

大数据时代的知识融合理论体系构建

于梦月¹, 申静²

(1. 上海大学文化遗产与信息管理学院, 上海 200444; 2. 北京大学信息管理系, 北京 100871)

摘要:构建大数据时代的知识融合理论体系,旨在为大数据时代知识融合提供基础理论和实践指导。根据大数据时代知识融合理论体系构建的目的与原则,基于拉卡托斯科学研究纲领,构建大数据时代的知识融合理论体系。大数据时代知识融合理论体系由知识融合基本理论、知识融合的辅助性假设和应用理论、大数据环境组成。其中,知识融合基本理论包括知识融合的概念和特征,知识融合的辅助性假设和应用理论包括知识融合的思维、支撑理论、需求、对象、结果、流程、技术、方法、工具等,在大数据环境的启发下,知识融合理论体系不断扩展和修正。构建的大数据时代知识融合理论体系不仅能指导大数据环境下知识融合的现实工作,还能为大数据时代的知识融合理论研究提供参考。

关键词:大数据时代;知识融合;理论体系;拉卡托斯科学研究纲领;基本理论

中图分类号:G35

文献标识码:A

大数据时代,海量的数据中蕴含着庞大的知识,这些知识规模巨大、类型众多、来源广泛、价值密度低且产生速度快。知识融合是从多个异构的网络资源中定位和抽取知识并对其进行转换的过程,并将这种知识组合结果应用于问题求解^[1]。所以,知识融合可以面向知识服务和决策问题,对这些问题知识进行有效地获取、匹配、集成和利用,进而挖掘出隐含的、有价值的知识,以提供精准的知识服务和高质量的决策方案。基于信息融合和知识工程的基础,随着本体、语义网、知识图谱、数据挖掘和自然语言处理等领域的快速发展,知识融合受到了国内外学者的广泛关注。不少学者就知识融合的技术和方法问题进行了较多研究,如知识融合的算法和技术手段改进^[2-5]、体系架构分析和设计^[1,6]等;有学者探讨知识融合与相关学科的交叉应用研究^[7-10],还有学者探讨大数据环境对知识融合的影响^[11,12];有学者关注知识融合的概念辨析和支撑理论问题^[13,14],但是知识融合的理论研究仍相对匮乏,尚未形成适用于大数据时代的知识融合理论体系。知识融合理论体系是包括知识融合的基本理论、应用理论和理论环境等在内的科学理论体系,各种知识融合理论按照一定的逻辑关系有机结合形成一个多

层次的整体。本研究试图构建大数据时代的知识融合理论体系,旨在为大数据时代知识融合提供基础理论和实践指导。

1 大数据时代知识融合理论体系构建的目的与原则

(1) 构建目的

大数据时代,不断增长的数据总量和数据类型,不断变化的数据需求,不断提高的技术水平,导致数据环境愈加复杂,丰富了知识融合的基础与理论,扩展了知识融合的应用与实践,使得知识融合的理论体系发生了改变。知识融合理论体系的建立有助于深入理解知识融合的内涵和外延,知识融合研究的对象和范围,知识融合的有关内容及其相互关系。尤其大数据时代的知识融合往往是多种来源渠道、结构与表现形式各异的、动态迭代更新并实时有效的,建立大数据时代知识融合理论体系,提炼出一个具有内在逻辑结构的知识理论体系,对于全面深入理解大数据环境下知识融合的本质以及完成复杂知识融合任务具有重要意义。因此,构建大数据时代知识融合理论体系,旨在对知识融合进行理论层面

的系统梳理,为大数据时代知识融合提供基础理论和实践指导。

(2) 构建原则

为实现大数据时代知识融合理论体系的构建目的,应遵循以下几条构建原则:①整体性原则,大数据时代的知识融合理论体系可看作一个由多个要素按照一定方式组成的整体,从要素与整体相互作用的关系中认识和把握该理论体系的特性和规律;②开放性原则,建立的知识融合理论体系需适应大数据时代开放式创新环境的变化,理论体系不是封闭的。③动态性原则,建立的知识融合理论体系需适应大数据时代多变的数据环境,面向复杂的知识融合任务,能够动态补充调整,理论体系不是一成不变的。

2 大数据时代知识融合理论体系结构设计

科学的理论必须运用科学的方法。科学理论的发展模式研究可追溯到 20 世纪初以石里克 (Moritz Schlick) 和卡尔纳普 (Paul Rudolf Carnap) 为代表的维也纳学派的逻辑实证主义;基于科学在动态过程中所表现出知识的猜测性、否定性和假说性,波普尔 (Karl Popper) 提出了科学发展模式的“证伪”原则;20 世纪 60 年代,库恩 (Thomas Samuel Kuhn) 基于社会历史主义,从科学的整体性和理论能动性上提出了科学进步的新特征,即一门学科只有形成了统一的“范式”才能被称为科学^[15];英籍匈牙利科学哲学家伊姆雷·拉卡托斯 (Imre Lakatos) 借鉴库恩的常规科学解题方法,修正了波普尔的朴素证伪主义,并提出了精致证伪主义的“科学研究纲领”^[16]。

拉卡托斯认为科学理论体系是一个有机联系的整体,而一组具有严密的内在结构的科学理论体系就是科学研究纲领,并由硬核、保护带、启发法、进化和退化标准等核心观点组成^[17]。其中,硬核是不容反驳和改变的、稳定而确定的基本理论;保护带则由可以随时调整与改变以应付反常情况的辅助性假设和应用理论构成;启发法包括在纲领中指导具体理论产生的正面启发法和阻挡知识受到直接挑战的反面启发法,正面启发法由部分得到明确表达的一组关于如何改变与发展研究纲领、如何修改与精炼保护带的提示或暗示组成,反面启发法本质上是一种禁令,禁止把反驳的矛头指向硬核,通过修正和调整保护带以保护硬核免遭经验的反驳。进化的研究纲

领必须满足具有超余的经验内容,能够说明先前的成功,且一些超余内容已得到确认,否则就是退化的研究纲领^[16]。科学研究纲领已应用于构建中国特色社会主义理论体系^[18]、信息化理论体系^[19]、数字时代目录学理论体系^[20],护理学的进步与退步^[21]等。鉴于科学研究纲领是构建理论体系的一种科学研究方法,本研究将应用该方法建立大数据时代的知识融合理论体系结构。

选择拉卡托斯科学研究纲领构建知识融合理论体系,而非其他科学哲学理论(如库恩范式理论、波普尔证伪主义),是基于以下原因:库恩范式理论强调科学发展的阶段性,认为科学在“常规科学”和“科学革命”之间交替发展。然而,知识融合理论的发展并非简单的范式更替,而是需要在大数据环境下不断调整和优化。拉卡托斯纲领通过“硬核”和“保护带”的划分,能够更好地解释这种动态调整过程。波普尔证伪主义强调科学理论的可证伪性,但知识融合理论的发展不仅需要证伪,还需要在大数据环境的启发下不断修正和扩展。拉卡托斯纲领通过“启发法”的引入,能够更好地适应这种动态变化。拉卡托斯纲领通过“硬核—保护带—启发法”的结构,能够系统地解释知识融合理论的动态演变。在大数据环境下,知识融合的理论体系需要不断调整和优化,拉卡托斯纲领的“保护带”机制能够通过调整辅助性假设来保护核心理论,使其更具适应性和韧性。“启发法”机制能够根据大数据环境的变化(如新的数据类型出现)动态调整理论体系,使其更好地适应实际需求。

基于科学研究纲领,设计大数据时代的知识融合理论体系结构,如图 1 所示。大数据时代知识融合理论体系分为硬核、保护带和启发法三个部分;知识融合基本理论是大数据时代知识融合理论体系的硬核和根本主张,构成了知识融合理论体系发展的基础,表现了知识融合理论体系的本质;知识融合的辅助性假设和应用理论是大数据时代知识融合理论体系的保护带,其作用是当知识融合理论体系与经验事实发生冲突时,通过对其自身的调整以保护知识融合基本理论不受经验事实的驳斥;大数据环境是大数据时代知识融合理论体系的启发法,即通过大数据环境的变化,不仅可以主动发现知识融合的新规律和解释知识融合的新现象,还可以动态修正知识融合的辅助性假设和应用理论,以保护知识融合基本理论免遭观察和经验的反驳。

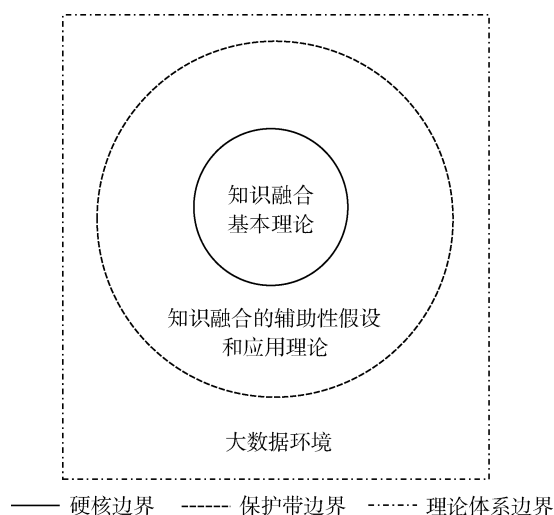


图1 大数据时代的知识融合理论体系结构

Figure 1 Structure of knowledge fusion theoretical system in the Era of Big Data

3 大数据时代知识融合理论体系结构分析

3.1 基本理论

知识融合基本理论是大数据时代知识融合理论体系不容反驳的硬核,是大数据时代知识融合理论体系构建和发展的逻辑起点,是对知识融合基本概念和特征的解读,包括知识融合概念和知识融合特征。

(1) 知识融合概念

目前学者们对知识融合概念有不同的认识,但都强调知识融合的过程,关注融合过程中的融合对象及其来源、融合流程、融合结果及融合价值。如Preece等通过分析一种用于知识融合的代理体系结构(KRAFT)提出,知识融合是指从多个异构的网络资源中定位和抽取知识,并对其进行转换,从而将组合的知识应用于解决问题的过程^[1];Smirnov等认为信息领域的知识融合是将松散耦合源的知识整合为一个组合的资源,以补充不足的知识并获得新的知识^[6];唐晓波等认为大数据时代的知识融合是面向需求和创新,通过对众多分散、异构资源上知识的获取、匹配、集成、挖掘等处理,获取隐含的或有价值的新知识,同时优化知识的结构和内涵,提供知识服务^[22];刘晓娟等通过概念辨析认为知识融合是面向知识服务和决策问题,以多源异构数据为基础,在本体库和规则库的支持下,通过知识抽取和转换获得

隐藏在数据资源中的知识因子(即知识融合过程中从数据中抽取的基本知识单元,具有语义关联性和可操作性)及其关联关系,进而在语义层次上组合推理创造出新知识的过程,并且这个过程需要根据数据源的变化和用户反馈进行实时动态调整^[13]。本研究认为大数据时代的知识融合是面向不同应用场景的知识创新和知识服务需求,在知识融合思维及支撑理论的指导下,基于知识融合的技术、方法和工具,对知识融合对象进行一系列的采集、转换和融合的处理,从而产生新的知识或应用于解决问题的过程。

(2) 知识融合特征

大数据时代的知识融合具有以下特征:①过程性。知识融合包括知识融合之前、知识融合过程和知识融合之后三个阶段,涉及知识的采集、转换、组合等一系列过程。②需求导向性。知识融合始于用户需求、特定事件和现象研究,终于产出能够满足用户个性化需求、解决特定事件和解释普遍现象的融合结果。③理论支撑性。知识融合必须在知识融合思维的指导下深刻地认识知识融合规律,且从多维度、多层次、多角度地观察与分析知识融合要解决的具体问题,并基于知识融合之前、知识融合过程和知识融合之后三个阶段的支撑理论,更好更快地完成知识融合任务。④技术实现性。面向知识融合流程,知识融合技术包含了从知识采集到最终形成融合结果并提供知识服务的全部技术,其中,实现知识融合的关键技术主要包括知识表示、知识存储和知识抽取。知识表示是一种描述知识的数据结构,通过以某种结构化的方式表示知识客体中的知识因子和知识关联,以解决多来源数据中的语义异质性问题;知识存储是将知识融合过程中的知识经过选择、过滤、加工和提炼后,按照一定的规则保存在数据库中,以随时保存和更新融合过程中的原始数据,使其成为具有知识特征的数据、语义知识和领域决策知识;知识抽取是对非结构化信息的语义标注过程,通过识别、理解、筛选、格式化,获得完整、正确的知识元及其语义关系,以发挥多来源数据的协同效果,在海量分散异构的数据中全面准确定位与抽取融合对象。

3.2 辅助性假设和应用理论

当大数据时代知识融合理论体系与观察、实验结果等有矛盾时,可以通过调整知识融合的辅助性假设和应用理论来保护知识融合基本理论,使得理

论体系具有适应性和韧性。其中,辅助性假设是一种能够独立检验并能显著增进科学知识的假说,用以解释科学理论与观察、实验结果之间的矛盾,主要包括支持知识融合开展的基础理论,如知识融合思维、知识融合支撑理论^[14];支撑知识融合实现的关键要素,如知识融合需求、知识融合对象、知识融合结果^[13];指导知识融合实施的基础知识,如知识融合流程等^[23]。应用理论涉及知识融合的具体应用,即在知识融合的实施过程中采用的手段和措施,主要包括知识融合技术、知识融合方法和知识融合工具^[24,25]。

(1) 支持知识融合开展的基础理论

① 知识融合思维

知识融合思维是应对复杂多变大数据环境的一组指导思想,由融合思维、分析思维、计算思维、创新思维、收敛思维共同组成,5种思维相辅相成^[26]。融合思维是指通过发现不同对象之间的共性,构建新的规则将不同对象及其所包含的内容组合成新的内容,从而将其应用于问题求解的思维方式。融合思维是开展大数据时代知识融合的核心方法论和基础。分析思维是指面向知识融合任务,从数据、信息和知识等对象中提取关键内容,制定可行的解决方案,以完成知识融合任务的思维方式。良好的分析思维可以提高大数据时代知识融合的效果,所以,分析思维是知识融合必不可少的重要思维。计算思维是运用计算机科学的基础概念进行问题求解、系统设计、以及人类行为理解等涵盖计算机科学之广度的一系列思维活动,对大数据时代知识融合的指导作用主要体现在,基于计算思维能够将知识融合的实际问题数学化和模型化,以期实现计算化、智能化和自动化处理。创新思维的本质在于使用新视角、新方法思考问题,提出与众不同的解决方案,从而产生新颖的、独到的思维成果,对大数据时代知识融合的指导作用主要体现在,基于创新思维能够产生差异化的创新性知识融合成果。收敛思维是指从已知信息中产生逻辑结论,从现成资料中寻求正确答案的一种有方向、有范围、有条理的思维方式,所以,基于收敛思维可以指导大数据时代的知识融合从不同来源、不同结构、不同层次,聚合各种信息,朝着一个方向获得知识融合任务的解决方案。

② 知识融合支撑理论

笔者在前期研究中,设计了大数据时代知识融合的支撑理论架构,该架构由3个层次、3个阶段、

15个支撑理论构成^[14]。其中,知识融合之前包括多元表示原理、协同理论、意义建构理论、相关性原理和情境认知理论等5个理论;知识融合过程包括本体论、形式语言理论、粒度计算理论、粗糙集理论、D-S证据理论、模糊集理论和知识空间理论(即一种用于描述知识结构和知识关系的理论框架,能够定义知识空间中的知识元素和关系)等7个理论;知识融合之后包括可视化理论、认知评价理论和信息扩散理论等3个理论。而且知识融合之前的多元表示原理、意义建构理论、相关性原理、情境认知理论和知识融合过程的本体论、形式语言理论还能为知识融合之后提供相关的理论支撑,知识融合过程的本体论、形式语言理论也能为知识融合之前提供相关的理论支撑。该架构从知识融合全过程视角,展示了知识融合之前、知识融合过程和知识融合之后3个阶段的相关支撑理论;清晰地呈现了相关理论在知识融合3个阶段中所起到的相应的支撑作用,以及在具体的知识融合应用场景下各相关支撑理论协同作用的理论集合,具有知识融合全过程的系统性,可以为知识融合的实现路径和服务模式提供理论指导。

(2) 支撑知识融合实现的关键要素

① 知识融合需求

知识融合是为了完成用户的需求或针对特定的人群而开展,或是为了解决实际生活中的特定事件而开展,或是为了解决/研究实际生活中有规律性或经常性发生的普遍现象而开展,所以用户、事件和现象是融合需求的主要来源。大数据时代知识融合的主要需求之一是知识创新需求,即通过知识的获取、组织和融合处理,挖掘出尚未被发现的、有价值的、新的知识单元,以提高知识的内涵、层次和置信度,实现现有知识的有效增值;另一个主要的需求是知识服务需求,即利用融合后经过提炼和验证的高质量的解决问题和决策的知识以提供知识服务。作为知识创新的重要工具,知识融合的需求通常具有创新性和变化性;作为提供知识服务的有效方式,知识融合的需求通常具有多样性和层次性,且不断向个性化和精准化方向发展。

② 知识融合对象

全面准确定位融合对象是知识融合的关键问题之一。大数据时代,数据总量、数据类型和新型数据源的增长扩展了知识融合对象的范畴^[11]。知识融合的对象可以来自充满大量传感器的信息物理系

统,如压力传感器、位移传感器、光电传感器、医疗器械等物理媒介中获取的由客观数据转化而成的信息;也可以来自拥有庞大信息的网络,如网络平台中丰富的网页内容知识、用户行为数据、社交媒体数据和网络舆情数据等;还可以来自多种知识资源的决策支持系统,如各种数据库、知识库、模型库、方法库、规则库,甚至包括专家经验,因此数据源、信息源和知识源都有可能成为融合对象^[27]。大数据环境下,数据体量巨大、增长速度快、类型繁多、精确性低,使得知识融合对象愈发具备多源性、异构性、动态性、分散性、碎片化、跨领域、大规模、多维度、多广度、不确定性和稀疏性^[22]。知识融合对象在知识融合的不同阶段具有不同的层次,知识融合之前属于原始状态的原始数据,知识融合过程中经过初步甄别后得到具有知识特征的数据,后又转换为统一描述的具有逻辑性的语义知识,知识融合之后通过算法使其升级为具有决策支持功能的领域知识,但是这些不同层次的数据和知识都可能成为新一轮知识融合中的融合对象。

③知识融合结果

知识融合的结果是产生有价值的、可用的新知识,或是用来补充不完全的、更加完整、更少冲突、更少不确定的知识,或是应用于问题求解的领域决策知识。知识融合的结果常常具有创新性,可以满足个性化的用户需求,或解决实际生活中的特定事件,或解释实际生活中有规律性或经常性发生的普遍现象。对知识融合结果的评价与反馈能够反映出知识融合的效果和效率,评价与反馈可以来源于用户,也可以通过一定的算法进行自适应评价。用户的评价反馈属于小样本,适用于针对特定用户的知识融合任务,因为对于个体来说,反馈是确定和准确的。自适应评价算法可以主动摒弃无效的、不合逻辑的融合结果,对其重新处理,直至产出符合逻辑的融合结果。为了确保知识融合结果的质量和有效性,可以考虑以下四项评估指标:准确性、创新性、实用性和实时性。准确性通过与真实数据或专家标注数据的对比来评估;创新性通过与现有知识的对比来衡量知识融合结果的新颖性和独特性;实用性依据用户满意度调查和实际应用效果进行评估;实时性则以数据更新频率和融合结果的时效性为依据。这些评估指标能够为知识融合的实施和优化提供科学依据。

(3)指导知识融合实施的基础知识

知识融合包括知识融合之前、知识融合过程、知识融合之后3个阶段。知识融合之前的主要任务是明确融合需求,知识融合之后的主要任务是对接用户提供知识服务和解决方案,所以,具体的知识融合流程主要集中在中间的知识融合过程阶段。不同的学者对知识融合的实现流程有不同的划分,Preece把知识融合流程划分为知识的定位、抽取和转换^[3];唐晓波和魏巍把知识融合流程划分为知识的获取、匹配、集成和挖掘^[22];赵丽梅等提出知识整合流程应包括知识整合目标设定、知识资源获取、知识资源识别、知识资源筛选、知识资源配置、知识重构等阶段^[28],所以,知识融合实施中主要包括3个核心流程,分别为:知识采集、知识转换和(狭义的)知识融合^[29]。知识采集包括对采集对象的知识定位与知识抽取,运用语义网、关联数据和数据挖掘等手段对对象知识进行定位与抽取,弱化对象知识的动态性、多源性和分散性的特点。知识转换包括对采集对象的知识过滤、知识分解和知识序化,过滤不相关的、不完整的、矛盾的知识,把过滤后的知识分解为同等粒度的知识因子,再通过对其进行统一的资源内容描述变无序为有序,获得隐藏在数据资源中的知识因子及其关联关系。规则库和知识融合算法是实现知识融合的关键支撑因素,知识融合需要基于不同的融合规则,并运用一定的融合算法,对获取与处理过的知识进行知识重组、知识分类、知识聚类 and 知识推理等运算,进而在语义层次上组合、推理、创造出新知识。知识表示贯穿于整个知识融合流程,是知识的符号化、形式化、模型化,是知识组织、关联、推理、融合的基础。

(4)知识融合的具体应用理论

①知识融合技术

知识融合技术面向知识融合流程,包含了从知识采集到最终形成融合结果并提供知识服务的全部技术,其中的关键技术主要包括知识表示、知识存储、知识抽取。知识表示是对知识的一种描述,一种计算机可以接受的用于描述知识的数据结构,能够以某种结构化的方式表示知识客体中的知识因子和知识关联。当前主要的知识表示方法包括逻辑表示法、产生式表示法、框架表示法、面向对象的表示法、语义网表示法、基于XML的表示法、本体表示法。本体表示法以一种显式、形式化的方式来表示语义,能够提高异构系统之间的互操作性,促进知识的共享与融合,被学界广泛采用。知识存储是将有价值

的知识经过选择、过滤、加工和提炼后,按照一定的规则保存在适当媒介内,以利于需求者更为便利、快速地使用,并随时更新和重组其内容和结构的活。知识融合中的知识存储的主要是通过数据库实现的。知识抽取是通过识别、理解、筛选、格式化,把文献中的各个知识点抽取出来,以一定形式存入知识库中的过程,这个过程同时又可以看作对现有的非结构化信息的语义标注过程。在知识融合中,通过知识抽取从非结构化、半结构化、结构化的多数据源中,经过命名实体识别、关系抽取和事件抽取的 3 个主要任务,获得采用某种知识表示形式的完整、正确、无歧义的知识元及其语义关系,进而作为后续知识融合的输入。

②知识融合方法

知识融合方法包括知识融合的通用算法、模式、模型和体系架构等。知识融合算法是以通用的算法设计与实现思路为基础所构建的一系列计算步骤和指令序列,是实现多源知识融合的关键,直接影响融合效果。目前,广泛采用的知识融合算法有基于贝叶斯网络的知识融合算法^[30]、基于 D-S 证据理论的知识融合算法^[31]、基于模糊集理论的知识融合算法^[32]、基于本体论的知识融合算法^[33]、基于主题图的知识融合算法^[34]等。知识融合模式是在知识融合某类应用中所使用到的一般性和普适性的规律和操作过程^[26]。例如,Smirnov 等根据信息/知识源、上下文的结构和自治的保存/更改 3 个方面的模式标准,发现了决策支持系统中基于上下文的 7 种知识融合模式,分别是简单融合、扩展融合、实例化融合、配置融合、自适应融合、平面融合和历史融合^[35]。知识融合模型随着面向不同的应用场景、基于不同的理论基础、处于不同的外界环境,可以处理不同的融合对象和内容,包括不同的功能模块和局部模块,具有开放性和较强的可扩展性。例如,Zhang 等面向解决方案的信息网络,提出了一个轻量级的基于语义的知识融合模型,该模型在自动组件方面依赖于自然语言处理,在不确定情况下依赖于众包,节省了从 Scopus 学术论文中提取信息的时间和精力^[36]。知识融合系统的体系架构是指实施知识融合时系统所采用的一般性原理和形式,包括实施知识融合的组成要素及其关系、要素之间的交互和相互作用的方式等。例如,最典型和最具影响力的知识融合 KRAFT 架构,采用开放灵活的代理体系结构,支持来自多个分布式异构源的知识融合,引

入共享本体作为公共知识交换格式,网络传递的消息必须使用共享本体内所定义的词汇^[37]。

③知识融合工具

知识融合工具通过预先打包集成的代码能够有效处理一些简单的知识融合任务。典型的知识融合工具有本体匹配工具 Falcon-AO、实体匹配工具 Dedupe/Limes/Silk 等。Falcon-AO 是一个基于 Java 的自动本体匹配系统,已经成为 RDF(S) 和 OWL 所表达的 Web 本体相匹配的一种实用和流行的选择。Falcon-AO 的匹配算法库使用 4 种匹配算法,分别是基于虚拟文档的语言学匹配 V-Doc 算法,基于编辑距离的字符串匹配 I-Sub 算法,基于本体 RDF 图结构的匹配 GMO 算法,基于分而治之的大本体匹配 PBM 算法。Dedupe 是用于模糊匹配,记录去重和实体链接的 Python 库。Limes 是一个基于度量空间的实体匹配发现框架,适合于大规模数据链接,框架形成了从知识库、样本选取、过滤、相似度计算、序列化再到知识库的环路。Silk 是整合异构数据源的开源框架,整体框架包括从知识库、预匹配、链接、过滤到输出 5 个步骤。

3.3 大数据环境

对大数据时代知识融合理论体系产生最大影响的是理论体系所处的时代环境,即大数据环境。大数据环境下,数据规模的海量性、数据类型的多样性、数据来源的广泛性、数据价值的稀疏性、数据产生的时效性等多种特点,丰富了知识融合的内容,将不断催发知识融合理论体系的适应性改变。大数据时代的知识融合可能在更多思维方式和支撑理论的作用下,面向更多需求特征,扩展更多来源渠道,融合更多类型的对象,产生更多样的成果形式,涉猎更广泛的学科领域或应用场景,以满足大数据环境的改变及其对知识融合的要求。面向大数据环境的复杂知识融合任务,传统的知识融合算法可能较难满足异构大数据环境中知识融合的准确性及实时性需求,要提出更高效的知识融合方法;为了控制知识融合结果的规模,提高知识的语义相关性和准确度,要建立更高效的知识融合框架;为了提高知识融合的可靠性与实时性,要不断改进知识融合的关键技术手段;为了实现更优的解决方案,要改进功能模块和局部模块结构,优化知识融合模型;为了提供灵活的自主服务,要开发更多实用性和简易性的知识融合工具。因此,受大数据环境影响,大数据时代知识融合理论体系可以动态补充、调整、修正、扩展。

4 大数据时代知识融合理论体系建立

根据大数据时代知识融合理论体系构建的目的

与原则,基于拉卡托斯科学研究纲领,结合上述对大数据时代知识融合理论体系结构的深入分析,建立了大数据时代的知识融合理论体系框架,如图 2 所示。

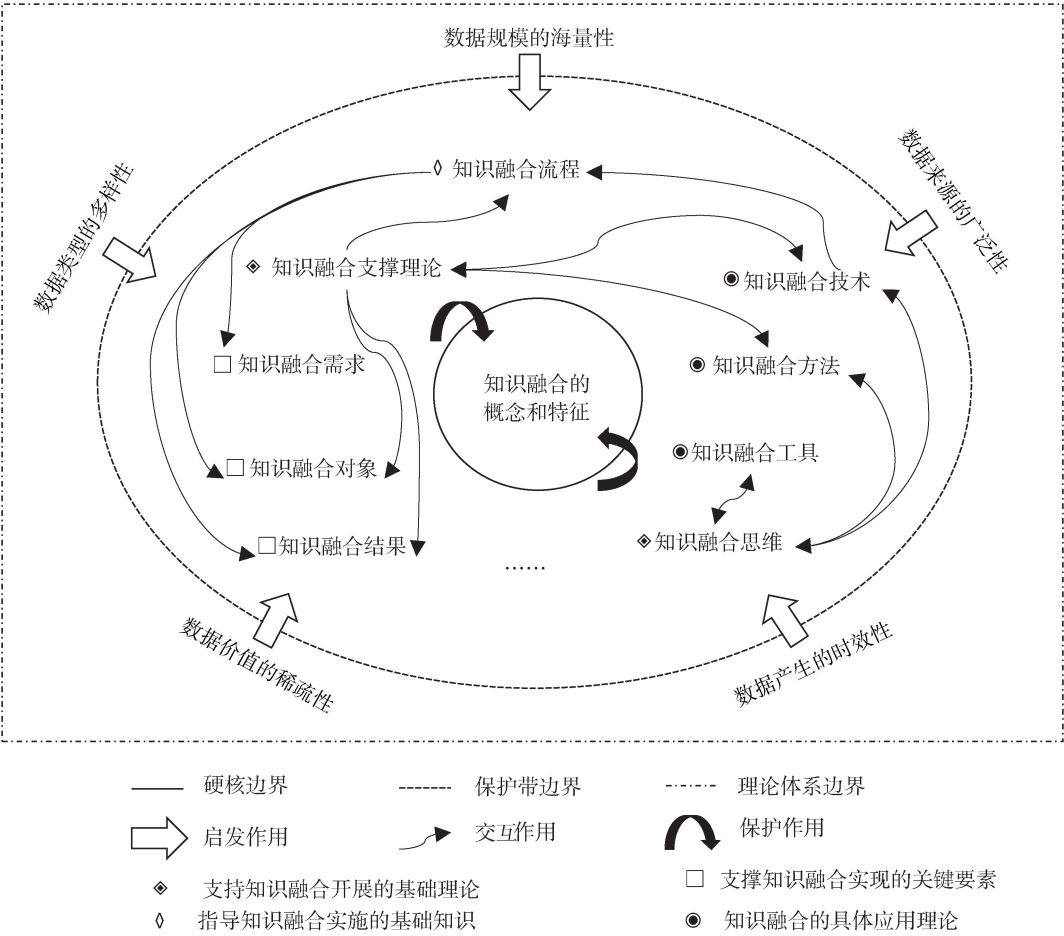


图 2 大数据时代的知识融合理论体系框架

Figure 2 Framework of knowledge fusion theoretical system in the Era of Big Data

由图 2 可见,大数据时代的知识融合理论体系是由知识融合基本理论、知识融合的辅助性假设和应用理论、大数据环境三个部分组成。其中,知识融合基本理论是大数据时代知识融合理论体系的硬核,是不容反驳的,包括知识融合的概念和特征(见图 2 的内核圆);知识融合的辅助性假设和应用理论是大数据时代知识融合理论体系的保护带,是可以改变和调整的(见图 2 的椭圆带),主要包括支持知识融合开展的基础理论(见图 2 的◇),支撑知识融合实现的关键要素(见图 2 的□),指导知识融合实施的基础知识(见图 2 的◇),知识融合的具体应用理论(见图 2 的●);大数据环境是大数据时代知

识融合理论体系的启发法(见图 2 的虚线方框),因为大数据环境具有数据规模的海量性、数据类型的多样性、数据来源的广泛性、数据价值的稀疏性、数据产生的时效性等特点,这将涌现更多的知识融合思维和支撑理论,面向更多的知识融合需求和更多类型的知识融合对象,产生更多形式的知识融合结果,开发更高效率的知识融合流程和更先进的知识融合方法技术,发明更多实用性和简易性的知识融合工具,从而推动知识融合理论体系的适应性演变。可见,大数据时代的知识融合理论体系是一个由内核圆的知识融合基本理论,到中间保护带的知识融合的辅助性假设和应用理论,以及受外界启发的大

数据环境形成的多层次有机整体,具有整体性和层次性的特点。

知识融合的思维、支撑理论、需求、对象、结果、流程、技术、方法、工具之间并不是孤立存在的,而是存在着复杂的交互关系。具体而言,知识融合思维能够指导技术、方法和工具的选择与实施。例如,在面对不同的知识融合任务时,融合思维可以指导多源异构数据的整合,创新思维则能推动新的融合算法的开发。在实际应用中,融合思维有助于选择合适的知识融合技术,如知识图谱技术,从而实现多源数据的语义关联和深度整合。知识融合支撑理论为技术、方法提供了坚实的理论基础。知识融合支撑理论贯穿知识融合流程,包括知识融合之前、知识融合过程、知识融合之后 3 个阶段的相关支撑理论。其中,相关性原理、意义建构理论和情境认知理论可以指导选择知识融合对象;本体论和形式语言理论可以提供统一的知识表示形式;D-S 证据理论、模糊集理论和知识空间理论可以设计不同的知识融合算法;认知评价理论和信息扩散理论可以应用于对知识融合结果的评价。知识融合技术、方法的实施结果能够反馈到思维和理论的优化中。例如,通过实践发现现有算法的不足,可以推动新的思维模式的形成和理论的修正。在实际应用中,如果发现某知识融合算法在处理特定类型的数据时效果不佳,可以促使研究人员重新审视融合思维和支撑理论,进而提出新的算法或改进现有算法。知识融合流程与各要素密切相关且相互交互。知识融合流程的 3 个阶段分别与知识融合需求、知识融合对象和知识融合结果紧密相连,因为知识融合之前的主要任务是明确知识融合需求,知识融合实施中旨在选择合适的知识融合对象进而实现知识的采集、转换和融合,知识融合之后的主要任务是提供满足融合需求的知识融合结果。知识融合技术也面向知识融合流程,包含了从知识采集到形成融合结果并提供知识服务的全部技术。由此可见,大数据时代的知识融合理论体系并非是知识融合理论的简单罗列,而是各个理论之间存在着一定的交互关系,具有交互性的特点。

受大数据环境启发,知识融合的思维、支撑理论、需求、对象、结果、流程、技术、方法、工具与大数据环境的具体特征之间也存在着复杂的交互关系。比如,数据规模的海量性、数据类型的多样性、数据来源的广泛性促使数据总量、数据类型和新型数据

源不断增长,进而扩展了知识融合对象的范畴,使得知识融合对象愈发具备大规模、异构性、多源性等特征。数据产生的时效性体现在数据的更新频率、延迟时间、有效期限和实时性等,衡量了数据在特定时间范围内的准确性和可用性,知识融合需求和知识融合结果通常具有创新性和及时性,需要利用最新的数据。由于数据规模的海量性、数据类型的多样性、数据来源的广泛性、数据价值的稀疏性、数据产生的时效性,促使数据体量巨大、类型丰富、多来源、跨领域、精确性低、增长速度快,这对于知识融合的技术、方法和工具提出了更高的挑战。如在类型丰富的数据中用更有效的知识表示方法提高异构系统之间的互操作性,在体量巨大的数据中用更合理的知识存储方式重组保存有价值的知识,在多来源的数据中用更高效的知识抽取模式完善语义标注过程,设计跨领域多源异构数据下更可行有效的知识融合算法,建立更多模块功能的可扩展性更强的知识融合模型,以及开发适用于更大数据量级的知识融合工具等。显然,在大数据环境的启发下,经过改变和调整的知识融合辅助性假设和应用理论,在理论上将对事实作出更多的解释,在经验上将发现新的事实,从而构成进化的知识融合理论体系。可见,受大数据环境启发,知识融合理论体系可以通过对原有辅助性假设和应用理论的重新解释或者提出新的辅助性假设和应用理论,进而动态地调整、改变和修正知识融合理论体系,具有开放性和动态性的特点。

显然,基于拉卡托斯科学研究纲领构建的知识融合理论体系,是一个由内到外的多层次有机整体,是一种科学发展观,且在大数据环境的启发下,该体系可以动态地调整、改变和修正。可见,笔者构建的知识融合理论体系具有整体性、层次性、交互性、开放性和动态性的特点,表明这一理论体系是一个系统化的知识理论体系。但是,由于事物动态发展的内在规律,科学研究纲领的进化和退化是相对的、动态变化的。因此,本文构建的知识融合理论体系也可能处于进化、退化、新旧交替、新进化的不同阶段。期望本文构建的知识融合理论体系能表现出明显的进化特征,即在理论上,通过对知识融合辅助性假设和应用理论的调整能更好地描述和阐释既定的知识融合规律,且在经验上能更好地承受知识融合现象的检验。当然,随着科学技术的不断进步,对于知识融合领域来说,会有若干个知识融合理论学派在其

进化阶段处于竞争对手的地位,存在着各种观点长期争议。可以说百花齐放式的知识融合理论之竞争是知识融合理论体系发展的根本动力,竞争的激烈程度决定了知识融合理论体系发展的速度。而在竞争中,各个知识融合理论学派之间可以相互借鉴、相互吸收、相互发扬,促进共同进步。

5 结 语

大数据时代,大数据和人工智能等理论和技术促使知识融合的技术和方法在政府、电子商务、医疗、教育等领域广泛应用,但知识融合的理论研究仍相对匮乏,尚未形成大数据时代特有的知识融合理论体系。本研究根据大数据时代知识融合理论体系构建的目的与原则,基于拉卡托斯科学研究纲领,建立了大数据时代的知识融合理论体系。该理论体系是由知识融合基本理论、知识融合的辅助性假设和应用理论、大数据环境组成,具有整体性、层次性、交互性、开放性和动态性的特点,是一个系统化的知识理论体系。本研究构建的大数据时代知识融合理论体系,更强调大数据环境对知识融合理论的启发作用以及大数据环境下知识融合的特征,该理论体系可以解决当前知识融合理论相对分散,难以全面解释与指导大数据环境下知识融合任务的现实问题,并为大数据时代的知识融合理论研究提供参考。本研究构建的大数据时代知识融合理论体系不仅为知识融合提供了理论指导,还具备实际应用的可行性。在实施过程中,该体系可应用于多个场景。在政府决策支持领域,通过知识融合技术整合多源异构的政务数据,为政策制定提供精准的知识服务。在企业知识管理领域,企业可借助知识融合方法,整合内部的业务数据、客户数据和市场数据,优化知识结构,提升知识利用效率。在医疗健康领域,知识融合可用于医疗大数据环境中的疾病诊断、治疗方案优化及医疗资源管理,如通过融合患者电子病历、基因数据和临床指南,提供个性化诊断建议。

与传统知识融合理论相比,本文理论体系在大数据环境下展现出显著的动态性。传统知识融合理论侧重于静态知识整合,假设数据和用户需求相对稳定,在应对快速变化的数据环境和动态更新的需求时,往往显得不够灵活。而本文体系强调实时响应数据变化和需求更新,更具灵活性和适应性。此外,基于拉卡托斯“硬核—保护带—启发法”框架,

本文体系能系统解释知识融合理论的动态演变,弥补传统理论在系统性描述上的不足,全面阐释知识融合的复杂动态变化。尽管如此,本文构建的大数据时代知识融合理论体系仍有待完善。后续研究可借助实证研究或专家访谈等方式,对理论体系的有效性予以进一步验证,进而提升结论的可信度。同时,由于知识融合理论体系构建是一个不断验证和演变的过程,后续可通过吸收和整合不同领域、不同场景的新兴知识融合的辅助性假设和应用理论,并在大数据环境的启发下,改变和调整该理论体系的保护带,从而不断完善大数据时代的知识融合理论体系。

参考文献:

- [1] Preece A, Hui K, Gray A, et al. KRAFT: An agent architecture for knowledge fusion[J]. International Journal of Cooperative Information Systems, 2001, 10(1-2): 171-195.
- [2] Tai C H, Chang C T, Chang Y S. Hybrid knowledge fusion and inference on cloud environment[J]. Future Generation Computer Systems, 2018, 87: 568-579.
- [3] Liu Y, Zhou J, Gan J. Heterogeneous knowledge fusion algorithm for minority cultural resources based on MapReduce[J]. International Journal of Perforability Engineering, 2019, 15(8): 2258-2266.
- [4] Seyedmohammadi S J, Atapour S K, Abouei J, et al. KnFu: Effective Knowledge Fusion[C]. 27th International Conference on Information Fusion (FUSION). Venice, ITALY: IEEE, 2024.
- [5] 王忠义,王泽人,李志鹏,等. 面向馆藏数字资源的跨学科事件知识融合研究[J]. 情报学报, 2025, 44(5): 577-591. Wang Z Y, Wang Z R, Li Z P, et al. Interdisciplinary event knowledge fusion research for library digital collections[J]. Journal of the China Society for Scientific and Technical Information, 2025, 44(5): 577-591.
- [6] Smirnov A, Pashkin M, Chilov N, et al. Multi-agent Architecture for Knowledge Fusion from Distributed Sources[M]. Berlin: Springer, 2002: 293-302.
- [7] Kawtrakul A, Khunthong V, Suktarachan M, et al. Development of an Information Integration and Knowledge Fusion Platform for Spatial and Time Based Advisory Services: Precision farming as a case study[C]. Annual SRII Global Conference. San Jose: IEEE, 2014: 241-248.
- [8] Liu J, Xu W, Zhan H. Multi-source and Heterogene-

- ous Knowledge Organization and Representation for Knowledge Fusion in Cloud Manufacturing[C]. Proceedings of International Conference on Soft Computing Techniques and Engineering Application. New Delhi:Springer, 2014:55-61.
- [9] Liu J, Li T, Ji S, et al. Urban flow pattern mining based on multi-source heterogeneous data fusion and knowledge graph embedding[J]. IEEE Transactions on Knowledge and Data Engineering, 2021,35(2):2133-2146.
- [10] 张彬,徐建民,吴姣. 跨域推荐中的知识融合研究进展[J]. 现代情报,2023,43(3):157-166. Zhang B, Xu J M, Wu J. Research progress of knowledge fusion in cross domain recommendation[J]. Journal of Modern Information, 2023,43(3):157-166.
- [11] 赵蓉英,张心源. 大数据环境对知识融合的影响研究[J]. 情报学报,2017,36(9):878-885. Zhao R Y, Zhang X Y. Review of impact study of big data on knowledge fusion[J]. Journal of the China Society for Scientific and Technical Information, 2017,36(9):878-885.
- [12] Xie N, Wang W, Ma B, et al. Research on an agricultural knowledge fusion method for big data[J]. Data Science Journal, 2015,14:7.
- [13] 刘晓娟,李广建,化柏林. 知识融合:概念辨析与界说[J]. 图书情报工作,2016,60(13):13-19,32. Liu X J, Li G J, Hua B L. Knowledge fusion: From the conceptual understanding to the system construction[J]. Library and Information Service, 2016,60(13):13-19,32.
- [14] 于梦月,申静. 大数据时代知识融合的支撑理论架构[J]. 信息资源管理学报,2020,10(3):18-26. Yu M Y, Shen J. Supporting theory framework for knowledge fusion in the era of big data[J]. Journal of Information Resources Management, 2020,10(3):18-26.
- [15] 杨颖春. 波普尔与拉卡托斯的科学发展模式比较[J]. 山西大同大学学报(自然科学版),2014,30(1):93-96. Yang Y C. Comparison and contrast of scientific development model—Based on the philosophy of popper and lakatos[J]. Journal of Shanxi Datong University(Natural Science Edition), 2014,30(1):93-96.
- [16] 杨波,王学东. 论拉卡托斯的科学研究纲领[J]. 西南科技大学学报(哲学社会科学版),2009,26(5):81-84. Yang B, Wang X D. On Lakatos' methodology of scientific research program[J]. Journal of Southwest University of Science and Technology(Philosophy and Social Science Edition), 2009,26(5):81-84.
- [17] Lakatos I. The Methodology of Scientific Research Programmes[M]. Cambridge: Cambridge University Press, 1978.
- [18] 欧健. 中国特色社会主义理论体系结构新探——基于拉卡托斯理论的分析[J]. 喀什师范学院学报,2012,33(2):1-4. Ou J. New exploration on the structure of chinese characteristic socialism theory system—Based on the analysis of the Lakatos's theory[J]. Journal of Kashi University, 2012,33(2):1-4.
- [19] 郑建明,经渊. 新型城镇化进程中的信息化理论体系研究[J]. 图书与情报,2016(1):1-6. Zheng J M, Jing Y. Study on the informatization theory system in the new urbanization process[J]. Library & Information, 2016(1):1-6.
- [20] 王锰,陈雅,郑建明. 数字时代目录学理论体系构建的相关理论研究[J]. 图书馆学研究,2014(17):2-6. Wang M, Chen Y, Zheng J M. A research on the basis to construct the bibliography theoretical system in digital age[J]. Research on Library Science, 2014(17):2-6.
- [21] Younas A. Examining progression and degeneration of nursing science using Imre Lakatos's methodology of scientific research programs[J]. Nursing Philosophy, 2020,22(2):e12342.
- [22] 唐晓波,魏巍. 知识融合:大数据时代知识服务的增长点[J]. 图书馆学研究,2015(5):9-14,8. Tang X B, Wei W. Knowledge fusion: The Age of Big Data knowledge service increasing point[J]. Research on Library Science, 2015(5):9-14,8.
- [23] 郭强,关欣,曹昕莹,等. 知识融合理论研究发展与展望[J]. 中国电子科学研究院学报,2012,7(3):252-257. Guo Q, Guan X, Cao X Y, et al. The development and overview of knowledge fusion theory research[J]. Journal of China Academy of Electronics and Information Technology, 2012,7(3):252-257.
- [24] 陈沫,李广建. 大数据环境下知识融合技术体系研究[J]. 图书情报工作,2022,66(20):20-31. Chen M, Li G J. Research on the knowledge fusion technology taxonomy in big data environment[J]. Library and Information Service, 2022,66(20):20-31.
- [25] 林海伦,王元卓,贾岩涛,等. 面向网络大数据的知识融合方法综述[J]. 计算机学报,2017,40(1):1-27. Lin H L, Wang Y Z, Jia Y T, et al. Network big data oriented knowledge fusion methods: A survey[J]. Chinese Journal of Computers, 2017,40(1):1-27.
- [26] 李广建,罗立群. 走向知识融合——大数据环境下情

- 报学的发展趋势[J]. 中国图书馆学报, 2020, 46(6): 26–40. Li G J, Luo L Q. Towards knowledge fusion: The development trend of information science in big data environment[J]. Journal of Library Science in China, 2020, 46(6): 26–40.
- [27] Smirnov A, Levashova T. Knowledge fusion patterns: A survey[J]. Information Fusion, 2019, 52: 31–40.
- [28] 赵丽梅, 孙艳华, 刘岩芳. 基于行为知识信息融合模型的高校科研团队内部知识整合流程研究[J]. 科技进步与对策, 2014, 31(21): 125–132. Zhao L M, Sun Y H, Liu Y F. A study on the internal knowledge integration process of university research teams based on behavioral knowledge information fusion model[J]. Science & Technology Progress and Policy, 2014, 31(21): 125–132.
- [29] 张卫东, 左娜, 陆璐. 政府网站信息资源知识融合体系架构设计[J]. 图书情报工作, 2018, 62(17): 112–119. Zhang W D, Zuo N, Lu L. Knowledge fusion system architecture design of government website information resources [J]. Library and Information Service, 2018, 62(17): 112–119.
- [30] Sameh A. Knowledge based bayesian network construction algorithm for medical data fusion to enhance services and diagnosis[J]. Journal of Computer Science, 2019, 15(5): 612–634.
- [31] 韩立岩, 周芳. 基于 D–S 证据理论的知识融合及其应用[J]. 北京航空航天大学学报, 2006(1): 65–68, 73. Han L F, Zhou F. Knowledge fusion based on D–S evidence theory and its application [J]. Journal of Beijing University of Aeronautics and Astronautics, 2006(1): 65–68, 73.
- [32] Yuan Y, Chai H. Knowledge fusion method based on fuzzy set theory[J]. Journal of Intelligent & Fuzzy Systems, 2020, 38(4): 3971–3979.
- [33] Smirnov A, Levashova T, Shilov N, et al. Knowledge Fusion in Context – aware Decision Support: Ontology – Based Modeling and Patterns[M]. Berlin: Springer International Publishing, 2014: 35–51.
- [34] 鲁慧民, 冯博琴, 李旭. 面向多源知识融合的扩展主题图相似性算法[J]. 西安交通大学学报, 2010, 44(2): 20–24. Lu H M, Feng B Q, Li X. Novel similarity algorithm of extended topic maps for multi – resource knowledge fusion[J]. Journal of Xi'an Jiaotong University, 2010, 44(2): 20–24.
- [35] Smirnov A, Levashova T, Shilov N. Patterns for context – based knowledge fusion in decision support systems [J]. Information Fusion, 2015, 21(1): 114–129.
- [36] Zhang Y, Saberi M, Chang E. A semantic – based knowledge fusion model for solution – oriented information network development: A case study in intrusion detection field[J]. Scientometrics, 2018, 117(2): 857–886.
- [37] Preece A, Hui K, Gray A, et al. The KRAFT architecture for knowledge fusion and transformation [J]. Knowledge – Based Systems, 2000, 13(2): 113–120.

Construction of knowledge fusion theoretical system in the Era of Big Data

YU Meng – yue¹, SHEN Jing²

(1. School of Cultural Heritage and Information Management, Shanghai University, Shanghai 200444, China;

2. Department of Information Management, Peking University, Beijing 100871, China)

Abstract: To construct the theoretical system of knowledge fusion in the big data era aims to provide basic theory and practical guidance for knowledge fusion in the big data era. According to the purpose and principle of the theoretical system of knowledge fusion in the era of big data, based on Lakatosian research programme, the theoretical system of knowledge fusion in the era of big data is constructed. The theoretical system of knowledge fusion in the era of big data is composed of basic theory of knowledge fusion, auxiliary hypothesis and application theory of knowledge fusion, and big data environment. Among them, the basic theory of knowledge fusion includes the concept and characteristics of knowledge fusion. The auxiliary hypothesis and application theory of knowledge fusion includes the thinking, supporting theory, demand, object, result, process, technology, method and tool of knowledge fusion. Inspired by the big data environment, the theoretical system of knowledge fusion can be expanded and modified continuously. The theoretical system of knowledge fusion in the era of big data solves the practical problems that the relevant theories of knowledge fusion are scattered and difficult to fully explain and guide the task of knowledge fusion in the environment of big data, and provides a reference for the theoretical research of knowledge fusion in the era of big data.

Key words: Big Data Era; knowledge fusion; theoretical system; Lakatosian Research Programme; basic theory