KS-Studio: 一个知识计算引擎

庄越挺，汤斯亮，吴 飞

浙江大学

背景

人工智能正再次成为国际国内学术界和产业界关注的热点，深度学习、迁移学习和增强学习等方法在诸多领域得到了成功应用。图灵早期对获得机器智能进行了一些设想[1]，即通过添加遗传物质，辅以变异、进化、教育与自然选择等手段来使得“the child machine”成熟，并进一步去模仿成人的思维，这一设想至今仍具借鉴意义。与孩童成长类似，这个“child machine”首先需要对语言、文字、图像等非结构化数据所蕴含的（常识性）概念进行理解，来感知外界环境，这是一个跨媒体（cross-media）的感知过程。在此基础上，人类会利用已有的概念、关系等知识，通过演绎和推理等方法来获取新知识。在这一过程中，知识图谱可发挥重要作用。可以说，从数据到知识，再从知识到决策服务是实现人工智能深度应用的主要途径。

知识图谱构建

知识图谱由实体、实体的属性描述以及实体和实体之间的关联构成。尽管其对于大数据人工智能的实现意义非凡，但其构造过程却极为困难。在早期，知识图谱构建单纯依赖于人类专家。在这一方法中，知识图谱中的实体、实体属性与实体关联关系完全由专家人工构造，此类知识图谱包括 WordNet[2]、CyC[3] 等。WordNet 定义了词汇之间的特定语义关系，包含约 15 万个词汇、20 万个词汇语义对；CyC 包含了 320 万条人工定义的断言，涉及 30 万个概念、1.5 万个谓词。随着互联网兴起，虽然依靠专家进行知识图谱构建能获得精度较高的知识，但其规模、构建的速度，以及构建成本已经完全无法适应大数据时代发掘大量涌现知识的需求。为此基于数据驱

动的自动知识图谱构建方法，逐渐成为国际知识图谱研究的主要方向。

目前，国际上主流的知识图谱构建方法根据其知识来源与顶层概念设计理念可大致分为以下四大类。

1. 基于 Wikipedia infoboxes 等结构化数据的构建方法

这一方法以百科作为知识的主要来源，抽取百科词条作为实体，利用词条中的 infobox 来填充实体的属性，其主要代表如 YAGO[4-6]、DBpedia[7-8] 和 Freebase[9] 等。此类构建方法的特点是质量较高，但更新较慢。

2. 基于开放文档构建 (schemaless)

这一方法以互联网开放网页文档作为知识的主要来源，其基本假定为，如果已知两个实体存在特定的语义关系，那么包含实体对的句子在某种程度上就存在表征二者语义关系的作用。于是可利用自然语言处理技术，从非结构化的文本中抽取名词短语作为实体、动词短语作为谓词，通过共现关联与句法分析发现实体之间的关系。其主要代表系统如 Reverb[10]、OLLIE[11] 和 Prismatic[12]。此类方法可以汇聚大量实体与实体间关系谓词，其主要缺点是发现的知识噪音很大。

3. 基于 fixed ontology/schema 的构建方法

这一方法以少量人工定义的抽象 ontology/schema 作为知识图谱的顶层概念设计，以此来充实、汇聚符合顶层概念的实体与实体关系，并在此之上进一步发现新的概念，其代表系统如 NELL[13]、PROSPERA[14] 和 DeepDive[15] 等。此类方法可用于构建面向特定领域的知识图谱。

4. 基于层次化本体 (ontology) 的构建方法

这一方法综合使用上述几种方法来构建知识图谱，尽管可以得到大量的实体、属性、实体关系，但其涉及的顶层概念数量往往较少，而且不能反映概念间的层次特性，为此，另一些研究试图从开放领域寻找构建具有层次化特性的顶层概念的可能性，其主要代表为 Probase[16]。Probase 从开放域汇聚了约 265 万个概念，并计算这些概念的上下位关系，最后基于概率的方法，从横向与纵向对这些概念进行合并，形成一个具有丰富层次的概念树。

上述这几种知识图谱的构建方法均基于文本，目前针对跨媒体数据的自动知识网络构建方法鲜有研究。总体而言，随着现代人工智能技术的发展，基于非结构化开放文档的自动知识图谱构建将是未来发展的主要趋势。

KS-Studio 知识计算引擎

2012 年，中国工程院启动建设“中国工程科技知识中心 (CKCEST)”项目[1]。该项目是我国工程科技领域重要的大数据项目，旨在打通和汇聚各类工程科技数据资源，通过技术分析处理形成知识库，并开发各种应用提供知识服务，推动国家工程科技战略思想库的建设，服务于国家的战略决策。

从数据的性质看，建设知识中心所需的

知识是高度结构化的，而分散在各工程科技领域的数据资源绝大部分属于非结构化数据。如何将无序繁杂的文本、图像、视频等原始的非结构化数据加工转化为有序、可用、标准的结构化知识，是知识中心建设的核心问题。这个问题的解决，需要数据汇聚、知识加工、图谱构建、数据可视化等诸多关键技术的支撑。

2015 年 4 月，作为中国工程科技知识中心的关键技术研发中心，浙江大学提出了 KS-Studio[2]（Knowledge Service-Studio）知识计算引擎的研发计划，旨在综合上述关键技术，探索有效解决这一问题的途径。通过技术手段，让计算机高效地完成从非结构化数据到知识这一过程。

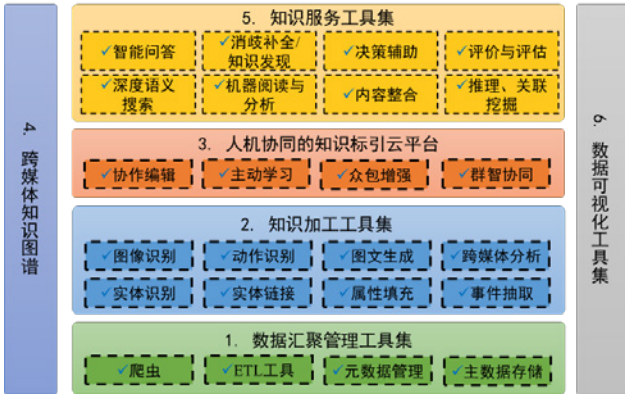

图 1 KS-Studio 功能架构示意图

如图 1 所示，KS-Studio 作为一种知识计算引擎，是将非结构化数据转换为结构化知识及提供创新服务的一系列 API 和工具的集合。KS-Studio 将涵盖从大数据到知识全过程中的核心功能，在知识深度计算基础上提供知识创新服务。目前 KS-Studio 支持从非结构化数据中的汇聚管理，以及从中识别概念、发现新实例与新关系，以构建完善知识图谱，其中用于知识图谱构建的核心包括实体检测、实体链接、属性填充、事件抽取、图像识别、图像文本描述生成，以及跨媒体分析等一系列 API 与工具。在这里，我们把人类社会中所描述的具体对象或概念称为实体。KS-Studio 在对人类社会常识性实体的内涵和外延分析理解基础上，构建知识图谱，从而具备知识深度计算能力，以提供知识创新服务。

具体而言，KS-Studio 的核心 API 分为自然语言 API、视觉 API、跨媒体 API 三部分。

1. 自然语言 API

自然语言 API 帮助用户对文档进行分析、对知识进行加工，更加方便地理解到文档中的所涉及的实体（专有名词）、实体类别（如人名、地名、机构名、疾病名称），以及关系定位等。对于一份非结构文档，自然语言 API 可以通过自动标引的方式将其转化为结构化知识。自动标引服务主要由以下三个功能级联的 API 构成。

- 提及检测 API：系统可以自动检测与识别出文本中与知识库所关心实体相关的专有名词短语，并给出该短语在上下文中的类别信息。比如说，系统可以从一份医学文档中检测并识别出相关的疾病名和药物名等。

- 实体链接 API：通过将一段文档中识别到的实体提及链接到知识库中所对应的条目，来消除歧义，以及发现知识库中未涉及的新实体，并从文档中挖掘出实体新的描述。比如说，我们的算法从文档中检

测得到"猴免疫缺陷病毒"这一关键词，并将其链接到 MeSH[17] 知识库中，便可在 MeSH 知识库中对"猴免疫缺陷病毒"这一概念实体增加下述新的描述，如"猴免疫缺陷病毒"是一种逆转录病毒，以及易感染 45 种非洲非人灵长类动物等。这样就可实现知识图谱的不断学习和扩充。

• 关系发现 API：结合“自然语言理解”和“监督式深度学习”特征提取技术，可挖掘出两个实体之间的关系。如识别出“苯巴比妥造成运动障碍”这一由药物引起病因的关系。

此外，系统还提供了人机协同知识加工与服务，可以将数据驱动的自动标引方法与专家众包机制有效结合起来，让专家对算法发现的知识进行补充、纠错，提升机器学习的效能。

自然语言 API 可以很好地扩展已有知识库，打通不同知识库之间联系。目前 KS-Studio 的实体链接 API，已可将检测出来的实体链接到 MeSH[17]（医学主题词表）、ChEBI[18]（生物化学实体本体）和 Wikipedia（维基百科）三个知识库中，以丰富对所检测实体的深度理解。

2. 视觉 API

视觉对象识别是图像语义理解的基础。KS-Studio 的视觉 API 目前支持对图像语义内容的概念识别，如输入一张图片，自动识别图像中出现的主要实体对象，给出相应的文本标签与确信度。

KS-Studio 目前的视觉 API 可完成部分工程科技领域图像内容的识别，帮助用户进行更有效的资源管理与基于内容的图像分类。除此之外还可用于某些特定的应用领域，例如在海关等检验检疫部门，可对现场拍摄的动植物照片进行分析，快速准确地判断出生物的种类并获取相关知识，帮助相关部门合理处置这类生物，从而避免外来生物入侵事件的发生。

3. 跨媒体 API

KS-Studio 整合跨媒体计算的技术，目前已提供了图文描述生成 API 服务，即在给定一图像后，算法自动生成对该图像的文本描述。跨媒体处理工具能够识别给定图像中的物体及其相互关联，其输出为能够描述图像的一些语句，进而实现从视觉图像到自然语言的跨媒体无缝转换。

在 2016 年由美国国家标准技术研究所（NIST）主办的国际知识库构建大赛[3]（TAC Knowledge Base Population）中，浙江大学的 KS-Studio 从来自 CMU、UIUC、IBM 等国内外知名高校与研究机构的 15 支参赛队伍中脱颖而出，获得了英文实体识别与链接比赛综合排名第一（8 个指标，6 个第一，2 个第二）的成绩。

目前 KS-Studio 已经实现了基于通用的 Wikipedia 与 Freebase 的知识自动标引工具，支持通用知识图谱 6 大类实体的识别，以及 30 多类属性自动填充，可对新闻文本进行 8 类事件的检测发现与关联识别。针对医学领域，实现了针对医学主题词表 MESH 与生物化学知识库 ChEBI 的实体链接与知识标引工具，支持疾病名称、症状、药物分子式等多种类型的实体识别与链接，以及药物导致的副作用、疾病与症状这两

类关系的自动发现。

结束语

下一代人工智能（AI 2.0）将改变计算本身，将大数据转变为知识以支持人类社会作出更好决策[19]。目前 KS-Studio 正在以知识图谱的自动构建为基础，不断丰富对于非结构化数据的知识加工处理的能力，并在不断探索将数据驱动方法与人类常识先验与隐式直觉有效结合起来的可能，我们认为只有如此才能实现可解释、更鲁棒和更通用的人工智能。

参考文献

[1] TURING A M. Computing machinery and intelligence [J]. Mind, 1950, 59(236): 433-60.

[2] FELLBAUM C. WordNet [M]. Wiley Online Library, 1998.

[3] LENAT D B, GUHA R V. Building large knowledge-based systems; representation and inference in the Cyc Project [M]. Addison-Wesley Longman Publishing Co., Inc., 1989.

[4] SUCHANEK F M, KASNECI G, WEIKUM G. Yago: a core of semantic knowledge; proceedings of the Proceedings of the 16th international conference on World Wide Web, F, 2007 [C]. ACM.

[5] HOFFART J, SUCHANEK F M, BERBERICH K, et al. YAGO2: exploring and querying world knowledge in time, space, context, and many languages; proceedings of the Proceedings of the 20th international conference companion on World wide web, F, 2011 [C]. ACM.

[6] HOFFART J, SUCHANEK F M, BERBERICH K, et al. YAGO2: a spatially and temporally enhanced knowledge base from Wikipedia [J]. Artificial Intelligence, 2013, 194(28-61).

[7] BIZER C, LEHMANN J, KOBILAROV G, et al. DBpedia-A crystallization point for the Web of Data [J]. Web Semantics: science, services and agents on the world wide web, 2009, 7(3): 154-65.

[8] LEHMANN J, ISELE R, JAKOB M, et al. DBpedia–A large-scale, multilingual knowledge base extracted from Wikipedia [J]. Semantic Web, 2014,

[9] BOLLACKER K, EVANS C, PARITOSH P, et al. Freebase: a collaboratively created graph database for structuring human knowledge; proceedings of the Proceedings of the 2008 ACM SIGMOD international conference on Management of data, F, 2008 [C]. ACM.

[10] CHEN Y, WANG D Z. Web-scale knowledge inference using markov logic networks; proceedings of the ICML workshop on Structured Learning: Inferring Graphs from Structured and Unstructured Inputs, F, 2013 [C].

[11] SCHMITZ M, BART R, SODERLAND S, et al. Open language learning for information extraction; proceedings of the Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, F, 2012 [C]. Association for Computational Linguistics.

[12] FAN J, KALYANPUR A, GONDEK D, et al. Automatic knowledge extraction from documents [J]. IBM Journal of Research and Development, 2012, 56(3.4): 5: 1-5: 10.

[13] MITCHELL T, FREDKIN E. Never Ending Language Learning; proceedings of the Big Data (Big Data), 2014 IEEE International Conference on, F, 2014 [C]. IEEE.

[14] NAKASHOLE N, THEOBALD M, WEIKUM G. Scalable knowledge harvesting with high precision and high recall; proceedings of the Proceedings of the fourth ACM international conference on Web search and data mining, F, 2011 [C]. ACM.

[15] NIU F, ZHANG C, R C, et al. DeepDive: Web-scale Knowledge-base Construction using Statistical Learning and Inference; proceedings of the VLDS, F, 2012 [C].

[16] WU W, LI H, WANG H, et al. Probase: A probabilistic taxonomy for text understanding; proceedings of the Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, F, 2012 [C]. ACM.

[17] LIPSCOMB C E. Medical subject headings (MeSH) [J]. Bulletin of the Medical Library Association, 2000, 88(3): 265.

[18] DEGTYARENKO K, DE MATOS P, ENNIS M, et al. ChEBI: a database and ontology for chemical entities of biological interest [J]. Nucleic acids research, 2008, 36(suppl 1): D344-D50.

[19] ZHUANG Y-T, WU F, CHEN C, et al. Challenges and opportunities: from big data to knowledge in AI 2.0 [J]. Front Inform Technol Electron Eng, 2017, 18(1): 3-14.

1. http://www.ckcest.cn
2. http://www.ksstuido.org
3. https://tac.nist.gov/2016/KBP/

庄越挺

浙江大学计算机学院院长，教授，博士生导师。国家杰出青年科学基金获得者、教育部长江学者特聘教授，国家级"新世纪百千万人才"，“973”项目首席科学家，中国人工智能学会常务理事。曾获国家科技进步二等奖。主要研究方向为人工智能、跨媒体计算、多媒体分析处理等。

汤斯亮

浙江大学计算机学院副教授，博士生导师。浙江省钱江人才计划特殊急需人才。主要研究方向为跨媒体内容整合、信息抽取、自然语言处理等。

吴飞

博士，浙江大学人工智能研究所所长，教授，博士生导师。国家杰出青年科学基金获得者（2016年）。主要研究方向为人工智能、跨媒体计算、多媒体分析与检索和统计学习理论。