基于知识图谱的习题资源构建

万林 鲍雯 江苏省徐州高级中学

摘要：在教育数字化背景下，为落实“双减”政策，作者尝试以“教育技术赋能教育”为切口，让教学走向“轻负高质”，并针对“五项管理”教学中突出的难题——作业，借助计算机的优势对习题资源进行管理，构建了基于知识图谱的习题资源系统，同时以高中信息技术学科为例进行论证，期望对其他学科的教学有一定的迁移帮助。

关键词：“双减”政策；知识图谱；习题资源

中图分类号：G434 **文献标识码：**A **论文编号：**1674—2117（2023）21—0084—06

● 问题提出

在教育教学中，习题是检测教师的教与学生的学的效果的最直接的手段，但由于作业量大，学生经常陷入“题海”中不能自拔，疲惫地应付作业，不仅学得很累，而且未能真正掌握知识。在“双减”政策下，教师该如何“轻负高质”地开展课堂教学呢？面对这些问题，笔者提出构建基于知识图谱的习题资源系统，期望能够对教学有一定的帮助。

● 核心概念界定

1.习题与习题资源

习题资源，一方面指的是随着网络的发展、知识的不断更新而出现的海量题目，另一方面指的是与习题相关联的资源（如文本解析、音频讲授、微课详解等）。目前的习题资源不再僵化于简单的题目和答案，已逐渐发展为丰富的富媒体化的习题资源，但这些习题资源分散在各处，处于数据异构化、无序化的状态，急需进行更有效的管理。

2.知识图谱

在习题资源管理领域中，知识图谱可用来描述复杂的关联关系，从语义层面理解习题隐含的信息。因此，笔者运用知识图谱对任教的高中信息技术学科的知识点进行碎片化重组、再造，构建高中信息技术的知识图谱，以便于师生对高中信息技术的知识点进行碎片化学习和对应的习题资源检索和复用。

● 习题资源知识图谱的模型构建

1.节点定义

在基于知识图谱的习题资源管理系统中，主要存在三个节点，分别为学生S、习题Q、知识点K，它们之间形成一个三元关系，其中，学生集本文用$S=\{s_1,s_2,...,s_m\}$表示，习题集用$Q=\{q_1,q_2,...,q_n\}$表示，知识点集用$K=\{k_1,k_2,...,k_o\}$表示。它们之间的关系如下页图1所示。

根据图1不难发现，习题、知识点、学生三者之间形成了稳定的三角关系。从一个节点出发，皆可以关联到其他任意节点，它们三者形成了一种直接关系，而它们与自身之间却需要通过非自身建立联系，故节点与自身之间是一种间接关系。

2.关系定义与权重设置

（1）直接关系

①习题-知识点$M=(Q,K)$。

习题与知识点之间的关系是通过学科专家的直接标记，先人为设置权重，再通过在实验中不断检测，调整权重值，来表示知识点对于该习题的难易程度，分别为0.1，0.3，0.5，0.7，0.9。同时，依据布鲁姆的六个学习目标理论，对知识点的深度进行标注。

②学生–知识点$M=(S,K)$。

学生与知识点之间的关系是学生在学习完知识点后，对与该知识点相关的习题进行作答，根据学生答对与答错该知识点的次数得出学生对知识点的掌握程度：$M(S,K)=\sum \alpha q_1+\beta q_2$，其中$q_1$、$q_2$分别表示学生答对、答错该知识点的次数，$\alpha$、$\beta$为权重系数。

③学生–习题$M=(S,Q)$。

学生与习题之间是通过学生答题的行为进行关联的。根据学生答对与答错该习题的次数得出学生对习题的掌握程度：$M(S,K)=\sum \alpha q_1+\beta q_2$，其中$q_1$、$q_2$分别表示学生答对、答错该知识点的次数，$\alpha$、$\beta$为权重系数。

（2）间接关系

①学生–学生$M(S_1,S_2)$。

学生与学生之间的关系是一种间接关系，他们是通过各自的答题集产生的答题序列，对比他们之间的相似度，得出他们之间的关系。学生S_1、S_2的答题集为Q_1、Q_2，令$Q=Q_1\cup Q_2=\{q_1,q_2,...,q_n\}$，定义学生$U_i$答题序列$W_i=[w_1^1,w_2^1,...,w_n^1,w_1^2,w_2^2,...,w_n^2]$，其中$w_i^1$表示该学生答对$q_i$题的次数，$w_i^2$表示该学生答错$q_i$题的次数。定义学生S之间的相似度：$M_{S_1,S_2}=\frac{W_1+W_2}{\sqrt{W_1^2+W_2^2}}$，设定阈值C，若大于阈值C，则两学生相似。

②习题–习题$M(Q_1,Q_2)$。

习题与习题之间的关系是通过知识点作为桥梁联系的。如图2所示，Q_1与Q_2之间不是直接联系的，

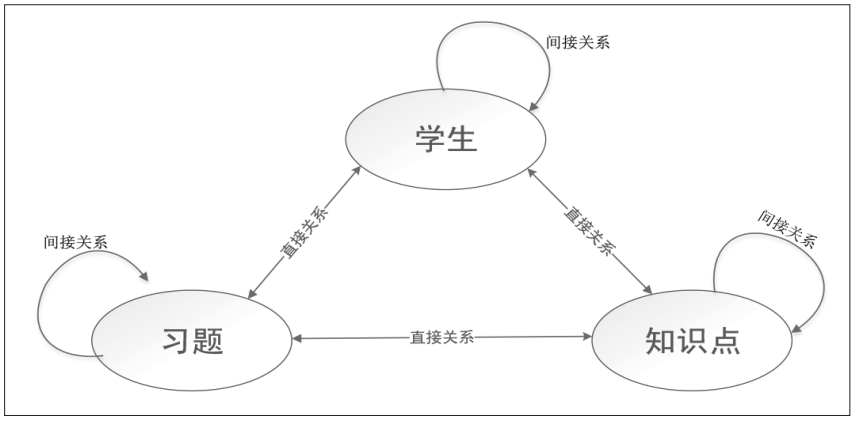

图1 学生、习题与知识点之间的关系

是通过知识点K_1建立了相应的联系。虽然Q_1与Q_2的题面不同，但考查的是同一个知识点，故可以将Q_1与Q_2作为一类习题进行聚类处理。

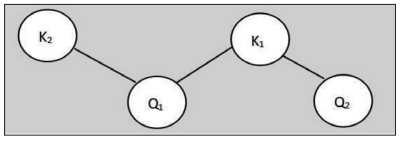

图2 习题–习题之间关系

③知识点–知识点$M(K_1,K_2)$。

知识点是知识体系中相对独立、完整，不能或不宜再进行划分的基本知识单位。知识点层次关系图（Knowledge Network Hierarchy Graph，KNHG）是指由内容上相关联的知识点组成的有向无环图。知识点层次关系反映了学生学习路径或教师讲授知识点的顺序关系，知识点层次关系图如图3所示。

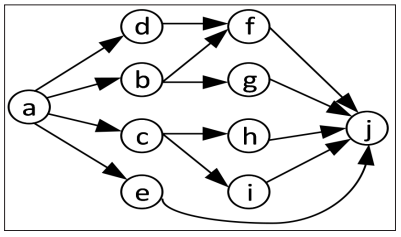

图3 知识点层次关系

知识点a所在的层次是指从起点出发到知识点a的横向最长路径的长度。如图3中b所在的层次是2，h所在的层次是3。在图3中，每一个顶点均代表一个知识点，带方向的箭头表示知识点间具有前驱和后继的关系，即两个知识点的内容或在教学顺序上具有前后关系，如图3中标记为b的知识点的内容依赖于标记为a的知识点，则称b为a的后继，a为b的前驱，其反映的实际意义是指，在学习知识点b之前，学生

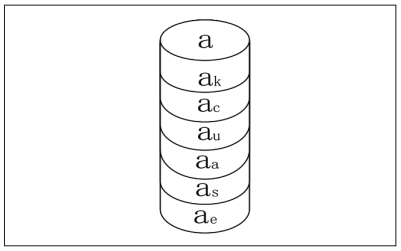

图4 知识点a深度表示

应该已经掌握了知识点a。知识点f的广义前驱知识点集合为{a，b，d}，知识点c的广义后继知识点集合为{h，i，j}。

知识点以及知识点层次关系图虽然可以较好地表示学生的学习路径，但无法表示学生对知识点的掌握程度。针对此问题，为单个知识点划分知识点深度，通过知识点深度表示知识点的掌握

程度。因此，引入布鲁姆教育目标分类分为知道（Knowledge）、领会（Comprehension）、应用（Use）、分析（Analysis）、综合（Synthesis）、评价（Evaluation）六个深度。从当前知识点出发依据布鲁姆教育目标分类到达层次的纵向最长路径称为知识点深度，用KND（Knowledge Network Depth）表示。

在上页图4中，每一圆柱代表一个知识点，知识点分层代表知识点深度。如图中知识点a，将知识点a根据布鲁姆教育目标分类分为六个深度，每个深度分别用a_k、a_c、a_u、a_a、a_s、a_e来表示。例如，“应用”对应的知识点深度为3，表示为$KND_u=3$。知识点深度同样具有单向传递性，即从知识点深度浅的向知识点深度深的单向传递。将知识点深度和知识点层次相联系，知识点之间关系为由前驱知识点要求达

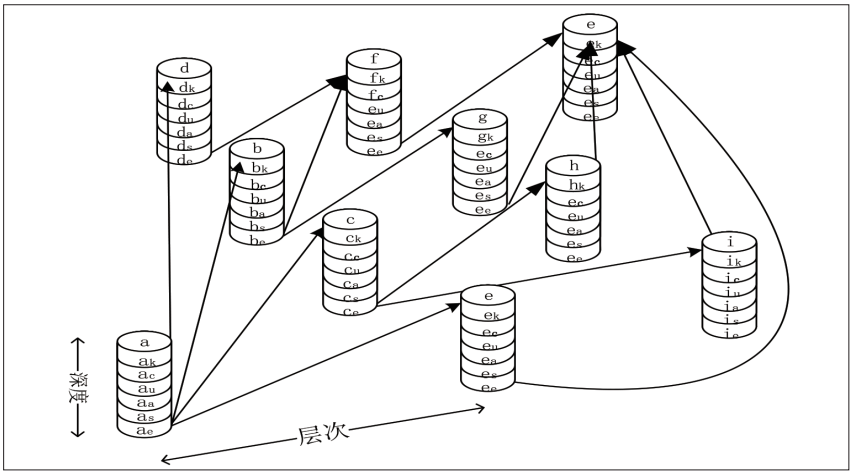

图5 知识点关系

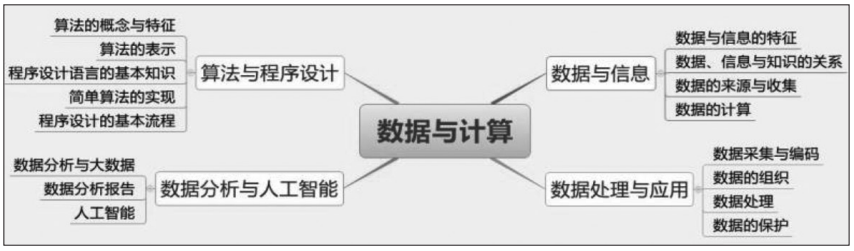

图6 高中信息技术——必修1《数据与计算》知识分类体系

图7 高中信息技术——必修2《信息系统与社会》知识分类体系

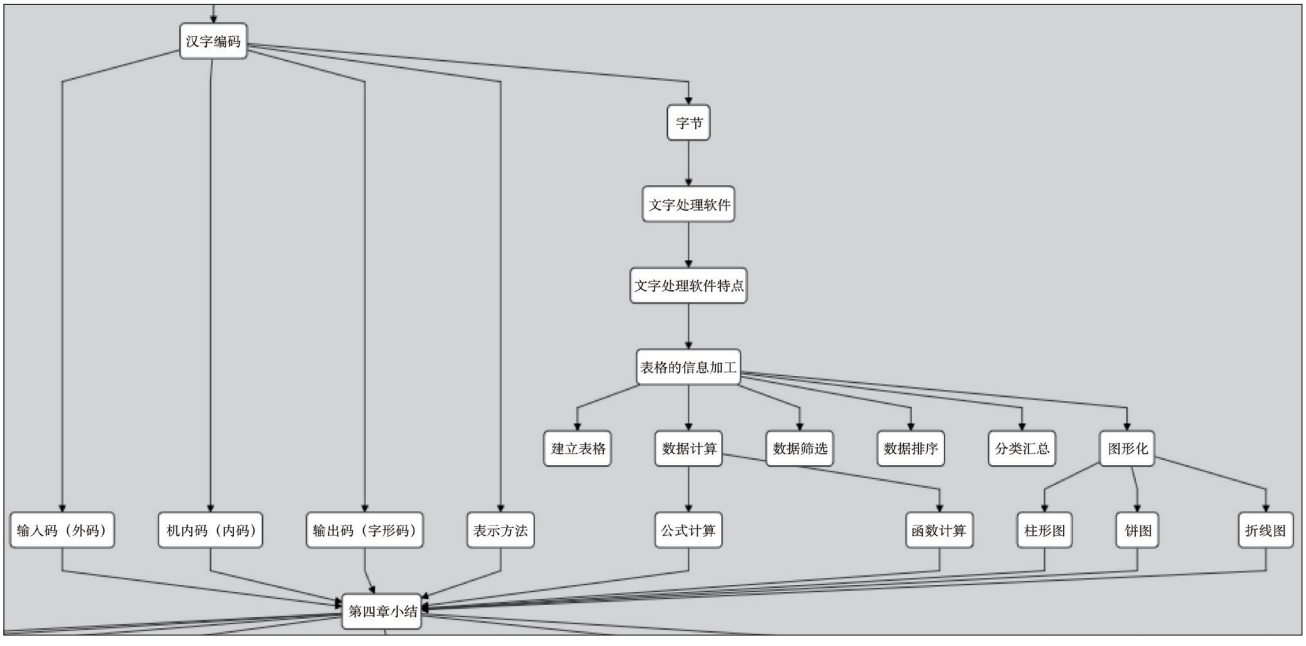

图8 《数据与计算》第二章知识点关系

到的最深深度向后继知识点的最浅深度传递，如上页图5所示。

● 高中信息技术学科知识图谱建构

高中信息技术学科图谱体系的建构包括本体构建和语义标注，高中信息技术学科本体构建主要按照从上到下的方式，对学科知识进行抽象表示、挖掘概念、关系整理、知识分类，规范信息技术学科领域内的知识点、相应关系和结构分类。语义标注是对本体内的概念进行三元组挖掘，处理与数据层的映射关系。

为了解决高中信息技术学科内的知识共享和复用，必须要有高中信息技术学科的知识本体，而目前没有已构建好的高中信息技术本体，因此，笔者通过研究，确定分六个环节来完成高中信息技术学科本体的构建。

1.知识领域范围

学科的知识领域范围，是由其学科的属性、国家课程标准和使用者决定的，《普通高中信息技术课程标准（2017年版）》明确信息技术学科的课程结构类别主要由必修、选择性必修和选修组成，高中信息技术学业水平测试主要考查必修内容，具体为《数据与计算》和《信息系统与社会》两本教材。必修教材上的知识点是构建本体的主要依据，网络上的知识作为扩展补充使用。

2.知识分类体系

为了将高中信息技术的知识点、知识点之间的关系呈现出来，笔者依据权威教材、标注迭代、专家意见三个视角，采用自上向下的方法构建了高中信息技术学科知识分类体系（如上页图6和图7）。

表1　　习题资源语义模型XML元素

元素	属性	说明
exercise_id	var	习题的来源，为根节点，标识习题，采用录入时自动随机产生 URL 进行标识，唯一性
exercise_name	var	习题名称，标识习题的题干
exercise_level	var	标记习题的难易系数
exercise_type	var	标识习题的类型
exercise_option	var	标识习题的选项
exercise_score	number	标识习题的分值
point_id	var	标识习题对应的知识点 ID

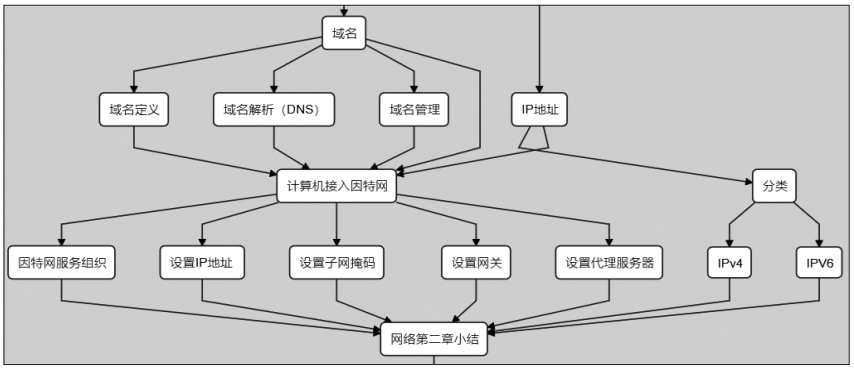

图9 《信息系统与社会》第二章知识点关系

3.知识概念体系

高中信息技术学科的每一章节都包含了很多的知识点，而知识点又可以细分为一个个概念，概念与概念之间有一定的关系，如前驱与后继的关系。只有学会了前面的知识点，才能进行下一个知识点的学习，否则会出现断层现象。

笔者在构建高中信息技术学科知识概念体系时最大限度地包含了高中信息技术学科中全部的核心概念，去掉个别冗余。上页图8、图9是高中信息技术学科的知识概念体系的部分内容，如“字节”“域名”和“IP地址”等，它们与其他概念之间产生关系，共同构成了结构化的高中信息技术学科本体。

4.学科实体的关系

知识概念体系的层次结构体现了知识概念之间的一种继承关系，依据朴素贝叶斯模型训练知识分类体系，识别高中信息技术学科知识概念体系的“实体－关系－实体”三元组。将知识点的概念化为空间中的一个点，概念与概念之间的关系称为边。点与点相连，边与边相接，构成了实体之间的庞大关系网。采用纯人工标注数据作为训练集，得到一个实体关系分类模型，训练高中信息技术本体中知识内容，使得标

注知识更有结构性和层次性。

5.学科本体的审核

针对高中信息技术学科领域纯手工构建的本体缺乏科学评价和管理机制的问题，学科专家应将国家出台的课程标准和省考试院的考试大纲作为审核的依据进行审核。

6.学科本体的再修改

根据学科专家对本体的审核，应适时适当做出再修改，保障高中信息技术学科本体的科学性、严谨性和全面性。

● 基于知识图谱的习题资源语义化处理

1.XML元素语义标注

XML（Extensible Markup Language）是一种以标记为核心的自然描述语言，具有元信息语义和很强的扩展能力，充分描述数据的逻辑性、严谨性、层次性和结构性的特点。笔者准备使用XML语言对高中信息技术习题资源进行语义化处理。

针对习题资源定义XML元素，利用定义完整的XML元素将收集的高中信息技术习题进行逐一标注，标注方法则按照知识图谱的思想进行，当习题经过标注后，计算机能通过解析XML元素获取习题隐藏的信息，具有语义功能。习题资源语义模型XML元素如上页表1所示。

在经过XML元素标记后，习题及其属性已经被有效分离。元素exercise_id的属性包含习题的ID，是标记习题唯一性的前提；exercise_name元素包含了习题名称及其题干，通过标注可以挖掘习题隐含的信息，更有利于计算机对习题的理解；元素exercise_level包含了习题的难易程度，为教师组卷，推荐习题资源做准备；exercise_type元素标记习题的类型，本文主要是针对高中信息技术学科，将习题类型分为选择题、操作题和简答题；exercise_option元素作为标记习题的选项或解题的提示，让学生从迷惑的答案中做出选择；exercise_score元素则表示习题的分值，标记出习题的每道题分值，有利于对学生做出评价和反馈；point_id元素是标记习题所对应的知识点ID，找到了知识点的ID，就是找到了知识点本身，及其资源。

表2 知识点

字段描述	数据库字段	类型	长度	是否为空	备注及扩展
知识点 ID	Id	varchar (32)	50	not null	主键
知识点编号	Knowledge Code	varchar (m)	50	not null	
知识点名称	Knowledge Name	varchar (32)	128	not null	
学段 ID	StageId	varchar (32)	50	not null	
学科 ID	StudyId	varchar (32)	50	not null	
最后同步时间	sysTime	char (m)	20		

表3 试卷表

字段描述	数据库字段	类型	长度	是否为空	备注及扩展
习题 ID	exercise_id	varchar (32)		not null	主键
习题名称	exercise_name	varchar (m)	200		
习题难度	exercise_level	varchar (m)	10		简单、一般、困难
习题类型	exercise_type	varchar (m)	10		选择题、操作题、简答题
习题选项	exercise_option	varchar (m)	200		
习题分数	exercise_score	number (m, n)	2, 0		
知识点 ID	point_id	varchar (32)			

本节不仅通过XML元素语义化习题的过程，挖掘习题隐藏的信息，充分发挥知识图谱具有的逻辑性、层次性等优势，而且还建立了习题、习题资源和知识点数据表，帮助系统更好地构建习题资源数据库。

2.知识点数据表

知识点属性主要包含知识点ID（id）、知识点编号（knowledgeCode）、知识点名称（knowledgeName）、学段ID（StageId）、学科ID（StudyId）与知识点最后同步时间（sysTime），如表2所示。其中，学段ID用于区分学生所处学段，学科ID用于区分学生使用系统时所学科目，一个学生可以对应多个学段与多个学科。

3.习题表

习题是学生在检测时使用的资源，习题表主要用于存储与习题相关的信息，主要包含习题ID、知识点ID、所属人ID、习题难度、习题类型等基本信息，如上页表3所示。

4.资源表

资源主要分为两类：讲解资源和习题资源。讲解资源主要包括微课、视频、课件、教案等对知识点进行讲解的学习资源；习题资源主要用于作业、练习与测试。

由于系统的使用不仅局限于某一具体学段及学科，各学段学科间可能会存在知识点重复的问题，但不同学段的对知识点掌握程度的要求也不同，因此习题资源除了要与知识点进行对应外还需要与学段对应，如表4所示。

表4　　学习资源

字段描述	数据库字段	数据类型	数据长度	是否为空	备注及扩展
资源 Id	resource _id	varchar (32)		not null	主键
资源名称	name	varchar (64)		not null	
资源类型	resource _type	varchar (m)	2		00 微课，01 视频，02 课件，03 教案，04 试卷，05 其他
所属字段	stage	varchar (32)			
所属学科	subject	varchar (32)			
所属版本	vertion	varchar (32)			
所属年级	grade	varchar (32)			
所属章	unit	varchar (32)			
所属节	chapter	varchar (32)			
所属知识点	point	char (m)			
上传时间	createTime	char (m)	20		
资源状态	status	varchar (32)	2		

小结

本文主要针对习题资源在实际的教学中存在的问题，提出利用知识图谱的技术理念，构建知识点、习题、学习者三者之间的三元关系模型，并以高中信息技术学科为例，构建了本学科知识图谱，且对习题资源进行语义化处理，且设计了习题、知识点和资源的数据表，为下一步系统的开发实现做准备。实践说明，构建基于知识图谱的习题资源系统能够提高教学质量，希望对其他学科的教学具有借鉴意义。

参考文献：

[1]百度百科.知识图谱[EB/OL].https：//baike.baidu.com/item/知识图谱/8120012?fr=aladdin.

[2]杨瑛，朱频频.语义技术应用与标准发展[J].信息技术与标准化，2015(04):18—20.

[3]杜方，陈跃国，杜小勇.RDF数据查询处理技术综述[J].软件学报，2013(06):1222—1242.

[4]于根，李晓戈，刘睿，等.基于信息抽取技术的问答系统[J].计算机工程与设计，2017，38(04):1051—1055.

[5]王冬青，殷红岩.基于知识图谱的个性化习题推荐系统设计研究[J].中国教育信息化，2019(17):81—86.

[6]徐雄峰.基于知识图谱的政治知识库构建及其应用研究[D].武汉：中南民族大学，2018.

作者简介：万林，教育硕士，中学一级教师，研究方向为人工智能、信息技术教学；鲍雯，中学高级教师，研究方向为课程教学。

本文系江苏省教育科学“十三五”规划2020年度重点自筹课题“基于语义技术的习题资源管理及应用研究”（B—b/2020/02/188）的研究成果。