基础教育知识图谱赋能智慧教育

□文／许斌、苏伟杰、刘阳

许　斌

清华大家计算机系副研究员，博士生导师，中国计算机学会（CCF）计算机应用专委会副主任，中国中文信息学会语言与知识计算专委会副秘书长。主要从事知识图谱研究，主持构建了中国第一个全学科基础教育知识图谱 edukg.org，发表了近百篇论文，引用逾千次，H 指数 18。主持科技部 863 计划课题“面向基础教育的海量知识库建设与构建关键技术及系统”等多项国家项目。担任国际万维网大会分会主席，中国云计算与大数据应用学术会议程序委员会主席，中国青年科学家论坛执行主席。曾获中国人工智能学会吴文俊人工智能科学技术进步一等奖、北京市科学技术进步一等奖。

苏伟杰

清华大学计算机系硕士研究生，主要研究方向为知识图谱中的知识链接。

刘　阳

清华大学计算机系硕士研究生，主要研究方向为知识图谱问答。

人工智能技术在基础教育领域的应用要求计算机必须具备基础教育领域的认知能力和理解能力。因此，如何在计算机中表示基础教育的知识成为了一个挑战。知识图谱是用于表示互联网知识的技术。它通过实体和关系来描述客观世界中的概念及其相互关系。清华大学的研究人员将知识图谱技术用于表示基础教育领域的知识，构造出一个包含 1000 多个类、160 万个实例、4000 多种属性、2200 万条三元组的基础教育知识图谱 edukg.org，并将其应用到智慧教育中。通过知识问答、知识搜索、知识梳理、知识链接等多种应用，基础教育知识图谱将不断为智慧教育提供知识动能。

英国人 Tim Berners Lee 于 1990 年发明的万维网是 20 世纪最伟大的发明之一，通过浏览器进行信息浏览极大地丰富了人类信息传播与产生的方式，Tim 也因此获得了 2016 年计算机领域的世界最高奖——图灵奖。1998 年，Tim 提出了“语义网”（Semantic Web）的概念，其核心思想是在网页数据中添加能够被计算机理解的语义信息，从而提升机器的理解能力。源自语义网的知识图谱是一个巨大的知识网络。网络中的节点表示实体，而网络中的边表示实体和实体之间的关系。实体包含概念和实例两种，每个实体还有很多属性－值对来描述实体的内在特性。三元组是知识图谱中描述知识的组织单元。以语文的部分图谱（见图 1）为例，三元组（《蜀道难》，作者，李白）描述了实体“李白”和“《蜀道难》”之间是一种“作者”关系，而实体“李白”和“李商隐”之间通过“三李”这个实体连接起来；另外一个三元组（李白，类型，诗人）描述了实体“李白”在基础教育知识图谱中的类型是“诗人”。通过三元组，可以将基础教育中的许多事实类知识组织成为一个巨大的知识网络，从中发现知识之间的联系。虽然知识图谱并不是计算机表示人类知识的唯一方式，但它可以将大量的事实类知识准确地表达出来，为实现计算机的认知智能提供坚实的基础。

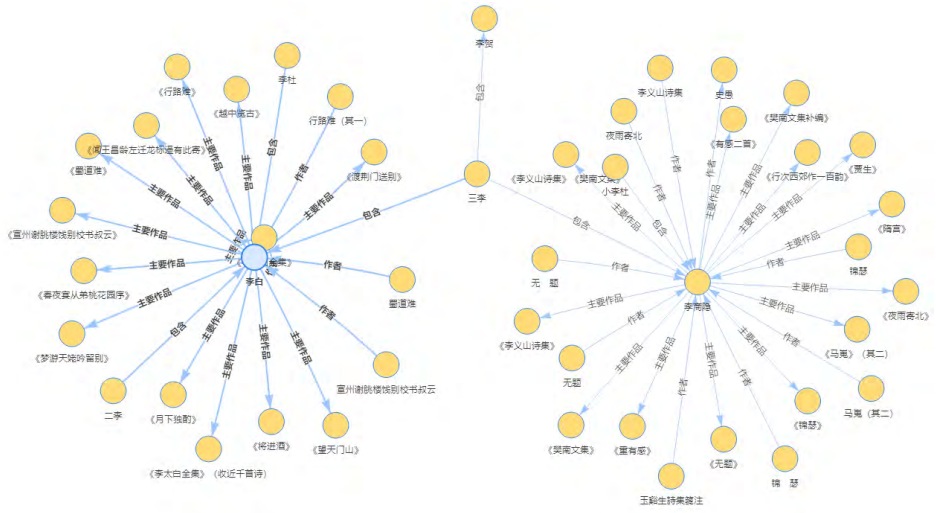

图 1：语文的部分知识图谱

一、基础教育知识图谱的构建

2014 年，国家高技术发展研究计划（863 计划）启动了对类人智能项目“高考机器人”的研究支持。该项目研究人工智能技术在基础教育的应用，要求计算机能够对语文、数学、地理、历史四门学科的高考试卷进行答题。基础教育知识图谱的构建目标就是为了支撑计算机的高考答题，也就是要将中国学生在基础教育阶段学习到的知识“教”给

计算机，让计算机能够“拥有”基础教育知识。其本质就是在计算机中对基础教育知识进行知识表示。在大多数情况下，人类并不能把自己头脑中所有知识的组织方式表达清楚，那又该如何构建计算机能够理解的知识库呢?

知识图谱一般分成通用知识图谱和领域知识图谱，基础教育知识图谱属于后者。通用的知识图谱较多，包括研究领域中的 DBpedia、YAGO、WordNet、Freebase 等，工程领域中谷歌的 Knowledge Graph、微软的 Probase、百度的“知心”、搜狗的“知立方”等。领域知识图谱也有很多被构建出来，比如地理信息领域的知识图谱 Geonames、“天眼查”的企业领域知识图谱等。在 2014 年“高考机器人”项目开始研究的时候，世界上还没有一个中文的全学科基础教育知识图谱。因此，开展基础教育知识图谱的构建工作显得十分必要。

知识图谱的构建方式包括全自动、半自动、全手工。完全自动化的构建方式由于当前的自然语言处理方法还不够好，准确率不能令人满意。而完全人工构建的方法虽然保证了准确性，但是却需要花费巨大的人力和时间成本。因此，如何协调准确率和效率，以便高效准确地构建出知识图谱，是需要解决的一大难题。

基础教育知识图谱的构建包括准确性、全学科、全覆盖、可用性四个方面的挑战。准确性是要求图谱中的基础教育知识必须准确，其知识来源必须是教材等权威资源；全学科是要求图谱必须覆盖中国 K12 教育的主要九门学科（语文、数学、英语、历史、地理、政治、生物、物理、化学）；全覆盖是要求每门学科的知识必须覆盖教育部颁布的学科课程标准中规定的全部知识点；可用性是指图谱中的知识检索与访问效率要足够高。

为了应对上述四项挑战，我们提出了“基础教育知识图谱构建的技术路线”（见图 2）。该路线结合了手动构建的高准确率以及自动构建的高覆盖率和易维护性，同时避免了手动构建方法更新困难和自动构建精度不够的缺点。构建流程包括以下主要步骤：

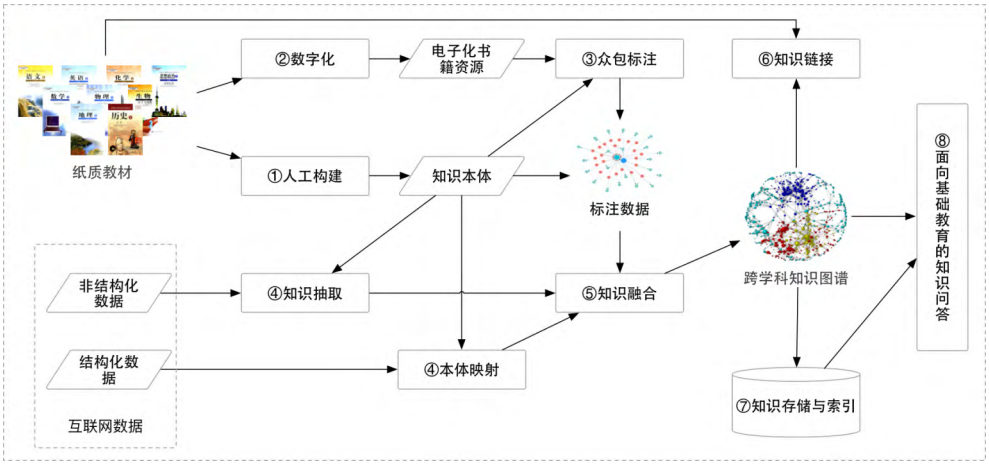

图 2：基础教育知识图谱构建的技术路线

1. 参考教育部颁布的课程标准和教材，知识工程师利用领域本体构建工具手工建立能够覆盖核心语义信息的核心本体概念模型，即基础教育知识本体 Schema；

2. 将国内主流的基础教学材料通过 OCR 技术和相关的文本处理技术进行处理，得到半结构化的电子教材资源；

3. 依据第 1 步中得到的基础教育核心本体概念模型，组织标注人员在电子教材资源上人工标注学科知识三元组；

4. 对互联网上获取的基础教育相关数据资源采用知识抽取和本体映射算法，自动获取基础教育相关结构化知识；

5. 采用知识融合方法，将人工标注和自动获取的结构化知识融合成跨学科的百万实例量级的基础教育知识图谱；

6. 利用算法对非教材电子书籍资源进行知识链接处理，实现课外书籍资源与基础教育知识库的关联；

7. 基于 RDF 和图数据库构建百万实例量级的知识库存储和索引系统；

8. 在基础教育知识库的基础上研究和构建基础教育知识问答系统。

经过上述八个步骤，我们建成了基于教材的涵盖基础教育全学科的知识图谱，以基础教育课程大纲为依据，以基础教育概念模型为骨架，以国家中小学教材为核心资源，包含了 1000 多个概念类、160 多万个实例、2200 万条三元组，并通过知识问答系统验证知识图谱的正确性和覆盖率。

二、基础教育知识图谱在智慧教育中的应用

基础教育知识图谱为计算机提供了实现该领域中认知智能的可能性，使计算机成为拥有基础教育知识的虚拟教师或者学习伙伴。该图谱可以应用到知识搜索、知识快照、知识问答、知识链接等方面。

知识搜索

知识搜索应用是一个垂直领域（基础教育）的搜索，提供了类似普通搜索引擎的搜索界面，用户输入感兴趣的知识点，返回图谱中的知识信息。以图 3 所示为例，用户对“牛顿”这个知识点感兴趣并进行搜索。在返回结果中，左半部是中国所有基础教育教材和教辅中包含了“牛顿”这个知识点的页面。比如，“牛顿”分别出现在“物理”、“历史”、“数学”这三门课程中。一般来说，在物理和历史教材中介绍牛顿是比较普遍的，但是在数学教材中介绍牛顿并不为人所熟知，很多人并不清楚牛顿与数学的关系。这个疑问在返回结果的右半部可以找到答案。右半部的中间部分列出了以“牛顿”为中

图 3：搜索知识点的返回结果

心的关系图谱，可以看到“牛顿”与“微积分”有关，因为牛顿是微积分的发明人之一。因此，“牛顿”出现在“数学”课程中是合理的。但是由于微积分往往到大学才学习，在中学阶段不知道牛顿与数学的关系也是情有可原的。

相比于在通用搜索引擎中搜索基础教育的内容，基于基础教育知识图谱的知识搜索有两个很大的优点：一是内容高度集中在基础教育领域的教材、教辅、课外读物中，不会出现不受控的内容，有利于青少年的健康成长；二是知识图谱能够呈现出知识点的关系图谱，帮助用户更好地进行知识关联与知识联想。

知识快照

知识快照是基础教育知识图谱中的子图，以图的方式呈现出部分实体与关系，其用途是帮助进行单学科以及跨学科的知识梳理。对于单门学科（如语文），可以将该学科一个学期内所讲授的知识点组织成若干张知识快照，十分有助于教师的教学与学生的学习。比如，输入一个历史知识点“江姐”，就可以在图谱中以实体“江姐”为起点生成一张图，呈现出与“江姐”有关的其他实体和关系；还可以输入“爱因斯坦”与“唐太宗”两个知识点，利用图谱生成一张包含实体“爱因斯坦”与“唐太宗”的图。更有意思的是，可以生成跨学科的知识快照，帮助教师与学生进行知识的融会贯通。以图 4 所示为例，总共有 34 个实体，涉及到语文、历史、地理、化学、政治、物理六

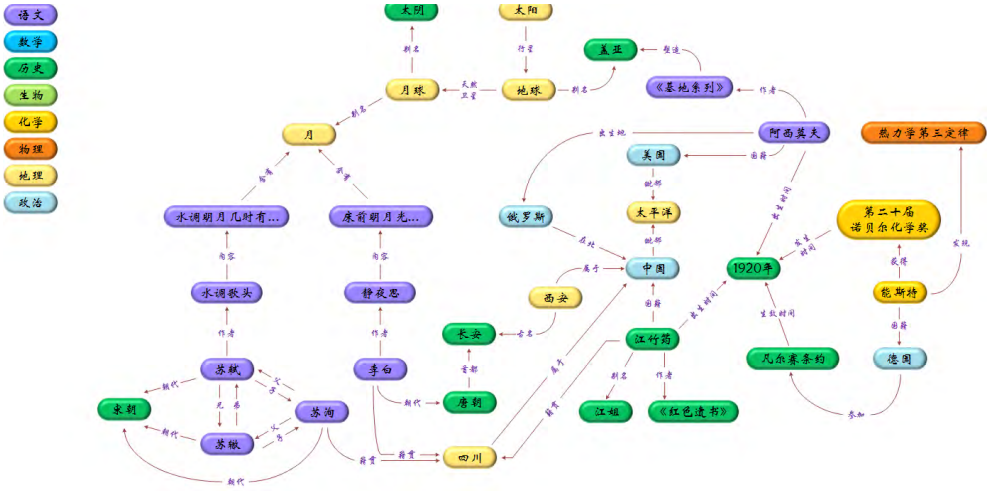

图 4：跨学科的知识快照

个学科的知识点，总共有 46 条边，涵盖了 26 种关系。从图中可以发现很多有意思的联系，比如“苏轼”和“诺贝尔奖”的联系、“盖亚”和“江姐”之间的联系。中学教师讲课都是按照单学科的方式来组织，而知识快照这种实现跨学科知识关联的能力对于教师与学生极有帮助。

知识问答

知识问答是知识图谱的一种智能化应用形式，用户给出自然语言问题，问答系统将其转化为能够对知识图谱进行查询的语句（如 SPARQL），将查询出的知识作为答案反馈给用户。它能够充分利用知识图谱中的结构化数据为用户提供非常简洁、精确的答案。知识图谱问答常用于专家系统、智能客服助理、生活助手等场景中。以图 5 所示为例，可以用自然语言对语文学科的事实类问题进行提问。比如，问“《题西林壁》的作者是谁？”，在问答系统返回答案“苏轼”的同时，还会给出回答这个问题时用到的实体“题西林壁”。这个问答系统是基础教育垂直领域的问答技术，并可通过问答 API 的方式服务于其他“教育机器人”应用。

图 5：用自然语言提问事实类问题的结果

知识链接

知识链接是将基础教育知识图谱中的实体与教材教辅或者参考书

进行链接，也就是当教材等书籍的文字中出现了知识图谱中的实体时，通过算法将其链接到图谱中的具体实体。以图 6 所示为例，课外书《中华民族遗传多样性研究》的书名中提到了生物课程的知识点“遗传”。通过实体链接算法，识别到该知识点应该链接到基础教育知识图谱中的实体“遗传”，于是我们在该书籍电子版中增加了一个可点击的网络链接（见书名中“遗传”的蓝色下划线）。当读者在阅读该书电子版时，点击书名中的“遗传”，就可以跳转到基础教育知识图谱中的生物课程中的实体“遗传”知识卡片，展示出“遗传”是一种“生物现象”，以及它的分类信息、定义和作用等。

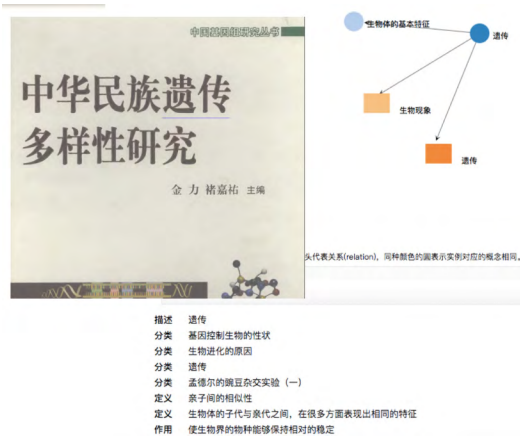

图 6：课外书与基础教育知识图谱间建立知识链接

我们使用知识链接算法对一万本课外读物电子版（由国家图书馆提供）进行链接计算，将这些图书的内容链接到知识图谱的实体。同理，知识链接算法可以对任意一本电子读物进行运算，为电子读物增加知识链接。知识链接的作用在于，当读者阅读这些添加了知识链接的图书时，可以方便地跳转到对应知识点的知识卡片中，增强了基础教育知识图谱的传播性。

三、总结

基础教育知识图谱的构建为计算机在基础教育的认知计算提供了强有力的支撑。这种构建基于中国基础教育教材教辅的知识图谱，以 1000 多个概念类来组织各种知识，抽取出 160 多万个实例，形成了 2200 万条三元组，基本覆盖了教育部课程标准中所规定的九门学科的所有知识点。随着教育部对基础教育教材的更新换代，基础教育知识图谱也将不断进行演化与更新。通过知识搜索、知识快照、知识问答、知识链接等多种应用方式，基础教育知识图谱将不断为智慧教育提供知识动能。

查看内容精选