[R/OL]. [2019-10-31]. http://www.formal.stanford.edu/jmc/history/dartmouth/dartmouth.html.

[12] McCulloch W S. Embodiments of mind[M]. Cambridge: MIT Press, 1965.

[13] Shannon C E. A mathematical theory of communication [J]. The Bell System Technical Journal, 1948, 27: 379-423, 623-656.

[14] Turing A M. Computing machinery and intelligence[J].

Mind, 1950, 59: 433-460.
[15] von Neumann J. The computer and the brain[M]. New Haven: Yale University Press, 1958.
[16] Wang Y. Software engineering foundations: A software science perspective[M]. New York: Auerbach Publications, 2007.
[17] Wang Y. On Abstract Intelligence: Toward a unified theory of natural, artificial, machinable, and computational intelligence[J]. International Journal of Software Science and Computational Intelligence, 2009, 1(1): 1-17.
[18] Wang Y. Cognitive robots: A reference model towards intelligent authentication[J]. IEEE Robotics and Automation, 2010, 17(4): 54-62.
[19] Wang Y. Keynote: Big data algebra: A rigorous approach to big data analytics and engineering[C]//17th International Conference on Mathematical and Computational Methods in Science and Engineering (MACMESE´15). Kuala Lumpur, 2015: 2.
[20] Wang Y. Concept algebra: A denotational mathematics for formal knowledge representation and cognitive robot learning[J]. Journal of Advanced Mathematics and Applications, 2015, 4(1): 62-87.
[21] Wang Y. In search of denotational mathematics: Novel mathematical means for contemporary intelligence, brain, and knowledge sciences[J]. Journal of Advanced Mathematics and Applications, 2012, 1(1): 4-25.
[22] Lewis H R, Papadimitriou C H, Elements of the theory of computation[M]. New York: Prentice Hall, 1998.
[23] Wang Y. On the informatics laws and deductive semantics of software[J]. IEEE Transactions on Systems, Man, and Cybernetics (Part C), 2006, 36(2): 161-171.
[24] Zadeh L A. Fuzzy logic and approximate reasoning[J]. Synthese, 1975, 30(3/4): 407-428.
[25] Wang Y. Fuzzy Causal Inferences based on fuzzy semantics of fuzzy concepts in cognitive computing[J]. WSEAS Transactions on Computers, 2014, 13: 430-441.
[26] Wilson R, Keil F. The MIT encyclopedia of the cognitive sciences[J]. Electronic Resources Review, 2013, 43 (4): 282-283.
[27] Codd E F. A relational model of data for large shared data banks[J]. Communications of the ACM, 1970, 13(6): 377-387.
[28] Cardelli L, Wegner P. On understanding types, data abstraction, and polymorphism[J]. ACM Computing Surveys, 1985, 17(4): 471-523.
[29] Guttag J V. Abstract data types and the development of data structures[J]. Communications of the ACM, 1977, 20 (6): 396-404.
[30] Martin-Lof P. An intuitionistic theory of types: Predicative part[J]. Studies in Logic & the Foundations of Mathematics, 1975, 80: 73-118.
[31] McKinsey B, Gartner D. Big Data means high value, not just volume[M]. Computer Weekly, 2011, 6: 1-2.
[32] McKinsey B, Gartner D. Big Data means high value, not just volume[J]. Computer Weekly, 2011, 6: 1-2.
[33] Mashey J R. Big data and the next wave of infrastress[J]. SGI, 1998: 1-46.
[34] Wang Y. On mathematical theories and cognitive foundations of information[J]. International Journal of Cognitive Informatics and Natural Intelligence, 2015, 9(3): 41-63.
[35] Wang Y. Keynote: The emergence of abstract sciences and brain-inspired symbiotic systems[C]//IEEE FDC Workshop on Symbiotic Autonomous Systems in SMC´18.1. Piscataway N J: IEEE, 2018: 3.
[36] Chapra S C, Canale R P. Numerical methods for engineers with software and programming applications[M]. Boston: McGraw-Hill, 2002.
[37] Chicurel M. Databasing the brain[J]. Nature, 2000, 406 (6798): 822-825.
[38] Snijders C, Matzat U, Reips U D. ´Big data´: Big gaps of knowledge in the field of internet[J]. International Journal of Internet Science, 2012, 7: 1-5.
[39] Sternberg R J. In search of the human mind[M]. 2nd ed. New York: Harcourt Brace & Co, 1998.
[40] Ullman J D, Widom J. A first course in database systems [M]. New York: Prentice Hall, Inc., 1997.
[41] Wang Y. On cognitive informatics[J]. Brain and Mind, 2003, 4(3): 151-167.
[42] Wang Y. Keynote: From information revolution to intelligence revolution: big data science vs. intelligence science[C]//Proceedings of 13th IEEE International Conference on Cognitive Informatics and Cognitive Computing (ICCI*CC 2014). London: IEEE CS Press, 2014: 3-5.
[43] Hartmanis J. On computational complexity and the nature of computer science, 1994 turing award lecture[J]. Communications of the ACM, 1994, 37(10): 37-43.
[44] Wang Y. On the cognitive complexity of software and its quantification and formal measurement[J]. International Journal of Software Science and Computational Intelligence, 2019, 1(2): 31-53.

[45] Wikipedia[EB/OL]. [2019-10-31]. https://en.wikipedia.org/wiki/Largest_known_prime_number.
[46] Wang Y. On the big-R notation for describing iterative and recursive behaviors[J]. International Journal of Cognitive Informatics and Natural Intelligence, 2008, 2(1): 17-23.
[47] Wang Y. Software science: On general mathematical models and formal properties of software[J]. Journal of Advanced Mathematics and Applications, 2014, 3(2): 130-147.

The cognitive and mathematical foundations of big data science

WANG Yingxu[1,2], PENG Jun[3]

1. National Engineering Key Lab for Big Data System Software, School of Software, Beijing National Research Center of Information Science and Technology, Tsinghua University, Beijing 100084, China
2. International Institute of Cognitive Informatics and Cognitive Computing (ICIC), Deptartment of Electrical and Computer Engineering, Schulich School of Engineering and Hotchkiss Brain Institute, University of Calgary, Calgary T2N 1N4, Canada
3. School of Intelligent and Technology Engineering, Chongqing University of Science and Technology, Chongqing 401331, China

Abstract The big data play an indispensable role not only in a wide range of science fields and engineering applications, but also in the cognitive mechanisms of the sensation, the quantification, the qualification, the estimation, the measurement, the memory, and the reasoning of human beings. This paper reviews the basic studies of the theoretical foundations of the big data science, as well as a coherent set of general principles and analytic methodologies for the big data systems. The cognitive foundations of big data are explored in order to formally explain the origin and the nature of the big data. A set of mathematical models of the big data are created to rigorously elicit the general essences and patterns of the big data across pervasive domains in science, engineering, and society. A significant finding about the big data science is that the big data systems in nature are a recursively typed hyperstructure (RTHS) rather than pure numbers. The fundamental topological properties of the big data reveal a set of denotational mathematical solutions for dealing with the inherited complexities and unprecedented challenges in big data engineering.

Keywords big data science; big data engineering; mathematical models; recursively typed hyperstructures (RTHS); cognitive computing; computational intelligence

（责任编辑　刘志远）