

信息学基础研究¹

邹晓辉

519125 广东珠海井岸桥东恒美花园 15-2栋 201号 0756-5505041 qhkjy@yahoo.com.cn

【摘要】信息学基础研究在定性和定量分析的前提下集中论述（信息的）内容（意）、形式（文）、本质（义）等问题，并在信息形式化、部门信息学以及一般信息学等领域提出“语义、信息与智”的统一理论（框架）。文章主要探讨信息概念（定义和分类）并采用协同智能科学的观点和方法，对当前有代表性的几个较为典型的信息观和方法论作了点评。

【关键词】一般信息学，协同智能科学，语义三棱，信息方程，基础研究¹

一、引言

最近几年，信息形式化和部门信息学诸学科，取得了很大进展。首先，基于“奈魁斯特、哈特莱及申农的通信信息理论、维纳的控制论和图灵及冯诺依曼的数字计算机理论”而发展的信息形式化数字技术，已为通信、自动控制和计算机工程实践所证明，非常有效。接着，数字技术的广泛应用，大大促进了部门信息学相关学科的形成和发展，其景象如雨后春笋，其结果是信息学科化，其影响十分广泛。在形式信息革命大潮下，学界活跃分子不禁纷纷反思：信息为何如此神奇？日常生活通用的信息概念与各个科学领域专用的信息概念能统一吗？一个人，除了熟悉本领域专用的特殊信息概念之外，对其他领域的特殊信息概念和公共领域的一般信息概念是否也都可界定清楚呢？于是，哲学领域信息学转向呼之欲出，科学领域统一信息理论以及信息科学交叉研究领域一般信息学也应运而生（酝酿中）。

与部门信息学的发展相比，一般信息学的进展却相当迟缓，很多重要的问题至今没有得到满意的解决。如：信息哲学提出：信息、语义、智（含：智慧、智力、智能）等概念需要系统阐述，统一信息理论提出：部门信息学与一般信息学的关系需要清晰界定，信息科学交叉研究提出：一般信息的概念、原理及方法也需要系统研究。处于一般信息学前沿，既要明确常识到部门信息学和相关技术领域信息概念，还要探究哲学、科学学乃至艺术等领域的信息概念。

有鉴于此，本文提出一种“语义、信息与智”的统一理论（框架），试图探究信息的内容、形式、本质等基础和核心理论问题。同时，对当前信息科学交叉研究领域普遍存在几个认识误区，提出简明扼要而又富有启迪或值得深思的意见或建议。希望这一信息学基础研究成果对一般信息学理论探索者们有抛砖引玉之功效！欢迎同行多提宝贵意见！不当之处敬请指正！

二、正文

1.信息学基础研究立足于协同智能的科学信息观、方法论和相应的信息处理原理及方法。

我认为：无论是人类智力，还是人工智能，其实就是指信息处理能力。继人类与人工智能之后出现的协同智能在互联网及计算机辅助的自然人和软件工程支持的计算机之间互助互补的基础上获得了空前的发展。知识信息处理方法及产品广泛普及的现象，就是通过人类的自然之智与人工之智“合理分工、优势互补，高度协作、优化互动”的融智原理、方法及实例而得以体现出来的，具体表现为人机交互作用（Human Computer Interaction）。

融智信息观（集中探讨“信息的内容、形式、本质”）是基于协同智能而表达的“语义、信息（含：知识、语义信息、数据）与智”的统一理论（框架）。

融智方法论（集中探讨“间接形式化和一体化管理”）是发现并补足了在还原论和整体论之间长期缺乏中间环节——域位论之后而发展起来的协同智能方法体系（统一的理论框架）。

融智信息处理原理及方法，就是协同智能系统的逻辑推理和数学计算的原理以及知识信息数据处理方法。互联网及计算机辅助和软件工程以及人机交互作用可视为其发展的初级类型。

我还认为：信息学具有理论与实践紧密结合的特点。因此，我所进行的信息学基础研究，始终是沿着应用基础、工程基础、理论基础（简称：三基）三部曲进行的。因此我的研究成果，

¹ 本专题涉及较多的内容，因为论文集篇幅规定，本文是作为《信息学基础研究》的“绪言”来撰写的。

首先由一种知识信息信息处理方法及产品的发明（1997-2000）而获得实质性进展，进而提炼出理论。正如科学革命史一书的作者科恩所总结和预言的那样，我亲身体会“思想革命、口头革命、纸面革命”的过程。经过2000-2005六年尝试，我感受到“科学革命”不是一件容易的事情。尤其是采用知识信息数据处理方法和信息学理论融“三基”于一炉的系统工程，其困难可想而知。更不用说担当知识信息数据处理系统工程总体方案或蓝图设计者这样的艰巨任务。于是2006年我决定把工作重心收敛到“纸面革命”的过程控制上并突出人机交互过程的基础控制。因为，人机合作的观点及方法可对（静态的）语义、（动态的）信息及其处理之智进行深入仔细和富有实效的科学探讨。

“语义信息新论”一文介绍了“数据、语义信息、知识”的关系式和语义信息的定义式。“两个基本信息公式及其算法的坏与好的比较”一文介绍了哈特莱及申农的形式信息量的直接计算公式和邹晓辉的形式及内容信息量的间接计算公式的区别及联系。“广义文本与序位本义（本真信息）”一文介绍了语义三棱模型（“重构‘概念分类体系’的新思路与新方法”一文介绍了由语义三角到语义三棱的学术渊源，“字本位与中文信息处理”一文介绍了基于双列表的分层集合与基于多列表的标志集合以及新的方法论基础——域位论）。以下仅做有关概述。

2.“语义、信息与智”的统一理论（框架）

2.1 领域：“语义、信息与智”的统一理论（框架）属于信息学基础研究领域。

2.2 特殊性：信息科学原理作者钟义信把信息比作多面体。这更凸显了一般信息学的困难。我的设想是这样，即：部门信息学诸学科好比从不同角度直接观察与分析这个多面体，（因为有协同智能的概念）我的策略及方法是：重点研究（语义）三棱锥、（信息）四面体、（智的）四要点，进而可借助协同智能系统高效率地间接观察与分析（语义）多棱锥、（信息）多面体、（智的）多要点。这样，自然可做到以简驭繁进而有望开辟一般信息学研究的新途径。

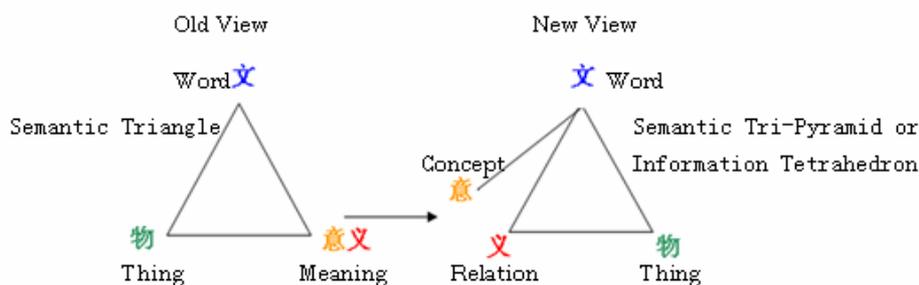
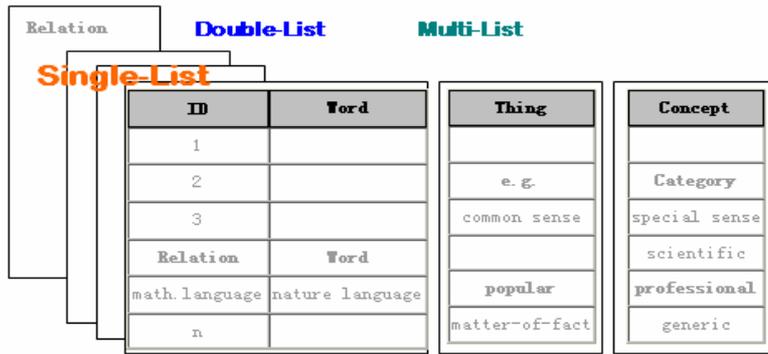


图 1是从“语义三角”到“语义三棱”的(几何模型)示意图。英文仅表示模型的实例。

2.3 重要性：众所周知，任何一个复杂的多面体都可归结为若干个简单的四面体，四面体是最简单最基本的多面体。毋庸置疑，掌握（信息）四面体可为更有效地探知（信息）多面体提供理论上的基础性指导。这比盲人摸象式地直接研究（信息）多面体的常规做法更加可取。理由之一：研究（信息）四面体可得唯一而确定的结果及结论，而研究（信息）多面体可得出多种多样不确定的结果及结论。与其直接为不可为之事，不如先为可为之事（即：打好基础、创造条件），进而，再把不可为之事化为可为之事而为之。如：先熟悉“三位一体”简单变换，然后，再借助计算机去探究“多位一体”的复杂变换，自然容易的多。一旦众人掌握“语义、信息与智”的统一理论（框架），对深入研究“信息、语义与智”的细节，对理顺部门信息学各个学科与一般信息学的关系，对研究一般信息学的框架和细节，就有了高屋建瓴的行动指南。这时，再借助互联网及计算机辅助知识信息数据处理方法及工具的支持（如：间接形式化方法及其系列产品乃至一体化管理方法及其系列服务），就可为整个信息学体系全方位全过程探讨创造更有利的条件。那样，标准化与个性化兼容的信息概念体系的总论及各论，也就可望早日建立建全。至少可加速一般信息学同仁达成共识（如明确研究对象、方法及任务）的进程。

图 2是“间接形式化”和“一体化管理”方法及其工具的(表格模型)示意图。



2.4 研究途径：（语义）三棱锥、（信息）四面体、（智的）四要点“三位一体”的几何模型（见：图 1），可形象地概括“语义、信息（知识、语义信息、数据）与智 [信息处理（分与合）机制]”的统一理论（框架）；“信息方程”的代数模型，可抽象地概括“基于双列表的分层集合以及关系数据库的软件工程”的知识信息数据处理（方法）的序位模型（见：图 2）。

方法论和基本方法，涉及（相对完全）归纳、（完全）演绎、（间接）计算。下面以中文信息处理和知识工程为例，说明“知识、语义信息、数据”的关系及其处理的基本方法。首先，通过对（自然人与计算机）通用的标准文本（如：国际统一编码 Unicode）的定性和定量分解，提炼出单一集合的各个子全域（如：字母表、笔画表、数字表、各种特殊符号表），进而，区分出分层集合的各个超子域的成员所归属的各个进阶层式（如：中文的“一，二，...，多”笔画的字——层面型结构和汉语的“一，二，...，多”音节的字与字组——线串型结构），以此作为对“语义信息和知识”实现“间接形式化”的基础（见：图 3），然后，按照“标准化与个性化兼容”的原则，对进一步提炼出基于各个学科分类的标志集合（如“语言文字、通用常识、专用知识”分科标注），实施“产、学、研、用、算”一体化管理，以此可实现“语义信息和知识”的获取、表达、以及有针对性地重用。此方法的原理和实施例，在“字本位与中文信息处理的基础”专著中将详细介绍。这里只概要介绍其中与本文有关的方法论和基本方法。

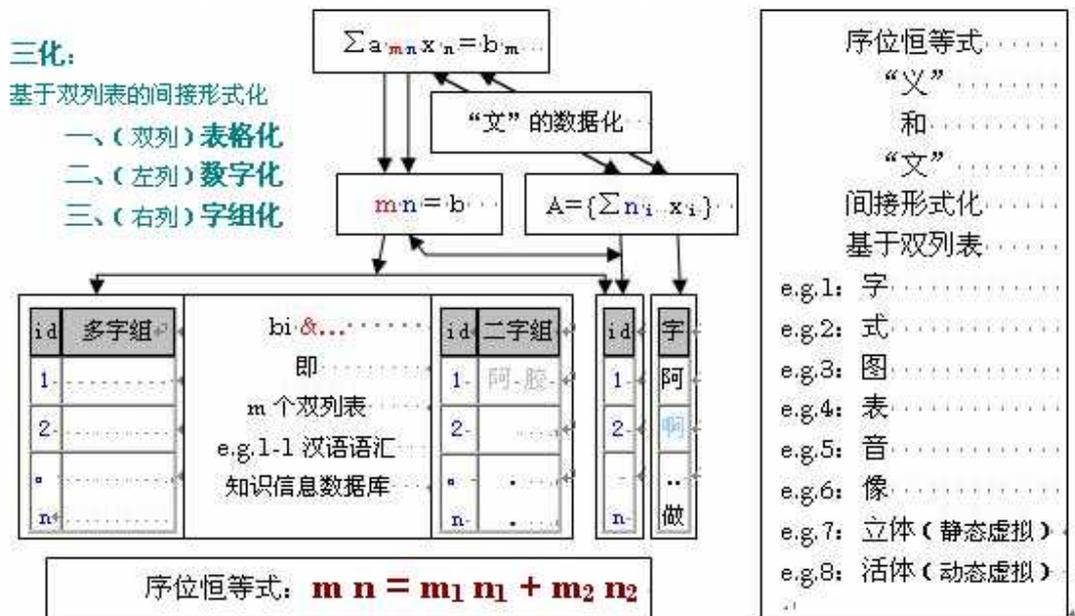


图 3 是以字与字组的关系为实例表示的“间接形式化”（三化）示意图。

“广义文本”（字、式、图、表、音、像、立体、活体）由“字”这一“狭义文本”推广而来。其中，涉及“文”与“物”两个基本范畴是显而易见的，而“意”这个范畴作为“主体的选择”必然可由“文”与“物”所蕴含或表达，这不证自明，可理解的难点在于“义”作为

控制“物、意、文”的根本范畴（本真信息）似乎只能以逻辑和数学的方式予以明确或掌控。“意义=意+义”的内涵及问题，至今仍未被人们所普遍重视。于是，由“广义文本”（物、意、文）再聚焦到“本真信息”（义）的过程，并非任何人（尤其在缺乏相应背景知识和创造能力的时候）所能快速驾驭并收敛到位的。好在协同智能的出现，带来了本真信息充分共享的希望。

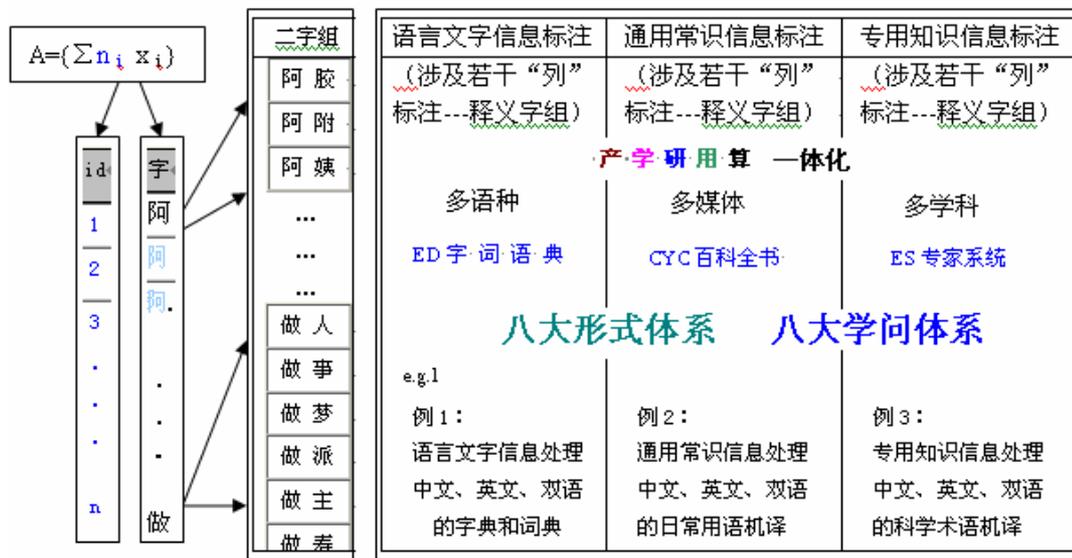
日常生活中的“集合”可这样被简化成为数学上的“集”。

方法论:	整体论	各就各位论	还原论
集合分类:	杂多集合	标志集合	分层集合
		商集	直集
			集
元素:	任意的	抽象的	简单对象
	复杂对象	标志的值	个体的数
		定性分析	结构分析
			定量分析
实例:	多语词典	分类词典	定位词典
			字母表

图 4是“方法论”与“集合分类”(相互关系)示意图。各就各位论是域位论的局部用例。

图 3和图 5的模型呈现一个基于 GLPS(全球语言定位系统)的 GKPS(全球知识定位系统)共享共建全域平台，以“产、学、研、用、算”人机协作为特征的融智过程（涉及文化基因工程）是标准化与个性化结合的开放式进化发展平台，其作用：既方便计算机辅助教学；又方便广大师生与遍布各级学校乃至校外的网络化计算机系统之间的大协作、大融通、大融合（涉及一种生产式合作型教学及科研的模式，有利于改掉各级师生乃至各类各级专家常有的孤芳自赏、固步自封或闭门造车的陋习）。

“三化”附加“三注”的义项大典与用例大全 CA 计算机辅助 多列标注与查询



字的(义项用例)直接呈现与间接标注(释义字组)

(计算机后台的分布函数与前台的标注字组满足同义并列的条件)

图 5是以字与字组的关系为实例表示的“一体化管理”(三注)示意图。

2.5基本假设(局限性):由于对信息多面体的归纳、演绎、枚举、类比乃至计算、统计、估计必受人的认识认知与实践优化的进程限制,因此,由信息多面体到信息四面体的收敛及其逆过程(发散)建立在间接形式化和相对完全归纳的逻辑推理及数学演算的基础之上。本研究对信息的讨论及处理,设订了可计算且有限的目标域,即:目标域=已知域+未知域, $D=K+1$;目标域=多表*多格, $D=m*n$ 或 $D=n*n$ (间接计量单位:ge)。

序位	形式	内容
m	$*$	n
$= D = K + I$		
义	文	意

图 5是本真信息(义)、形式信息(文)、内容信息(意)的(逻辑与数学)关系示意图。

2.6可能贡献的创新知识点：

2.6.1 信息概念体系有三个层次，一，信息的现象和本质(序位)；二，现象的形式信息(即：数据，计量单位 bit比特)和内容信息；三，内容的已知部分(即：知识)和未知部分(即：语义信息)。间接计量单位 ge格。如：当每个格仅占 1比特时，bit就是 ge的特例；当每个格仅有一个知识点时，个就是格的特例(即：形式信息与内容信息的非对称可忽略不计)。

2.6.2信息学基础研究把一般与特殊兼容的信息定义在形式上可数字化，内容上可知识化，本质上可序位化的未知域，其中，“三可”属于理论方面的基础研究领域；“三化”属于实践方面的基础研究领域；基本分类有：形式信息(文)、内容信息(意)、本真信息(义)。

2.6.3广义文本，体现：信息形式的丰富性(字、式、图、表、音、像、立体、活体)，其中，既涉及：物(质、能、时、空)的“虚拟映射”并子集(物象信息)，又涉及：意(含：静态的“知”和动态的“情、意”及其交融的“个性化”“选择”)的“虚拟映射”并子集(含：潜在的意向信息和显现的意识信息)。

2.6.4一些值得深入探讨的新课题：

1) 语义信息和知识是可通过数据而间接计量的

$I = H = N \log S$ (Hartley) 基于指对数转换的熵和形式信息计量公式， $H_s(p_1, \dots, p_n) = -K \sum p_i \log p_i$ (Shannon) 基于概率的熵和形式信息计量公式，与 $I = D - K$ 且 $D = n * n$ (ZXH) 基于双列表分层集合及关系数据库的语义信息计量公式相比较，在假设 $K = 0$ 的同等条件下，后者间接形式化的计算结果等价于前两者且有可直接使用一切已知算法的优点。

2) 信息量 = 能量与质量的比值

(表示 n 个表和 $n * n$ 个格的)序位模型 $n * n$ 在形式上与光速的平方 $c * c$ 相似，提示：当 $K = 0$ 且 $n * n = c * c$ 时，可由 $I = D - 0 = n * n$ 和 $E/m = c * c$ 推知“信息(量)、能(量)、质(量)”的关系式 $I = E/m$ 即：能(量)与质(量)的比值就是物的自然信息(量)。

3) 信息方程

狭义信息方程 $I = D - K$ (即：语义信息定义式)是(间接形式化条件下)“语义信息、数据、知识”关系式的变形。它是“获取(语义)信息、处理数据、重用知识”实施总量控制的理论基础。在科学理论上揭示(语义)信息量、数据量、知识量三大变量之间的相互关系，可得：信息学基础研究领域的三大基本原理。在技术实践上揭示知识信息数据处理方法，可得：数据(间接计算与直接呈现)及知识(间接计量与间接呈现)处理的高效方法。

(满足条件 $w = a+bi+cj+dk$ 的)广义信息方程 $f(x, y, z, ict) = w$ 似乎可揭示“空间、时间、能量、质量、信息量”五个变量的相互关系，而且似乎还可体现“本真信息，唯一守恒”法则和“物的序位与文的序位以及可选之意的序位”的等价性或恒定性。理论上，其中蕴藏的自然法则以及逻辑和数学原理值得深究。实践上，当 $w = a + bi$ 即 z, t 为 0 时特例 $f(x, y) = w$ 有意义；当 $w = a$ 即 y, z, t 为 0 时特例 $f(x) = w$ 不仅有意义而且很管用 [如： $I = D - K$ 及其相应的序位恒等式 $m_2 * n_2 = m * n - m * n_1$ (体现各论域序位模型的一致性)与基于双列表分层集合的间接形式化方法结合，可支持“选域”(推理)与“定位”解(超大规模)信息方程组]。

4) “语义、信息和智”的统一理论的几何模型

语义(三棱及多棱)、信息(四面及多面)和智(四点及多点)三位一体几何模型，采用关键少数掌控多数，如：由 4 范畴到 8 体系再到 n 系列的细分(反之则是粗分乃至合一)。

5) 有待探讨的问题

对广义文本多元数表达式 ($a + bi \dots$) 和本真信息定义式 (序位恒等式 $m_2 * n_2 = m * n - m * n_1$) 以及智的三个发展阶段——哲学的智慧、心理学的智力、认知科学的人工智能——的探讨。

7) 信息的名与实之辨

信息 (概念), 既不限于已有“信息”词条, 也不限于是否叫“信息”。与其说“信息”是名词, 不如说它是一个极为特殊的超级代词。因它几乎可指代任何“未知的或不确定的领域”。

8) 本真信息的重要性及其系统表述的困难

无论知或行, 支配“广义文本”的“本真信息”都是最重要的。尽管“文”形形色色、千变万化 (如: 文符、意念、物象), 但起支配作用的却是“本真信息” (即: 法、义、理、道, 如: 文法、意义、物理、生理、心理、哲理乃至文载之道)。对此, 古今中外感悟深刻者大有人在, 只因时代和知识信息数据处理能力的局限, 至今未见整体上与序位模型同质的科学研究。

3. 当前信息科学交叉研究中普遍存在的问题

3.1 本体问题

徐院士采用基于哈特莱及申农而创立的公式计算本体。我采用自己的公式计算本体。我个人认为: 真正能被计算的本体是“ (物的) 质能时空、 (文的) 数码、 (意的) 类例”的序位。

3.2 本质问题

徐院士在信源、信道、信宿的基础之上增加了信的、信值。我认为: 信源、信道、信宿的研究是由“物”研究“信”, 信的、信值的研究是由“意”研究“信” (这对人类具有特殊而重要的作用), 数码研究是由“文”研究“信”, 而“信”或信息的本质可序位化或“义”化。

3.3 现象问题

钟教授在研究信息科学原理几十年之后得到“信息好比是一个多面体”的结论。这无异于说一般信息学的研究还像是在“盲人摸象”。这一结论, 应该促使人们反省以往的研究路线! 我认为: 可采用由“四面体”切入的方式来建立一般信息学的统一理论 (框架), 而具体研究“多面体”各个面的任务, 主要属于部门信息学各个学科的研究任务。

3.4 名实问题

有人以是否 (如: 何时、何地、何人) 使用过“信息”这个“词 (外文)”或“辞 (中文)”, 作为信息学探源的依据。我认为: 不妥。因为, 同一个对象 (物) 或概念 (意) 或原理 (义), 完全可采用不同的词或辞 (文) 来表述。信息, 也不例外。当然, 研究初期这样做是必要的。

3.5 一般信息学研究的时机问题

也有人认为: 在部门信息学还没有完全研究透彻之前不宜开展一般信息学研究。我认为: 这对从“多面体”入手的观点来看, 很有道理。但如从“四面体”入手的观点来看, 一般信息学显然已有一条新的发展途径, 因而, 也就有了与部门信息学齐头并进协同研究的充分理由。

3.6 关于一些重要概念的语言表述问题

还有人时常谈“信息熵”, 甚至有人谈“信息能”。我认为: 这涉及一些知识领域的认识与语言表述的冲突问题。因为, 所谓“信息熵”实质是说信息学的熵 (它既不是信息也不是热力学的熵)。所谓“信息能”实质是说“信息的作用或功能”和“信息处理的能力” (其内涵区别于物理学的能和力, 其外延区别于: 哲学智慧、心理学智力、认知科学的人工智能)。

三、结语

我认为: 要夯实一般信息学的基础, 至少须有: 统一的研究对象 (1)、方法 (2) 和任务 (3), 即: (1) 由可涵盖所有特殊信息的一般信息的存在所决定。a 哈特莱和申农实际上提出了 (形式) 信息 (量) 的统一定义。b 本文提出 (形式及内容) 信息 (量) 的统一定义。(2) 由信息基本公式决定。a bit 是 (形式) 信息 (量) 的一个常量计算单位。b ge 将是 (形式及内容) 信息 (量) 的一个间接计算工具 (相当于一系列砝码, 基于分层集合的双列表就是相应的天平)。(3) 由语义信息处理 (智) 的基本方法所决定。a 基于 bit 的数字计算

技术的发展，由于具备在世界范围内达成共识乃至形成共为的理论基础和实践条件，因此取得了长足的进展。b 基于 ge 的知识信息数据处理原理、方法及实例，还在准备达成共识和形成共为的理论基础和实践条件。因此，其推广普及过程的加速或催化是必要的但要防止欲速不达。

统一的（形式及内容）信息（量）定义及其基本公式及计算单位及工具的明确，意味着：知识和语义信息的处理也可像数据处理一样得心应手。众所周知，信息科学主要是沿着形式化、数字化、数学化的可计算（形式）信息（量）bit 这一传统的直接计算途径而发展起来的，由于在语义、信息与智的基本问题上遭遇的大量歧义难以消除，所以，我在继承传统方法的基础上，另辟蹊径——提出：基于“双列表”分层集合的序位模型，其本意是为解决语言文字和知识信息的定量分析提供一套科学的量具。结果却开辟了知识和语义信息的“间接形式化”新途径。

至于提出广义信息方程，物（质、能、时、空）与信息（意、文、义）的关系，基于物的科学与基于信息的科学的关系，则是额外的发现或收获。

参考文献

1. 邹晓辉“融智学初创时期的交叉研究文选(20篇)”[C]《潜科学(前沿科学)》2005第48期[EB]
2. 邹晓辉“融智学精华介绍——融智学的知识创新点与基础实施例”[EB]《潜科学》2005第48期[EB]
3. 闫学杉“当代学科发展中的信息问题之考释”[J]《信息科学研究》2002年第4期第59-64页
4. 李宗荣“理论信息学：概念、原理与方法(博士论文)”[J]《潜科学》2005第49期[EB]
5. 苗东升“申论作为四论之一的信息科学”[J]《北京大学学报》(哲学社会科学版)2000年第6期
6. 张学文“组成论”[M]《潜科学》信息科学专栏(2005) Information Science Magazine [J] [EB]

注 1:《潜科学(前沿科学)》学术期刊网址 <http://survivor99.com/pscience/>

注 2:本文点评的几个问题是由“信息科学交叉研究学术研讨会”2005(北京)的部分论文中发现的。

注 3:本文最后修订时采纳了闫学杉先生的意见，文章的内容和形式都得到了相应的精简。特此致谢！

FUNDAMENTAL RESEARCH ON INFORMATICS

Zou Xiao Hui 0756-5505041 qhkjy@yahoo.com.cn

Room 201 Building 15-2 in Eternal Beautiful Garden In Zhu Hai 519125

Abstract: Fundamental Research on Informatics focuses on discussing content (concept), form (word), and essence (relation) based on Qualitative Analysis and Quantitative Analysis, and put forward the Unified Theory (Frame) of Semantic Information Processing in the domain of Formalization, Branch and General Informatics. Here we mainly discussed the definition of information according to my recent research and also discussed several currently typical viewpoints and methodologies on Informatics by using the viewpoint and method according to the Theory of Co-Intelligence.

Keywords: General Informatics, Co-Intelligence Science, Semantic Tri-Pyramid, Information Equation, Fundamental Research

特别说明：

本文不仅参考“潜科学网站”信息专栏和“前沿科学”学术期刊以及“熵、信息和复杂性网站”而且还对信息科学交叉研究学术研讨会(2005北京)文选的论文以及“潜科学论坛”对话中本文作者发现的几个问题给予了点评。在此，向有关作者(如：邬焜、徐光宪、姜璐等)、译者(如：刘刚)和对话方(如：鲁晨光、陈雨思、金玉成等)一并表示致谢！