碎片化知识处理与网络化人工智能

汪建基[†], 马永强[†], 陈仕涛, 刘子熠, 郑南宁*

西安交通大学人工智能与机器人研究所, 西安 710049

* 通信作者. E-mail: nnzheng@mail.xjtu.edu.cn

[†] 同等贡献

收稿日期: 2016-05-06; 接受日期: 2016-09-24; 网络出版日期: 2017-02-13

国家重点研发计划 (批准号: 2016YFB1000903) 和国家高技术研究发展计划 (863) (批准号: 2014AA01A301) 资助项目

摘要 互联网的出现带来了信息量的爆发式增长, 加快了信息的传播速度, 其传播方式也更加多样化. 人类社会进入了一个前所未有的碎片化知识时代. 互联网中的碎片化知识是人类大量行为、思考与交互活动的产物. 传统的人工智能试图通过知识工程所建立的专家系统实现机器的智能, 但无法有效地解决碎片化知识的处理与高效利用. 因此, 如何对这些巨量的碎片化知识进行归纳重组与新知识的再发现, 是当前信息科学和人工智能研究领域面临的重要科学问题. 本文分析了传统人工智能方法面对大规模碎片化知识处理时存在的问题, 讨论了碎片化知识处理组织与学习的基本结构, 以及网络化人工智能的概念, 介绍了大数据演化规律的公式发现方法, 并对未来的研究热点进行了展望.

关键词 碎片化知识, 网络化人工智能, 知识发现, 机器推理, 群智网络, 公式发现

1 引言

知识处理是人工智能研究的重要内容, 在互联网发展的驱动下, 知识碎片化时代的到来对知识处理的能力和能效提出了更高的要求. 而且知识表达的方式早已从传统的静态文本数据格式发展到了声音、动画、图像和视频等多种模式^[1,2], 知识的数量更是呈指数增长形式加速倍增. 知识处理^[3]包含了知识的获取、知识的加工、逻辑判断、推理与知识的输出等全部过程. 在这样的背景下, 大数据推动下的新一代知识工程就应运而生, 进而催生了新一代人工智能方法的出现——网络化人工智能.

传统的专家系统根据某一领域一个或多个专家提供的知识和经验, 进行推理和判断, 模拟人类专家的决策过程, 以便解决那些需要人类专家处理的复杂问题. 专家系统是一种模拟人类专家解决领域问题的计算机程序系统^[4~6]. 它主要由知识库和推理机两部分组成, 完成一个专家系统的过程被称为“知识工程”, 它包括了知识的获取、表示以及软件的设计与编程实现. 传统意义上的知识获取主要由

引用格式: 汪建基, 马永强, 陈仕涛, 等. 碎片化知识处理与网络化人工智能. 中国科学: 信息科学, 2017, 47: 171–192, doi: 10.1360/N112016-00120

Wang J J, Ma Y Q, Chen S T, et al. Fragmentation knowledge processing and networked artificial intelligence (in Chinese). Sci Sin Inform, 2017, 47: 171–192, doi: 10.1360/N112016-00120

相关领域的专家提供。然而,在互联网的快速发展下,领域知识已不仅仅被局限在少数的专家身上。大量的碎片化知识中蕴含着丰富的信息,其中不乏有价值的领域知识,但其本身碎片化的结构给知识获取造成了极大的难度;知识表示是建立在知识获取的基础上选择合适的方式准确地将知识展现给用户,人类科技的不断发展进步推进了各个学科领域间的相互融合,出现了多种交叉学科。因此,知识的表达也不能只通过传统的单一形式展现,而更需要结合相关知识信息^[7],为用户提供更全面的内容,这对专家系统提出了更高的要求。为了很好地应对新环境下人类对专家系统的需求,在知识工程的设计过程中可以选择网络化结构^[8]。网络化结构在知识获取过程中更加适合应对碎片化知识本身所具有的高速增长性、传播的社会性、稀疏无序、内容冗余与隐喻的基本属性。碎片化知识间存在着细微的内在联系,每一个碎片本身的价值十分有限,但将其联系起来则会得到更有意义的信息,通过网络化结构将碎片信息联系起来更利于我们获取知识。同时,由于知识间的相互依赖,使用网络化结构进行呈现也是一种更加直观有效的表达方式。

传统知识工程所建立的专家系统无法很好地解决在碎片化知识背景下为人类提供最准确有效的决策建议和高效的知识利用,如何通过网络化结构设计建立合适的人工智能系统是我们所面临的一个充满挑战的命题。本文将从传统知识工程与人工智能所面临的瓶颈问题出发,分析碎片化知识与网络化人工智能的结构特点,探讨网络化人工智能的可能实现途径与方法,并展望未来互联网+知识自动化以及网络化人工智能在不同领域的应用前景。

2 人工智能与知识工程

人工智能的核心是知识处理问题^[9~11]。伴随着人工智能的概念在1956年被首次提出,知识处理也应运而生,在一系列实际需求的不间断推动下,相应的知识工程也不断出现(如图1所示)。

Feigenbaum在1977年第五届国际人工智能联合会议上首次提出了“知识工程”^[12,13]这一概念。知识工程是人工智能的原理和方法,对需要人类专家才能够解决的领域应用难题提供求解途径。知识工程已在诸多领域得到应用,处理对象已从规范化知识扩展到非规范化知识,处理方式和规模也从封闭式扩大为开放式,从中小规模发展为能进行海量知识处理的大规模工程。

知识工程以知识为基础,通过计算机程序实现并建立专家系统。第一代知识工程出现在1968年到1980年间,以中小规模的知识工程为主,DENDRAL系统^[14]就是其中典型的代表。随着各类信息的不断增长与实际应用的需求催生了第二代大规模知识工程的诞生,其中Lenat^[15]于1984年开发的Cyc大型知识库尤为引人注目。第三代超大规模知识工程出现于2000年,至今仍在不断发展,加之大数据时代的到来,一些互联网企业和研究机构都在建立各自的知识图谱系统^[16,17](如图1所示)。

2.1 传统知识工程的局限性

目前,知识工程主要遵循着“事先(ex-ant)”模式^[13],即问题先于知识库构建。这种模式的特点是“以问题为驱动,以专家知识为核心”。在这种模式中,用户首先提出问题或需求,然后由知识工程师(knowledge engineer)与领域专家沟通,共同建立知识库与推理求解机制。这种模式存在两方面的瓶颈问题。

(1) 由于知识工程师与领域专家沟通困难、知识表示复杂,难以构造大规模、高质量知识库且成本高,产生了知识获取(knowledge acquisition)^[18]瓶颈问题。以Cyc知识库为例,经过900个人年的工作量,OpenCyc 1.0中仅包含了47000个概念,306000个事实。此外,人类知识增长与老化速度都不断

图 1 (网络版彩图) 知识工程的发展历程与未来

Figure 1 (Color online) The history and future of knowledge engineering

加快 (如纳米、生物等领域的知识倍增周期仅有 2~3 年), 知识高度的动态性加剧了传统知识工程中的知识获取与更新的瓶颈问题。

(2) 对已有知识库的评价、维护和重用等难度较高, 造成知识再工程 (knowledge reengineering) [19] 瓶颈问题。在 20 世纪 90 年代, 人们将本体 (ontology) [20] 引入知识工程, 开展了本体库构建的研究与实践, 代表性的本体库如 WordNet 和 HowNet 等。然而, 本体库在解决再工程瓶颈问题上仍具有局限性, 例如, 缺少求解问题所需的断言性知识 (assertional knowledge)。

上述两个基本问题导致传统知识工程难以满足大数据互联网时代、以知识为核心要素的服务新业态的需求。在新的时代发展背景下, 人们需要将人工智能与互联网进行充分的融合, 寻找更加合适高效的解决方案。

2.2 专家系统与用户产生知识的融合

云计算 [21]、互联网和物联网 [22] 正在对经济社会产生着深远的影响。其中, 用户生成内容 (user generated content, UGC) [23] 作为其中典型的一类大数据, 正在成为人们获取知识的重要来源。由于具有开放和协作等特点, 这类大数据极大地促进了知识的传播与应用。

UGC 中的知识大多是以碎片化知识 (fragmentized knowledge) 形式存在。这里, 碎片化知识是指特定主题在某个分面 (facet) 上的知识。例如, “痛风”这一知识主题具有“病因”、“症状”、“并发症”和“治疗”等分面, 关于“痛风症状”的知识是“痛风”的一个碎片化知识。碎片化知识以文本、图像

等不同媒体形式高度分布在不同数据源的 UGC 中。

充分利用 UGC 中的海量碎片化知识有助于突破传统知识工程中的“知识获取”^[24]和“知识再工程”^[19]两个瓶颈问题。利用碎片化知识,需要构建一种大数据驱动的知识自动化(简称为互联网+的知识自动化)新模式,该模式有望建立“数据-知识-智慧”转化的桥梁。

这种新模式具有融合群体智慧的特点,单一的碎片化知识存在完备性、一致性等质量问题,但海量的碎片化知识蕴含了大量用户对事物的认识。利用群体智能理论方法^[25~27]有望实现信息从量到质的转变从而获取有价值的知识,实现碎片化知识替代专家知识,建立与专家知识质量相当的知识库。除此之外,以数据驱动的知识自动化有利于知识库的更新迭代,具有自我进化的特点。UGC 是不断累积动态变化的,从而带动数据源的不断更新,对完善知识库适应知识快速增长与剔除冗余具有积极作用。同时实现了用户共同参与动态知识维护,避免专家知识因少数专家的倾向性而产生的局限性。传统的专家知识所建立的知识库更多关注于特定领域的知识构建而忽略了知识间的相互关联,难以完成知识的推理演化。UGC 的碎片化知识在互联网+的知识自动化中可以通过信息的关联、融合和推理等操作生成新知识,提升系统的处理能力,实现知识推理演化。

然而,碎片化知识自身特点以及用户需求的多样化,导致互联网+的知识自动化在技术与模式上还面临一系列挑战。例如,多源分布的碎片化知识的多模态、多粒度、时变等特性对知识的准确表示提出了更高的要求。同时,碎片化知识的真实性、无序性也导致当前的研究成果不能适应用户需求的多样性与场景的多变性,而互联网+的知识自动化对于解决传统知识工程的局限性问题具有积极意义。

3 碎片化知识的学习与组织

3.1 碎片化知识的基本属性

近些年来,由于互联网、人工智能、信息科学等科学技术的进步与发展,使得信息知识数据得到了大量的积累。同时,数据的复杂程度加深,信息的可靠性减弱,有价值知识的提取难度逐渐增大。人类被带入了碎片化的时代^[28,29],即时间的碎片化与知识的碎片化,如何充分使用这些碎片化的信息是人工智能未来亟须面对的难题之一。碎片化知识具有多源分布、传播的社会性、无序与非完整性、冗余与隐喻等特点。

3.1.1 多源分布

碎片化知识分布零散,且信息来自多源(如图2所示),知识隐藏在零碎的信息片段之中,并在不同的信息源与接收者之间进行传递。例如,与人类生活息息相关的交通信息^[30,31],它可以来自于媒体平台、社交网络以及人们的经验判断,这些不同来源的信息片段最终共同构成了更加可靠的交通信息知识库。再如生物医疗领域中,大量的碎片化知识分布广泛,来源多样,相互联系错综复杂,如图3所示。

我们可以对特定领域的知识加以归类区分,并进行学习研究,但信息的来源复杂多样使得寻求特定的内在的联系成为建立智能知识库所面临的重要问题^[32]。多源信息的融合^[33]是描述碎片化知识的有效方法,并且在图像认知、信号追踪等领域得到了发展与积累^[34~36]。

3.1.2 传播的社会性

碎片化知识传播的社会性^[37~39]特性主要表现为在相同的主题下,知识的内容来自不同地区、不

图 2 (网络版彩图) 特定领域知识图谱

Figure 2 (Color online) Domain-specific knowledge map

图 3 (网络版彩图) 多源数据

Figure 3 (Color online) Multi-source data

同群体、不同个体和不同时间, 在各种社交平台或网站传播, 这是一种碎片化知识的社会性形态, 而这些来源不同且看似没有关联的信息往往又存在着某种复杂的内部关联。例如, 在社交平台上, 来自不

同行业与不同地域的人共享了大量的信息, 其中不乏与经济金融有关的内容. 对于相同的内容, 在传播的过程中, 不同群体的关注点不同, 往往会使得碎片化知识呈现出不同的社会形态. 农村居民更关注农产品市场价格; 普通群众的关注点主要集中在物价的变化; 金融从业者更多地分析证券走势; 商业人士则在分享投资机遇、对比收益增减; 城市居民关注工资收益、房价变化. 这些碎片化知识从不同侧面描绘了一个国家或地区的经济走势, 体现了其经济状况. 同样, 碎片化知识的传播社会性在医疗领域也有很好的体现, 人们可以通过互联网自发地传递不同医学专家所积累的临床知识和患者的体验与感受, 同时也可以通过网络寻求自己希望了解的医学知识.

3.1.3 无序与非完整性

由于碎片化知识的多源分布与传播的社会性, 其内容表征之间也存在不同的方式且具有局限性, 导致知识之间的相关程度呈现出无序与非完整性^[40]的特征. 人工智能系统需要对碎片化知识进行有序化组织, 面向特定用户的需求与场景推送需要的及明确的知识. 传统的数据挖掘算法难以完全适用于非有序性的碎片化知识, 碎片化知识可能包含文字、数据、表格、图片、视频等任何形式的信息. 不同数据源之间的关系也较为隐藏. 尽管通过神经网络与深度学习的方法已经能够实现针对特定数据类型碎片化知识的有效学习检索, 但如何从非完整性知识的表象或特征学习到综合知识推理, 进而实现知识的结构化预测, 这是一个值得深入研究的方向.

3.1.4 冗余与隐喻

互联网的发展让任何人在任何地点都可以通过网络表达传递自己的观点, 它们构成了碎片化知识. 其中出现大量的冗余信息是不可避免的, 同时信息的表达往往具有隐喻的性质. 因此有价值的知识会间接隐藏在繁杂的内容当中.

冗余和隐喻^[41]是碎片化知识数据的低质、低效和混乱之源, 带来碎片化知识的真实性、完备性等问题. 如果不能很好地处理碎片化知识的冗余和隐喻, 则会对碎片化知识的合理利用带来负面影响.

Google 的流感预测系统 (Google flu trends, GFT)^[42]的预测结果严重偏离真实情况就是典型的碎片化知识的冗余造成的. Google 在 2009 年 *Nature* 上发表论文介绍了该系统, 曾经因其精准的预测 (2008 年预测数据) 轰动一时. 然而到了 2013 年, 该系统再次登上了新闻头条, 而原因则是其长时间错误的预测, 对民众和社会造成了一定的负面影响. 2014 年, Lazer 等^[43]分析了大数据的预测为何会背离事实. 因此在利用碎片化知识的时候, 应该尤其关注碎片化知识的冗余和隐喻等特性. 也正是由于碎片化知识的冗余, 碎片化知识并不应该是传统知识获取方法的替代品, 而应该是一个合理的补充. 因此, 如何将碎片化知识与传统知识进行合理结合, 做出精确的判别, 则是在碎片化知识环境下, 人工智能领域所需要面对的重要问题.

传统的数据挖掘方法无法有效地处理碎片化知识, 对碎片化知识的学习挖掘与表征描述是知识发现的关键. 2014 年, Wu 等^[44]介绍了大数据背景下的数据挖掘, 提出了 HACE (heterogeneous, autonomous, complex, evolving) 理论, 描绘了大数据的异质性、自治性、复杂与进化的特点. 碎片化知识是大数据的重要组成部分, HACE 理论很好地刻画了碎片化知识的属性, 对其学习挖掘与表征描述提供了有效方法.

3.2 碎片化知识的学习挖掘与有序组织

人类对知识的记忆, 通常需要学习的过程, 并且人类对于有组织性的内容学习相对容易且记忆深

图 4 (网络版彩图) 碎片化知识的组织与学习

Figure 4 (Color online) Organization and learning of fragmented knowledge

刻. 对机器而言, 这种规律尤为明显. 无论是 Bayes 算法^[45,46] 还是神经网络算法^[47~51] 都需要通过学习才能进行实际应用. 学习是对知识进行归纳的过程, 被归纳的知识更容易形成组织化结构. 传统机器的存储记忆对知识的组织化要求苛刻, 具有明显的结构特征, 否则无法保证信息被准确地存储或查询.

要建立完备的碎片化知识的归纳与组织体系, 需要对知识的组成特性与发展规律有充分的认识, 根据其特性规律结合相应的理论模型才有可能找到合适的方法. 这里需要运用知识世系模型和知识回溯探究知识演化规律, 然后依据知识演化规律, 对知识置信度进行评估, 提取所需要的知识. 图 4 描绘了碎片化知识的组织与学习的基本结构.

传统的人工智能主要基于特征学习^[52,53] 和定量学习的模式实现规则推理, 它需要对问题本身抽象出一个精确数学意义上的解析式的数学模型, 若无法抽象出这样的模型则归纳为不可解问题. 然而碎片化知识是具有动态时空关联性的, 抽象出精确的数学模型几乎是不可能的. 同时, 对数学模型的求解算法设计易引入诸如 NPC (non-deterministic polynomial complete problem) 类问题. 碎片化知识由于存在信息的冗余与隐喻问题, 数据的质量密度低, 无法进行向量化, 传统的人工智能用“度量”来区分模式的方法无法有效处理, 如图 5 所示. 碎片化知识通常都是无序的、内部关联稀疏的, 与用户的自身知识背景、认知感受密切相关, 而图灵意义下的可计算问题都是可递归的, 即有序的, 最终得到唯一的结果, 这与现实世界所固有的测不准性与不完备性相违背.

这些局限性使得传统的人工智能计算模型无法有效处理碎片化知识, 其中一个典型的应用案例为医疗碎片化知识处理. 在医学上, “同病不同因”与“同因不同病”是一种基本常识. 当人工智能系统对病人的病因和治疗方案进行决策时, 需要处理大量的非有序、难于表达的碎片化知识. 这些知识有的来源于病人本身的生理化验结果、医生的临床诊断结果以及过往的医疗专家数据, 而更多的是来自于互联网上患者自身的非有序的、难于组织的感受知识. 如何将这些庞大且复杂的碎片化知识进行合理的整合, 即去除冗余和低价值的知识, 对有用的知识进行统一化的表达, 并进行合理的学习与组织, 将是未来人工智能在处理医疗领域碎片化知识时需要面对的重要问题. 人工智能未来在医疗领域需要解决的是对疾病的有效预防, 寻找人体健康的临界状态、外界环境、生活习惯、工作强度等多方面对健康的综合影响问题.

4 网络化人工智能

碎片化知识如何有效地组织表示是未来人工智能首先需要解决的问题. 由于碎片化知识结构相对耦合松散, 关系相对复杂, 且无论是数据还是求解过程都表现为网络拓扑结构^[8], 其系统的输出则以

图 5 (网络版彩图) 传统人工智能方法无法有效处理碎片化知识

Figure 5 (Color online) Traditional artificial intelligence methods cannot effectively handle fragmented knowledge

图 6 (网络版彩图) 网络化人工智能的基本过程

Figure 6 (Color online) The basic process of networked artificial intelligence

知识图谱的形式给出,称之为“网络化人工智能”(如图 6 所示). 它可实现多源或复杂系统(确定领域的知识自动化,如自动驾驶、医疗健康、国防等)的认知处理^[54~56].

网络化人工智能的核心是如何利用大规模网络化碎片知识实现人工智能,在自治数据源的基础上实现表达演化建立拓扑结构完成知识自动化,其可以分解为多个方面的科学问题(如图 7 所示).

4.1 群智网络表征

知识特征表达是网络化人工智能的基础. 碎片化知识具有多源分布性、传播的社会性、冗余与隐喻性以及无序与非完整性的特征. 这些特征决定了传统的向量化方法不能有效表达和处理碎片化知识. 因此,如何以可存储、可计算的方式表达碎片化知识的本质特征和信息是网络化人工智能面临的首要问题.

4.1.1 群智网络拓扑表征

网络化人工智能系统是一个复杂的群智网络. 该网络包括结点、连接、信息传播、动态演化、外界交互等属性. 如何设计及表达这种复杂的拓扑结构是网络化人工智能面临的一个重要问题.

图 7 (网络版彩图) 网络化人工智能研究的科学问题

Figure 7 (Color online) The scientific problems of networked artificial intelligence

图 8 (网络版彩图) 群智网络的拓扑表征示例

Figure 8 (Color online) An example of topological representation for collective intelligence network

群智网络以群体智慧为基础, 碎片化知识是群体智慧的具体表现, 碎片化知识根据自身的相关性可以构成某一领域知识的结点, 各个领域通过信息的关联进行连接实现信息传播互换. 图 8 以痛风病为例说明了一种群智网络拓扑表征的结构.

图 9 (网络版彩图) 知识图谱的结构表征

Figure 9 (Color online) Structural representation of knowledge map

4.1.2 知识图谱结构表征

与传统人工智能的输出确定性向量描述不同, 网络化人工智能输出结构化知识图谱 (如图 9 所示). 知识图谱^[57~59] 蕴含更丰富的信息, 具有更强的非确定性信息描述能力, 以及具有更强的可拓展性. 但同时, 其也具有更复杂的连接和结构. 因此, 如何设计及表征结构化知识图谱是网络化人工智能需要研究的关键问题. 以碎片化知识为节点, 多个强关联的节点组成知识簇. 知识簇又通过连接形成网络化的知识图谱. 知识簇的合理连接有利于信息的融合演化, 多个知识簇的共同作用是进化出新知识的有效方法, 同时, 知识簇间的相互作用也可以完善自身的不足, 弥补碎片化知识本身所具有的不准确性与不完备性. 这样的表征结构实现了以碎片化知识为基础的信息, 可动态更新维护, 并且易于推理分析.

4.2 从机器学习到机器推理

人工智能目标的实现依赖于问题的可计算性, 建立有效的计算模型是解决人工智能问题的关键. 网络化人工智能具有知识碎片化、网络群智化、输出图谱化等几个特点. 这几个特点造成了传统的可计算模型无法充分有效地描述网络化人工智能结构. 因此, 建立合理的计算模型和理论是网络化人工智能研究的科学问题之一.

传统的计算模型大多依赖于算法的迭代更新完成参数的学习^[60,61], 实现机器学习. 未来的网络化人工智能将不止停留在机器学习阶段, 需要通过新型的计算模型拓展到机器推理阶段. 机器推理不仅局限于对现有知识的学习, 完成学习后还能够研究隐藏在现有知识中的内部关联, 进而推断并发现新知识. 这也是实现机器自我智能的基础.

4.2.1 模型学习与知识发现

模型学习研究的是如何从大量的碎片化样本中学习模型的参数、结构和形式等. 传统的机器学习方法以数据为输入, 网络化人工智能中以碎片化知识为输入. 如何从碎片化知识中学习更高层的知识以使人工智能系统具有更大的智慧是模型学习^[62] 和知识发现^[63] 需要研究的问题.

4.2.2 机器推理与知识演绎

机器推理是研究如何在一个具有不完整信息下求解问题, 知识演绎研究模型的可扩展性和知识增殖性. 模型学习与知识发现使网络人工智能的计算模型具有了可拓展条件. 因此, 通过学习得到的高

层次知识并不是静态和一成不变的,它应具有可拓展性和增殖性.如何在大规模的碎片化知识环境下进行机器推理^[64]和知识演绎^[65]是网络化人工智能计算要解决的科学问题.

4.3 计算模型与直觉融合

人与机器交互的环境以及互联网中存在的海量视听觉等非结构化大数据,将人对这类数据的直觉感知与智能机器在逻辑推理、演绎、归纳等计算处理的优势相结合,探索高效人机协同计算方法,实现类人的大规模视听觉数据的理解成为计算模型与直觉融合需要解决的问题.人类的视听觉是认知外部世界的主要途径,外部的刺激信号经过传输进入大脑后形成知识,对人的直觉感知产生影响.智能机器的主要优势在于对数据的精确计算与严谨的逻辑推理,将二者融合,充分发挥各自优势,有利于机器进一步接近人类智能.

4.4 认知过载与冗余剔除

大规模无序碎片化知识条件下易形成知识的溢出及误导,造成“认知过载”问题^[66].海量碎片化知识在信息质量方面存在着缺陷,数据量的巨大并不能代表其内容的可靠程度,甚至可能出现数量占优的信息限制了认知的全面性造成过载或形成误导的现象.另一方面,独立的碎片化知识具隐喻的特点,不能直接地表达所蕴含知识内容,同时不同的碎片化知识表述形式有所不同,但可能是在描述同一概念形成的信息冗余.如何将大规模知识的过载与独立知识的隐喻性融合,剔除冗余信息以实现知识的有序组织和表达是网络化人工智能面临的一个重要问题.

5 大数据演化规律的公式发现

5.1 公式发现领域研究现状

寻求大数据的演化规律是机器推理的主要研究内容之一,其中公式发现是大数据机器推理和知识自动化领域的重要科学问题,并在众多其他领域中有着广泛的应用^[67,68].对于自变量 x 和因变量 y ,公式发现的目标就是找到合适的函数 f 使得 $f(x, y) = 0$ 能够较好地描述变量 x 和 y 在不同状态的观测值.随着人们对知识推理的深入研究,已有多个基于观测数据的公式发现系统被提出,例如,Langley 和 Zytkow^[69]开发的 BACON 系统和 Chen 等^[70]提出的 FDD (formula discovery from data) 系统.这些方法的提出推动了公式发现的研究,但人们仍然希望能够找到一种可以有效揭示多变量公式的公式发现方法.

事实上,已有的基于相关系数、熵和多项式拟合的方法,仅考虑了两个变量之间的依赖关系或者仅对单变量函数有效.当变量增多时,往往只能固定其他变量而仅考虑一个变量方向的数据变动情况^[71,72].因此,传统方法在实际应用中具有相当大的局限性.随着知识碎片化程度的深入,信息之间的关联呈现愈加复杂的态势,这对传统的公式发现方法提出了更高的要求.例如,相对于双变量的情形,人们更加关注多个变量数据之间的内在规律性.能否提出一种适用于处理碎片化知识的理论或框架,是突破当前研究瓶颈的关键.

5.2 对公式发现的新要求

为了能够在碎片化知识中寻找数据之间的规律性和数据的公式化产生规则,我们需要更加有效的公式发现工具.而在大数据背景和多变量公式的需求下,这种有效性至少应该满足以下 3 个方面:

表 1 黑箱数据
Table 1 Black-box data

α	x	y	α	x	y
10.1	0.2	16.8	44.2	1.2	12.7
13.1	2.9	2.0	46.9	2.3	7.4
15.5	1.1	6.8	47.2	4.0	3.1
18.6	0.4	15.9	48.8	0.4	23.4
18.9	0.1	27.5	51.0	0.2	28.6
21.2	2.0	4.9	51.5	2.0	9.0
24.5	0.9	11.1	53.7	2.9	6.0
26.6	3.3	3.0	55.1	0.8	17.9
28.4	2.7	4.3	61.1	0.4	24.6
28.8	1.2	9.9	69.5	1.1	16.0
28.9	0.6	15.9	71.3	3.3	5.6
30.0	5.0	1.0	72.4	0.9	18.1
35.0	0.5	19.1	74.6	0.9	18.2
36.1	1.6	9.0	80.7	0.2	29.9
37.8	1.0	13.3	85.1	0.3	27.6
40.9	1.2	12.2	90.0	2.5	8.5

多变量要求: 在当前背景下, 知识与信息呈现出愈发复杂的关联关系, 仅考虑两个变量的系统是不具有实用性的. 可以处理多关联、多纠缠信息的公式发现方法亟待被提出.

强表征能力: 知识的复杂关联不仅体现在多变量层面, 也体现在函数形式本身的复杂性上. 单一函数已经不足以表达这种复杂的关系, 我们需要拥有更加丰富表征能力的系统.

高计算效率: 随着大数据影响的深入, 我们被淹没在信息的海洋中, 却渴望知识的淡水. 适用于大规模数据集的公式发现架构才是具有实用价值的.

5.3 基于多元相关系数的公式发现

作者的工作中, 使用了 MCC (multivariate correlation coefficient) 和 MIC (multivariate incorrelation coefficient) 作为多元相关性的度量标准^[73]. MCC 和 MIC 是一对耦合的系数, MIC 越小, 则相关性越大. 该方法天然满足多变量和高计算效率的要求, 因为 MIC 就是为这种环境而设计的指标.

而对于上述“强表征能力”的要求, MIC 可以直接应用于线性公式, 也可以很容易地拓展到多项式公式以及更一般的公式发现领域中. 该方法的知識表征能力是其他方法所不具备的.

5.4 公式发现的应用

5.4.1 黑箱系统探测

我们构造了一个黑箱, 产生了表 1 中的数据. 对于这个黑箱系统, 我们只有这样的先验知识, 即 α (代表角度). 图 10(a)~10(c) 分别给出了黑箱系统两两变量之间的关系. 由它们导出黑箱产生数据的公式是非常困难的, 但是利用基于 MIC 的方法 (如图 10(d) 所示), 我们可以很容易地得到三者之间的关系 $f(\alpha, x, y)$, 进而可以用一个公式描述这个黑箱.

图 10 (网络版彩图) 黑箱数据

Figure 10 (Color online) Black-box data. (a) represents the projection of α onto the x plane; (b) represents the projection of α onto the y plane; (c) represents the projection of x onto the y plane; (d) represents a three-dimensional plane obtained by the MIC-based method

表 2 WHO 数据集
Table 2 WHO data set

Variables	MIC	Equations	Data set
x_1, y_1, z_1	0.00030	$x_1 = y_1 - y_1 z_1$	WHO
x_2, y_2, z_2	0.00006	$\ln x_1 - \ln y_1 = \ln z_1$	UNdata

x_1 : Per capita government expenditure on health (PPP int. \$)
 y_1 : Per capita total expenditure on health (PPP int. \$)
 z_1 : Private expenditure on health pct. total expenditure on health (%)
 x_2 : CO₂ emissions (thousand metric tons)
 y_2 : CO₂ emissions per capita (metric tons)
 z_2 : Total population (in thousands)

依据先验知识, 我们可以构造下列公式字典:

$$\cos \alpha, \sin \alpha, x \cos \alpha, x \sin \alpha, y, y \cos \alpha, y \sin \alpha, x^2, x^2 \cos \alpha, x^2 \sin \alpha, y^2, y^2 \cos \alpha, y^2 \sin \alpha, xy, xy \cos \alpha, xy \sin \alpha.$$

随后, 根据字典, 计算所有组合的 MIC, 得到 $x^2 \sin \alpha + xy + y \sin \alpha = 36 \sin \alpha$. 事实上, 这正是黑箱产生数据的公式. 它证明了 MIC 算法的有效性.

5.4.2 真实大数据集公式发现

世界卫生组织 (WHO) 数据集 (如表 2 所示), 包含了 202 个国家的 356 组数据. 联合国 (UNdata) 数据集包含了 34 个数据库, 600 多万条记录. 我们以这两个数据集为输入, 得到了一些公式. 这些公式具有直观的物理意义. 但是还有许多不具有直观物理意义的公式被发现出来, 这就为相关领域方面的学者提供了启发与进一步研究的方向.

6 网络化人工智能的未来应用

无论是碎片化知识的学习与组织, 还是网络化人工智能的发展, 都将以互联网为基础平台, 并且人

图 11 采用卷积神经网络的深度学习结构

Figure 11 A deep learning structure based on convolution neural

工智能将更加趋近于智能化, 从思维模式到行为表现也将趋于类人化. 目前, 人工智能在诸多领域已经得到了广泛应用, 但仅局限于特定领域, 并且对数据集要求苛刻, 在实际应用中受到了诸多的限制. 同时, 我们也发现在人工智能领域发展的过程中还有很多方法值得去探索.

6.1 快速在线学习分类

人类目前对智能这一概念的定义还相对较为模糊, 但普遍的认识是, 智能是从学习中得到的. 人类对新鲜事物的学习通常首先需要观察事物本身的某些性质进行记忆, 再通过反复的练习增强记忆, 最后将其转化为自己的知识. 人工智能借鉴了人类学习的过程, 并希望通过这样的学习过程使机器也同样获得智能. 这样的模拟在一定程度上取得了成功. 例如, 2012 年 Hinton 及其团队使用卷积神经网络 (convolutional neural network, CNN^[74]) 大幅提高了对 ImageNET 数据集^[75] 识别的准确度^[76], 神经网络即是根据人类大脑的结构而提出的 (如图 11 所示). 2015 年 12 月举行的 ImageNet 和 MS COCO 计算机视觉挑战赛上, 来自微软亚洲研究院的孙剑团队采用 152 层深度卷积神经网络获得了超过人眼辨识的正确率^[77].

利用卷积神经网络是目前实现深度学习的一种比较成熟而且有效的方法. 通过大量的学习与层数的不断加多, 这种方法给人们带来了一个又一个的惊喜, 并在一定程度上解决了图像分类、物体识别等问题^[78~82].

与深度学习不同, 人类的学习往往是小样本多样化的, 并且消耗资源极小, 但目前的神经网络尚且无法做到这些^[83,84]. 并且神经网络存在过度拟合的风险, 只有采集到充分多的标注数据且数据维度足够高时, 才能缓解复杂模型的过度学习问题. 因此, 深度学习性能依赖于海量的学习样本, 在小样本数据下无法获得有效的知识. 由于深度学习神经网络对大样本空间学习的依赖程度较高, 2015 年 12 月 *Science* 杂志上登出的国外学者通过概率推理针对“单样本学习”方法的文献^[85] 显得更加有意义. 由于先验知识的存在, Bayes 方法非常适合小样本的学习, 但其同时又引入了先验知识的需求.

已有的成熟机器学习算法主要以有监督学习为主, 与人类相比缺少自主性, 将非监督学习的自组织神经网络 (self-organizing maps, SOM) 算法^[86,87] 应用在学习分类当中可以增强机器分类的自主性.

在这方面, 我们已经使用 SOM 算法在图像场景分类与 fMRI (functional magnetic resonance imaging) 脑信号数据重构中进行了尝试并得到了预期的效果.

目前, 深度学习还是停留在统计学习和复杂模式识别与分类层面上, 比起人的学习能力还有很多局限. 例如, 人的举一反三、触类旁通、无师自通的学习能力是现有统计学习所远不能达到的. 深度模型与知识的融合、外部记忆的增强、深度学习与推理的结合应该是其未来的研究方向.

6.2 精准治疗与预防医疗

人类的生命健康已经成为人工智能研究重点关注的领域之一^[88], 利用人工智能技术可以从医疗领域产生的大规模数据中挖掘出有关疾病治疗和预防的知识信息, 进而实现精准治疗和预防医疗.

作为认知计算系统的杰出代表, IBM (international business machines corporation) 的智能问答系统 Watson^[89,90] 在危机边缘智力竞赛节目中战胜人类选手之后已经被应用于了医学治疗领域. 德州大学安德森癌症中心通过 Watson 的认知计算能力, 从患者病历和丰富的研究成果库中寻找资料, 为临床医生提供有价值的信息, 帮助医护人员寻找最有效的治疗方案. 虽然目前类似于 Watson 这样的平台还不能独立地做出决策, 最终的治疗方案还需要医师来决定, 但是它们已经验证了人工智能在医疗领域能够给予人类巨大的帮助.

在当今社会, 人类对自身健康问题的重视程度在逐步增加, 医疗也从治疗型向预防型的方向转变. 人工智能在预防医疗领域近些年已经受到部分研究机构的关注. 1995 年美国的芝加哥热浪事件之后, 人们开始思考研究气候变化与人体健康之间的关系, 随后发现类似事件的出现并非偶然, 在其他地区也有发生由于温度的升高造成的疾病甚至死亡的现象^[91]. 但这些现象的出现仅靠人工的方法去分析天气与人体健康之间的联系是十分困难的, 网络化人工智能的应用将更易于分析揭示它们之间的关联. 不仅是天气, 饮食习惯、生活作息、人际关系、社会压力都可能会影响我们的身体健康, 而这些信息用传统的方法进行统计分析几乎是不可能的, 但它们往往存在于互联网的碎片化知识当中, 网络化人工智能将是有效的分析手段.

人工智能在医疗领域未来的发展依赖于有效的碎片化知识处理机制. 未来不仅需要解决临床医学问题, 而且要更加趋向于预测性研究. 下面通过一个例子来说明通过跨域知识无序与非完整性, 寻求新的知识发现. 丙肝与肌无力在医学领域属于两个完全不同的病症, 但我们在网络中进行查找发现, 有大量的用户在网络中咨询有关丙肝与四肢无力之间的关系, 然而对此的专业知识却少之又少, 在查阅了大量资料后我们发现, 早在 1996 年就有医学界的专家从专业角度对丙肝病毒与肌无力之间的联系进行过分析并发表了文献^[92], 但研究深度有限并未引起足够重视. 事实上, 有关的信息在互联网产生的碎片化知识中体现的更多, 这些碎片化知识可能来自患者或其亲友而非专业的医务工作者. 如何从碎片化知识中有效提取类似于丙肝与肌无力之间联系的有价值信息是人工智能未来必须面对的难题.

6.3 自博弈训练与国防

2016 年 3 月, Google DeepMind 研发的 Alphago 程序在围棋上成功战胜了韩国著名围棋职业九段选手李世石举世瞩目. 围棋是人类发明的最为复杂多变的棋牌类游戏, 并且源自于古代战争. Alphago 证明了人工智能具有自主学习的能力, 它通过对大量历史棋局的学习与自我博弈形成了自我学习过程^[93]. 既然智能机器能够在棋盘上战胜人类, 也就有理由相信它们也将在战场上把人类淘汰出局. 如果我们建立类似的系统, 通过深度学习与 Monte-Carlo 模型^[94,95] (如图 12 所示) 建立战场决策分析的平行军事系统, 则可对历史战场或演习数据进行学习, 模拟博弈并发现现有国防体系中存在

图 12 Monte-Carlo 树搜索模型. 每一个节点包含基于仿真结果的评估值与该节点被访问的次数这两部分内容. (a) Selection: 从根节点开始, 递归选择最优子节点, 直到到达叶子节点为止; (b) Expansion: 如果叶子节点非终止节点, 则创建一个或多个子节点并选择其中一个; (c) Simulation: 从所选择的节点开始仿真运行直到得到结果; (d) Backpropagation: 更新当前序列与仿真结果

Figure 12 Monte-Carlo tree search model. Each node contains an evaluation value based on the simulation results and the accessed times of the node. (a) Selection: From the root node down to the leaf nodes, recursively selecting the prior child nodes. (b) Expansion: If the leaf node is not a terminal node, creating one or more child nodes and selecting one. (c) Simulation: Running a simulation at the selected node. (d) Backpropagation: Updating the current sequence and simulation results

的不足之处以便及时加以改进. 同样地, 对未来战争的胜率也可以进行提前预测, 通过合理的调整确保国防体系的可靠性.

如上文所述, 近年来, 将人工智能方法应用于国防安全领域, 受到人们关注. 全球各国对安全问题高度重视, 发展面向信息物理系统 (cyber-physical systems, CPS) [96] 的军事体系成为重要的国防任务, 但 CPS 系统还缺少社会及人的因素 S (social) 的参与, 未来需要将这个重要因素加入其中从而组成面向社会物理信息系统 CPSS (cyber-physical-social systems) [97] (如图 13 所示) 的军事, 以便应对更加复杂的战争需要. 通过近年来的反恐战争、军事侦察反侦察、联合军事演习救援等活动的经验表明, CPSS 使得物理域、网络域和感知域等进行跨域联合作战成为现实.

现代战争已不仅是传统意义上的物理空间的作战, 网络环境成为另一个作战的主要空间, 同时社会环境与人类活动的因素也不可忽视. 在这类复杂空间环境中, 为实现针对多位一体的新型作战形式及演习训练的有效指挥控制, 催生了“平行军事体系”的概念. 平行军事体系的特点是以数据为驱动, 搭建模拟军事组织系统, 与实际军事组织系统进行互动.

复杂的实时战场环境事态依靠传统的人为因素进行分析判断已经无法完全胜任. 平行军事系统通过全方位的数据分析, 对大量数据处理得到碎片化知识, 通过建立网络化模型能够更加及时有效地传递战场信息. 平行军事系统不仅仅能够对战时战场指挥决策起到帮助, 它也是很好的军事训练平台.

6.4 智能农业

人类在开始学会使用工具之初就与获取食物有关, 随后人类的进步也大多与农业生产息息相关, 人类每一次的科技进步也同样会带动农业的发展. 人工智能在农业生产当中也得到了广泛的应用 [98], 如最常见的各类智能农业生产收割工具、智能引水灌溉技术和可控温室建设等.

图 13 (网络版彩图) CPSS 指挥系统

Figure 13 (Color online) CPSS command system

在现代化农业应用当中,早在 20 世纪 80 年代,人们就使用当时盛行的专家系统,帮助农业生产者做出更合理和有效的决策^[99].人们可以通过人工智能和计算机视觉方法,自动区分杂草和农作物,在很大程度上减少繁琐的人工劳动^[100].同样,一些大数据和人工智能分析方法也能够在自然和生态环境的管理上起到重要作用^[101],使农业生产可以更合理地利用资源,保证农业生产做到环境友好.传统人工智能方法当前在智能农业中的应用,起到了将人类从重复工作中解放出来的作用,由机器替代人完成简单重复的事情,同时通过智能监测调控改善生产过程中农作物所需的环境要求.这样的应用目前对农业生产的促进作用是有限的,农业生产过程中更多的信息还没有完全被利用,这些知识是以碎片化的形式存在的.

现代农业生产在依靠土壤环境与气候变化的同时,也需要密切关注人们的需求.前者通过传统人工智能的分析即可为农业生产工作者提供有价值的信息,但后者使用传统的方法无法有效处理,而网络化人工智能对此将更加有效.例如,从碎片化知识中提取人们在某段时间内对各类水果的兴趣以及该水果未来可能的种植数量作为指导,准确地引导果农合理进行种植.这样,不但可以有效地提高收益,还可避免过剩种植造成的损失.

7 结束语

人工智能在互联网的大环境下,已经涉及了人类生活、工作、学习等方方面面,人工智能所要面对的是具有多源分布、传播的社会性、无序与非完整性、冗余与隐喻等特性的碎片化知识,如何有效地对碎片化知识进行归纳组织是关键性难题.本文提出了建立网络化人工智能的思想,从碎片化知识中获取有价值信息实现知识自动化,提出了群智网络拓扑表征以及从机器学习到机器推理等需要解决的一系列重要问题,分析了计算推理与直觉融合、认知过载与冗余剔除的必要性.在大数据的背景下,公式发现是推动知识自动化发展的重要方法,有效的公式发现能够更加直观地表达隐藏在大数据内部

的知识。人工智能在分类学习、医疗、国防、农业等领域的应用已经受到了关注, 但与人类智能相比, 人工智能还有自身的局限性。本文所介绍的面向碎片化知识处理的大数据推动网络化人工智能还需要在理论与实践两个方面进一步发展, 以解决智能 + 知识自动化所面临的一系列重大问题。

参考文献

- 1 Kompella V P, Pasquale J C, Polyzos G C. Multicast routing for multimedia communication. *IEEE/ACM Trans Netw*, 2015, 1: 286–292
- 2 Akyildiz I F, Melodia T, Chowdury K R. Wireless multimedia sensor networks: a survey. *IEEE Wirel Commun*, 2008, 14: 32–39
- 3 Fargues J, Landau M C, Dugourd A, et al. Conceptual graphs for semantics and knowledge processing. *IBM J Res Dev*, 1986, 30: 70–79
- 4 Duda R O, Shortliffe E H. Expert systems research. *Science*, 1983, 220: 261–268
- 5 Liao S H. Expert system methodologies and applications—a decade review from 1995 to 2004. *Expert Syst Appl*, 2005, 28: 93–103
- 6 Lennartson B, Sundberg K. Expert system. US Patent No. 8,064,901, 2011
- 7 Zheng N N. Information processing for cognition process and new artificial intelligent systems. *China Basic Sci*, 2000, 8: 9–18 [郑南宁. 认知过程的信息处理和新型人工智能系统. *中国基础科学*, 2000, 8: 9–18]
- 8 Reagans R, Mcevely B. Network structure and knowledge transfer: the effects of cohesion and range. *Admin Sci Quart*, 2003, 48: 240–267
- 9 Holland J H. *Adaptation in Natural and Artificial Systems: an Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. Ann Arbor: University of Michigan Press, 1975
- 10 Marr D. Artificial intelligence—a personal view. *Artif Intell*, 1977, 9: 37–48
- 11 Russell S J, Norvig P, Canny J F, et al. *Artificial Intelligence: A Modern Approach*. 2nd ed. Upper Saddle River: Prentice Hall, 2003
- 12 Feigenbaum E A. The art of artificial intelligence: themes and case studies of knowledge engineering. In: *Proceedings of the 5th International Joint Conference on Artificial Intelligence*. Burlington: Morgan Kaufmann Publishers, 1977. 1014–1029
- 13 Studer R, Benjamins V R, Fensel D. Knowledge engineering: principles and methods. *Data Knowl Eng*, 1998, 25: 161–197
- 14 Lindsay R K, Buchanan B G, Feigenbaum E A, et al. DENDRAL: a case study of the first expert system for scientific hypothesis formation. *Artif Intell*, 1993, 61: 209–261
- 15 Lenat D B. CYC: a large-scale investment in knowledge infrastructure. *Commun ACM*, 1995, 38: 33–38
- 16 Ong T H, Chen H, Sung W K, et al. Newsmap: a knowledge map for online news. *Decis Support Syst*, 2005, 39: 583–597
- 17 Nickel M, Murphy K, Tresp V, et al. A review of relational machine learning for knowledge graphs: from multi-relational link prediction to automated knowledge graph construction. *ArXiv:1503.00759*, 2015
- 18 Wagner C. Breaking the knowledge acquisition bottleneck through conversational knowledge management. *Inf Resour Manage J*, 2006, 19: 70–83
- 19 Hoekstra R. The knowledge reengineering bottleneck. *Semant Web*, 2010, 1: 111–115
- 20 Gruber T R. A translation approach to portable ontology specifications. *Knowl Acquis*, 1993, 5: 199–220
- 21 Armbrust M, Fox A, Griffith R, et al. A view of cloud computing. *Commun ACM*, 2010, 53: 50–58
- 22 Atzori L, Iera A, Morabito G. The internet of things: a survey. *Comput Netw*, 2010, 54: 2787–2805
- 23 Wofford J. User-generated content. *New Media Soc*, 2012, 14: 1236–1239
- 24 Inkpen A. Learning, knowledge acquisition, and strategic alliances. *Eur Manage J*, 1998, 16: 223–229
- 25 Yamins D L K, Dicarlo J J. Using goal-driven deep learning models to understand sensory cortex. *Nat Neurosci*, 2016, 19: 356–365
- 26 Lévy P. *Collective Intelligence*. New York: Plenum/Harper Collins, 1997
- 27 Bonabeau E. Decisions 2.0: the power of collective intelligence. *MIT Sloan Manage Rev*, 2009, 50: 45
- 28 Gray P H, Meister D B. Introduction: fragmentation and integration in knowledge management research. *Inf Tech People*, 2003, 16: 259–265

- 29 Cox S A, Perkins J S. The role of E-collaboration systems in knowledge management. In: *E-Collaboration: Concepts, Methodologies, Tools, and Applications*. Hershey: IGI Global Publishes, 2009. 600–607
- 30 Li L, Wen D, Zheng N N, et al. Cognitive cars: a new frontier for ADAS research. *IEEE T Intell Transp*, 2012, 13: 395–407
- 31 Zheng N N, Tang S, Cheng H, et al. Toward intelligent driver-assistance and safety warning system. *IEEE Intell Syst*, 2004, 19: 8–11
- 32 Knoblock C A, Minton S, Ambite J L, et al. Mixed-initiative, multi-source information assistants. In: *Proceedings of the 10th International Conference on World Wide Web*. Hong Kong: ACM, 2001. 697–707
- 33 Dasarathy B V. Multi-sensor, multi-source information fusion: architecture, algorithms, and applications—a panoramic overview. In: *Proceedings of the 2nd IEEE International Conference on Computational Cybernetics*, 2004, doi: 10.1109/ICCCYB.2004.1437643
- 34 Atrey P K, Hossain M A, El Saddik A, et al. Multimodal fusion for multimedia analysis: a survey. *Multimedia Syst*, 2010, 16: 345–379
- 35 Chong C Y, Kumar S P. Sensor networks: evolution, opportunities, and challenges. *Proc IEEE*, 2003, 91: 1247–1256
- 36 Hall D L, Llinas J. An introduction to multisensor data fusion. *Proc IEEE*, 1997, 85: 6–23
- 37 Greenberg J, Pyszczynski T. The effect of an overheard ethnic slur on evaluations of the target: how to spread a social disease. *J Exp Soc Psychol*, 1985, 21: 61–72
- 38 Wu X, Liu Z. How community structure influences epidemic spread in social networks. *Phys A Stat Mech Its Appl*, 2008, 387: 623–630
- 39 Song S, Li Q, Bao H. Detecting dynamic association among twitter topic. In: *Proceedings of the 21st International Conference Companion on World Wide Web*. Lyon: ACM, 2012. 605–606
- 40 Zhu J, Lu J, Yu X. Flocking of multi-agent non-holonomic systems with proximity graphs. *IEEE Trans Circuit Syst I-Regular Paper*, 2013, 60: 199–210
- 41 Axley S R. Managerial and organizational communication in terms of the conduit metaphor. *Acad Manage Rev*, 1984, 9: 428–437
- 42 Ginsberg J, Mohebbi M H, Patel R S, et al. Detecting influenza epidemics using search engine query data. *Nature*, 2008, 457: 1012–1014
- 43 Lazer D, Kennedy R, King G, et al. The parable of google flu: traps in big data analysis. *Science*, 2014, 343: 1203–1205
- 44 Wu X, Zhu X, Wu G Q, et al. Data mining with big data. *IEEE Trans Knowl Data Eng*, 2014, 26: 97–107
- 45 Cheeseman P, Stutz J. Bayesian classification (AutoClass): theory and results. In: *Advances in Knowledge Discovery and Data Mining*. Menlo Park: AAAI Press, 1996
- 46 Jeffrey R C. *The Logic of Decision*. New York: McGraw-Hill Book Co. Inc., 1965
- 47 Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. *J Biomed Inf*, 2002, 35: 352–359
- 48 LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*, 2015, 521: 436–444
- 49 Rowley H A, Baluja S, Kanade T. Neural network-based face detection. *IEEE Trans Pattern Anal Mach Intell*, 1998, 20: 23–38
- 50 Hecht-Nielsen R. Theory of the backpropagation neural network. In: *Proceedings of International 1989 Joint Conference on Neural Networks*. Piscataway: IEEE, 1989. 593–605
- 51 Kawato M, Furukawa K, Suzuki R. A hierarchical neural-network model for control and learning of voluntary movement. *Biol Cybern*, 1987, 57: 169–185
- 52 Piater J H, Grupen R A. Feature learning for recognition with Bayesian networks. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society, 2000. 17–20
- 53 Lee H, Pham P, Largman Y, et al. Unsupervised feature learning for audio classification using convolutional deep belief networks. In: *Proceedings of the 23rd Annual Conference on Neural Information Processing Systems*, Vancouver, 2009. 1096–1104
- 54 Wang Y X. The theoretical framework and cognitive process of learning. In: *Proceedings of the 6th IEEE International Conference on Cognitive Informatics*, California, 2007. 470–479
- 55 Luger G F. *Artificial Intelligence: Structures and Strategies for Complex Problem Solving*. New York: Pearson

- Education, 2005
- 56 Brachman R J. What's in a concept: structural foundations for semantic networks. *Int J Man-Mach Stud*, 1977, 9: 127–152
- 57 Lu C Y. Knowledge map: an approach to knowledge acquisition in developing engineering expert systems. *Eng Comput*, 1987, 3: 59–68
- 58 Shiffrin R M, Börner K. Mapping knowledge domains. *Proc Natl Acad Sci*, 2004, 101: 5183–5185
- 59 Chen Y, Liu Z Y. The rise of mapping knowledge domain. *Stud Sci Sci*, 2005, 23: 149–154
- 60 Zhou D, Bousquet O, Lal T N, et al. Learning with local and global consistency. *Adv Neural Inf Proc Syst*, 2004, 16: 321–328
- 61 Ahn H S, Chen Y Q, Moore K L. Iterative learning control: brief survey and categorization. *IEEE Trans Syst Man Cy C*, 2007, 37: 1099–1149
- 62 Nguyen-Tuong D, Peters J. Model learning for robot control: a survey. *Cogn Process*, 2011, 12: 319–340
- 63 Fayyad U. From data mining to knowledge discovery in databases. *AI Mag*, 1996, 17: 37–54
- 64 Ishizaka H, Arikawa S. Model inference. *Ipsj Magazine*, 1991, 32: 236–245
- 65 Yang C, Wang G, Li Y, et al. Study on knowledge reasoning based on extended formulas. *IFIP Intl Federat Inf Proc*, 2005, 187: 797–805
- 66 Kirsh D. A few thoughts on cognitive overload. *Intellectica*, 2000, 30: 41–42
- 67 Tene O, Polonetsky J. Big data for all: privacy and user control in the age of analytics. *Nw J Tech Intell Prop*, 2012, 11: 1–27
- 68 Townsend A M. *Smart Cities: Big Data, Civic Hackers, and the Quest for A New Utopia*. New York: WW Norton & Company, 2013
- 69 Langley P, Zytkow J M. Data-driven approaches to empirical discovery. *Artif Intell*, 1989, 40: 283–312
- 70 Chen W, Zhang S. Experience formula discovery system FDD. *Mini-Micro Syst*, 1999, 6: 410–413
- 71 Dempster A P, Laird N M, Rubin D B. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc B (Methodological)*, 1977, 39: 1–38
- 72 Bach F R, Lanckriet G R G, Jordan M I. Multiple kernel learning, conic duality, and the SMO algorithm. In: *Proceedings of the 21st International Conference on Machine Learning*. New York: ACM, 2004
- 73 Wang J, Zheng N. Multivariate linear correlation analysis. *ArXiv:1401.4827*, 2014
- 74 LeCun Y, Boser B, Denker J S, et al. Backpropagation applied to handwritten zip code recognition. *Neural Comput*, 1989, 1: 541–551
- 75 Deng J, Dong W, Socher R, et al. Imagenet: a large-scale hierarchical image database. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Miami Beach: IEEE, 2009. 248–255
- 76 Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. *Adv Neural Inf Proc Syst*, 2012, 25: 2012
- 77 He K, Zhang X, Ren S, et al. Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: *Proceedings of 2015 IEEE International Conference on Computer Vision (ICCV)*. Chile: IEEE Computer Society, 2015. 1026–1034
- 78 Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks. In: *Advances in Neural Information Processing Systems*. Quebec: Neural Information Processing Systems Foundation Inc., 2015. 91–99
- 79 Sun Y, Wang X, Tang X. Deep learning face representation from predicting 10000 classes. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Guangzhou: Computer Vision Foundation, 2014. 1891–1898
- 80 Wang L, Qiao Y, Tang X. Action recognition with trajectory-pooled deep-convolutional descriptors. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, 2015. 4305–4314
- 81 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *ArXiv:1409.1556*, 2014
- 82 Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, 2015: 3431–3440
- 83 Vapnik V, Izmailov R. Learning using privileged information: similarity control and knowledge transfer. *J Mach Learn Res*, 2015, 16: 2023–2049

- 84 Chris E, Stewart T C, Xuan C, et al. A large-scale model of the functioning brain. *Science*, 2012, 338: 1202–1205
- 85 Lake B M, Salakhutdinov R B, Tnenbaum J. Human-level concept learning through probabilistic program induction. *Science*, 2015, 350: 1332–1338
- 86 Kohonen T. *Self-Organizing Maps*. New York: Springer, 1995
- 87 Kohonen T, Somervuo P. Self-organizing maps of symbol strings. *Neurocomputing*, 1998, 21: 19–30
- 88 Patel V L, Shortliffe E H, Stefanelli M, et al. The coming of age of artificial intelligence in medicine. *Artif Intell Med*, 2009, 46: 5–17
- 89 Gerola H. IBM-Watson. *Bull Amer Astron Soc*, 1980, 12: 131
- 90 Shader R I. Some reflections on IBM watson and on women's health. *Clin Ther*, 2016, 38: 1–2
- 91 Tian Y, Zhang S Y, Luo B, et al. Research progress in impact of heat wave on human health. *Adv Meteorol Sci Tech*, 2013, 2: 49–54 [田颖, 张书余, 罗斌, 等. 热浪对人体健康影响的研究进展. *气象科技进展*, 2013, 2: 49–54]
- 92 Garabedian V D, Pawlotsky J M, Authier F J, et al. Myasthenia gravis and hepatitis C virus infection. *J Viral Hepatitis*, 1996, 3: 329–332
- 93 Silver D, Huang A, Maddison C J, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 2016, 529: 484–489
- 94 Ghahramani Z. Probabilistic machine learning and artificial intelligence. *Nature*, 2015, 521: 452–459
- 95 Ferrenberg A M, Swendsen R H. Optimized monte carlo data analysis. *Phys Rev Lett*, 1989, 63: 1195–1198
- 96 Baheti R, Gill H. Cyber-physical systems. *Impact Control Tech*, 2011, 13: 1–6
- 97 Liu Z, Yang D S, Wen D, et al. Cyber-physical-social systems for command and control. *IEEE Intell Syst*, 2011, 26: 92–96
- 98 Murase H. Artificial intelligence in agriculture. *Comput Electron Agric*, 2000, 29: 1–2
- 99 Barrett J R, Jones D D. Knowledge engineering in agriculture. *Knowl Eng Agric*, 1989, 8: 214
- 100 Aitkenhead M J, Dalgetty I A, Mullins C E, et al. Weed and crop discrimination using image analysis and artificial intelligence methods. *Comput Electron Agric*, 2003, 39: 157–171
- 101 Rykiel E J. Artificial intelligence and expert systems in ecology and natural resource management. *Ecol Model*, 1989, 46: 3–8

Fragmentation knowledge processing and networked artificial intelligence

Jianji WANG[†], Yongqiang MA[†], Shitao CHEN, Ziyi LIU & Nanning ZHENG*

Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an 710049, China

* Corresponding author. E-mail: nnzheng@mail.xjtu.edu.cn

[†] Equal contribution

Abstract The Internet is responsible for the explosive growth of information and its accelerated and diversified dissemination. Human society has entered an unprecedented era of fragmented knowledge. Fragmented knowledge on the Internet is the product of diverse human behavior, thinking, and interactions. Traditional artificial intelligence methods try to intellectualize machines through an expert system established by knowledge engineering, which cannot effectively process and utilize fragmented knowledge. The summarization, reorganization, and re-discovery of the huge amount of fragmented knowledge are thus crucial scientific challenges in information science and artificial intelligence research. We analyze the limitations of traditional artificial intelligence methods associated with the large-scale processing of fragmented knowledge, discuss the basics of processing, organizing, and obtaining fragmented knowledge and the concept of networking artificial intelligence, introduce a law discovery method that can reveal the evolution of big data, and explore the hot topics of future research.

Keywords fragmented knowledge, networking artificial intelligence, knowledge discovery, machine reasoning, collective intelligence network, law discovery

Jianji WANG received his Ph.D. degree from Xi'an Jiaotong University, China, in 2013. He has been a lecturer at the Institute of Artificial Intelligence and Robotics at Xi'an Jiaotong University since 2013. His research interests include law discovery in big data, multivariate correlation analysis, and machine learning.

Yongqiang MA received his master's degree in software engineering from Xi'an Jiaotong University, China, 2015. He is a Ph.D. candidate in automation at the Institute of Artificial Intelligence and Robotics. His primary research interests include neuromorphic computing, neural network, and data mining.

Shitao CHEN received his Bachelor's degree in automation at Xi'an Jiaotong University, China, 2015. He is currently a Ph.D. candidate at the Institute of Artificial Intelligence and Robotics of the Xi'an Jiaotong University, China. His primary research interests include big data, data mining and machine learning.

Ziyi LIU received his Bachelor's degree in automation from Xi'an Jiaotong University. He is a Ph.D. candidate at the Institute of Artificial Intelligence and Robotics. His primary research interests include deep learning, neural networks, and pattern recognition.

Nanning ZHENG is a member of the Chinese Academy of Engineering, IEEE fellow, and professor at Xi'an Jiaotong University. He received his Ph.D. degree from Keio University in Japan in 1985. His research interests include artificial intelligence, pattern recognition, computer vision, and machine learning.