面向高校课程体系的多源教育知识图谱构建与融合

李泽举，程琳雅，张春红*，朱欣宁，赵辉 北京邮电大学 lizeju@bupt.edu.cn, chenglinya@bupt.edu.cn,zhangch@bupt.edu.cn, zhuxn@bupt.edu.cn, hzhao@bupt.edu.cn

arXiv:2305.04567v1 [cs.AI] 8 May 2023

摘要——**人工智能（AI）的快速发展使教育领域发生了重大变革。在各种人工智能技术中，利用自然语言处理（NLP）的知识图谱（KG）已成为整合多维信息的强大可视化工具。在大学教育背景下，大量专业课程和复杂的学习资源往往导致学生的学习效果不佳。本文提出了一种面向电子信息专业的自动化框架，用于知识抽取、可视化知识图谱构建和图融合。此外，我们还通过数据分析研究课程间的相关度及关系，对热门知识概念进行排名，并探索课程之间的交叉关联。我们的目标是提高学生的学习效率，并探索人工智能赋能的新型教育模式。通过提供对各类概念及课程之间关系的全面理解，所提出的框架有望帮助学生更好地理解和认识其专业领域的复杂内涵。**

索引词—**知识图谱构建，知识图谱融合，电子信息**

一、引言

由于技术的快速进步以及对创新教学方法 [1]–[3] 日益增长的需求，人工智能（AI）融入教育受到了广泛关注。专家提出，应利用智能技术加速人才培养模式和教学方法的改革，构建包括智能学习和交互式学习在内的新型教育体系。对于大学教育而言，每门课程都包含大量复杂的学习材料，例如教学大纲、教科书和幻灯片，这使学生难以确保学习成果的质量和效率。因此，有必要探索将人工智能融入大学教育的新颖且实用的方法。

知识图谱 [4], [5] 将教材中分散在各处的知识概念联系起来，形成大型知识库，以结构化形式描述客观世界中实体之间的复杂关系。此外，知识图谱的应用场景十分广泛，例如自然语言理解、基于知识图谱的问答系统、推荐系统，以及提供可视化的知识表示。

*张春红为通讯作者

近年来，越来越多的研究人员通过嵌入知识进行表征学习，将知识图谱引入教育领域 [6], [7]。然而，当前的研究成果仍存在一些局限性。首先，构建课程知识图谱所需的人力和物力资源过于庞大，且知识抽取的质量可能不尽如人意。此外，目前为教育领域构建的知识图谱大多是小规模图谱，或针对特定章节、特定课程构建的图谱。因此，我们将自然语言处理（NLP） [8] 和知识图谱 [5] 引入电子信息领域的大学教育中。电子信息专业的学生主要学习电路基础知识，并掌握利用计算机处理信息的方法。我们整合课程中的知识概念，自动构建电子信息领域的可视化知识图谱。学生可以检索知识概念、查看知识节点的属性等，从而更加高效、快速地掌握课程的核心知识。同时，学生还可以掌握相关的知识概念，加深对其他关联课程的理解和认知。

我们的贡献总结如下:

- 我们为数百门电子信息课程构建多源、大规模知识图谱，并将其自动导入图数据库进行可视化，涵盖超过 6 万个实体和 8 万个三元组。
- 我们提出了一种适用于异构学习资源的自动化知识抽取、数据清洗和图融合 [9], [10] 框架，该框架也适用于其他学科。
- 我们在构建的图中开发了辅助信息统计功能，例如课程体系相关性计算、热门知识概念排名、知识节点及其属性的相关关键词索引。

II. 方法

我们将知识图谱的自动构建与图融合分为以下步骤: 分别将教科书和课程幻灯片从原始格式转换为可编辑格式。接下来，我们使用“知识单元、知识章节、知识块、知识点”，自动将教科书和课程幻灯片文档划分为相应部分。

正则化表达式和命名实体识别（NER）[11] 方法。之后，我们将章节和知识点自动导入“Neo4j”图数据库，以构建大规模知识图谱。为了提升图的质量，我们使用由深度学习支持的语言纠错模型 [12]进行数据清洗。最后，我们对构建的图谱中的重要信息进行统计。

A. 学科知识图谱的本体构建

本体构建 [13] 为构建层级性强、冗余度低的统一知识结构提供指导。为了构建能够反映大学课程体系知识层次和知识内容核心概念的本体结构，我们定义了类及类继承。一个类包含多个知识实体，不同的类代表不同的层级或粒度。我们分析教学资源的知识层级划分，并设计知识实体的类划分，以获得本体类及其属性。随后，我们定义实体间的关系类型，从而获得能够以结构化方式表示知识关联的本体关系。

1）学科知识图谱实体类型定义：实体类型是对一组具有相同特征或属性的实体的抽象，用于区分课程内容和知识粒度。由于知识体系依据学科规则按照不同粒度层级进行处理，因此我们以教学资源的层级目录作为构建知识图谱本体的主要依据。具体而言，我们从最粗粒度的知识入手，再根据教学资源的目录层级逐步细化，以确定知识体系的最大粒度（知识单元）和最小粒度（知识点）。

$KG_{textbook}$ 表示教材中的知识图谱， KG_{slide} 表示课堂讲义中的知识图谱， $KG_{syllabus}$ 表示课程大纲中的知识图谱。在 $KG_{textbook}$ 和KG_{slide}中，我们进一步将各学科的知识体系细化为五级结构：课程-知识单元-知识章节-知识块-知识点。

知识点 是粒度较细的概念级知识，而其他则是课程级知识。在 $KG_{syllabus}$中，知识粒度被划分为“课程-教学内容-知识点”。

2）学科知识图谱实体属性定义：实体属性的定义是对实体语义信息的补充，其覆盖范围和准确性会影响知识图谱表示信息的准确性。不同类型的本体属性可以增强类别信息，并便于区分类别。因此，我们为每种知识实体设计了相应的属性集。这些属性全面、具体地呈现了实体信息。

例如，我们可以从表 I 中看到，知识单元设置了四个基本属性（名称、排序值、级别、URL），知识章节、知识块 和教学内容。对于实体知识点，我们设置了七个基本属性（名称、排序值、起始位置、词频、URL、级别、描述）。属性“描述”可分为三类：维基百科英文描述、维基百科中文描述和百度描述。

表 I 实体类型及相应属性

实体类型	实体属性	属性编号
课程	学期，name，背景，URL，课程先修要求，教育对齐	6
知识单元	名称、排序值、层级、URL	4
知识章节	名称、排序值、层级、URL	4
知识块	名称、排序值、层级、URL	4
教学内容	名称、排序值、层级、URL	4
知识点	name，排序值，开始，词频、URL，级别、维基百科英文描述，维基百科中文描述，百度描述	9

3) 学科知识图谱的实体关系定义：实体关系体现了不同实体类型之间的内在联系。对于知识体系而言，探究知识实体之间的关联至关重要，这有助于学生形成对课程知识体系更深入的累积性理解，并为教师开展课程改革提供清晰的知识结构。因此，在我们的课程知识体系中，知识实体彼此相互依存。我们通过教学资源中出现的章节以及每个知识实体的层级结构，建立知识关联。

$KG_{textbook}$、KG_{slide} 和 $KG_{syllabus}$ 都与名为“*haspartOf*”的关系相连，用于表示知识粒度和层级结构。从上到下， $KG_{textbook}$ 和 KG_{slide} 所连接节点的实体类别为 [$Course \xrightarrow{haspartOf} KnowledgeUnit \xrightarrow{haspartOf} KnowledgeChapter \xrightarrow{haspartOf} KnowledgeBlock \xrightarrow{haspartOf} KnowledgePoint$]。

类似地，在 $KG_{syllabus}$ 上，将 [$Course \xrightarrow{haspartOf} TeachingContent \xrightarrow{haspartOf} KnowledgePoint$] 相连，以表示自顶向下的知识系统。

B. 数据清洗

知识概念的质量决定了我们构建的图的质量、图融合过程以及学生的学习效果。因此，数据清洗是保证知识图谱质量不可或缺的一步。构建知识图谱时，可能会出现文本转换错误、格式编码问题及其他问题。此外，实体及其属性中也存在实体异常，例如文本中夹杂不完整的公式和符号、中英文混杂、拼写错误以及实体语法错误。此外，由于命名实体识别模型和用于知识概念抽取的正则化表达式存在算法局限性，提取出的实体可能会出现语法错误。

基于上述内容，我们需要对构建的原始图进行数据清洗。数据清洗是确保实体及其属性准确性和图融合效果的重要步骤，因为未经清洗的数据可能导致错误的决策和不可靠的数据挖掘。我们使用语言纠错模型来检查并识别实体中出现错误词语的位置。对于疑似错误的词语，根据语言模型对纠正结果进行排序，以选出最佳结果，然后自动完成错误纠正。

C. 知识图谱融合

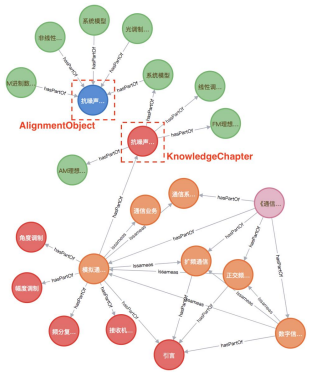


图 1. 这是知识图谱融合的一个示例。高亮显示的节点是名称相同的节点。在图中，粉色节点表示课程，橙色节点表示知识章节，红色节点表示知识章节，绿色节点表示知识块，蓝色节点表示对齐对象层级，即最低层级的知识点。

我们进行知识融合，以减少知识图谱中冗余的知识节点并提高数据质量。同时，为了保护知识概念内容的多样性，指向同一知识概念的实体将被融合。它们的属性和连接关系会予以保留，以维持知识图谱的信息内容。我们的主要融合方法分为四个部分：实体名称匹配、定位粗粒度实体、属性并集，以及保留原始边，如下所示：

a. 实体名称匹配

首先，我们对实体名称进行匹配，以确定两个或多个实体是否指代同一个知识概念。我们使用字符串匹配、字面相似度和语义相似度进行名称匹配及后续整合。如图1所示，知识图谱融合的第一步是将这些指代相同知识的碎片化节点整合为一个唯一节点。

b. 定位粗粒度实体

如上一节所述，知识概念（实体）通常存在于教科书知识体系的不同层级中。层级越高，覆盖范围越大，粒度也越粗。我们将待融合的实体定义为 $E = \{e_1, e_2, ..., e_n\}$，其层级为 $L = \{l_1, l_2, ...l_n\}$。例如， e_f 用于表示融合后的知识节点（实体）， l_f 表示 e_f 的唯一层级。为了为了更完整地保留知识图谱中的层级结构， e_f 位于 E 和 $l_f = highest（L）$ 中最粗粒度的节点处。

c. 对属性取并集

为了在知识融合工作后保持知识图谱的信息量，除了保留层级关系外，我们还需要保留融合前每个知识概念（实体）的属性信息（“url”除外）的并集运算。

我们定义属性集 $A = \{A_1, A_2, ..., A_n\}$，表示待融合实体集 E 对应的所有属性信息。 $A_1 = \{..., a_1, b_1, ...\}$, $A_2 = \{..., a_2, b_2, ...\}$， A_i ={$..., a_i, b_i, ...$}，其中 a_i 表示实体 $e_i \in E$ 的 a 属性， b_i 表示实体 $e_i \in E$ 的 b 属性。 A_f 表示融合后节点 e_f 的属性集，可表示为：

$$A_f = A_1 \cup A_2 ... \cup A_n = \{..., \sum_{i=1}^{n} a_i, \sum_{i=1}^{n} b_i, ...\}$$

例如，两个“抗噪性能”知识概念分别具有图1所示的属性，那么融合工作后得到的唯一知识概念将取两个节点属性的并集，即包含名称、描述、词频和排序值属性，从而最大限度地保留属性信息。

d. 保留原始边

在不同教材中，知识节点（实体）之间的拓扑结构和连接关系各不相同。除了保留实体信息外，还应对关系信息进行相应处理，以实现信息保留。由于融合后，同一概念在教材中只会对应一个节点，因此融合前存在的所有连接关系都应予以保留，并连接到融合后的唯一知识概念。

我们的图融合类型分为不同来源融合和不同课程融合。我们利用上述方法对同一课程进行图融合。此外，我们通过实体属性匹配，保留原始边并创建等价边，以实现课程间的图融合。

III. 实验与分析

A. 两门课程之间的相关度分析

在本部分中，我们通过计算两门课程之间等价关系边的数量占知识概念总数的比例，来判断任意两门课程之间的相关程度。首先，我们在课程体系内部以及同一门课程内进行整合，同时保持原始知识图谱结构系统不变。根据融合结果，我们认为课程之间存在相似关系，因此提出了课程间相关度指标，其定义为课程之间知识概念的数量占该课程节点总数的比例。

我们通过统计两门课程之间建立的等价关系边的数量，来检查课程之间知识概念的覆盖情况。

表 I 通信原理的相关度计算结果

$Course_i$	$Course_j$	$S_u(i,\ j)$	$S_v(i,\ j)$	$S(i,\ j)$
通信原理	信号与系统	0.11389	0.14008	0.12698
通信原理	通信电子电路	0.11676	0.05589	0.08632
通信原理	信息论基础	0.09912	0.05589	0.07751
通信原理	概率论基础与数学	0.06768	0.05078	0.05923
通信原理	线性代数与几何	0.08368	0.01761	0.05065
通信原理	数字电路与逻辑设计	0.2265	0.05723	0.03994
通信原理	电子与电路基础	0.02702	0.04336	0.03519
通信原理	数字信号处理	0.01923	0.04813	0.03368
通信原理	C/C++ 编程	0.04177	0.02548	0.03362
通信原理	电磁场与电磁波	0.03956	0.00998	0.02477
通信原理	大学物理	0.03242	0.00540	0.01891

边。 $KG_{textbook}$ ={$...V_i, V_j, ...$}, KG_{slide} = {$...U_i, U_j, ...$}, V_i 表示教材知识图谱中的课程，该课程的节点总数为 Nv_i； U_i 表示幻灯片知识图谱中的课程，课程节点总数为 Nu_i。课程 V_i 与课程 V_j 之间建立的映射等价边数量表示为$sim\ v_{i,\ j}$，课程 U_i 与课程 U_j 之间建立的映射等价边数量表示为$sim\ u_{i,\ j}$。分别融合两个数据源的知识图谱，并定义整体课程相关度为 S:

–

$$S(i,j) = \frac{Su(i,j) + Sv(i,j)}{2}$$

幻灯片数据库计算出的相关度为相关度 Su，教材数据库计算出的相关度为相关度 Sv。我们结合两个数据源的数据对数据结果进行了综合分析，并将其作为两门课程相关度的平均总相关度。Su和 Sv 的计算如下:

$$Su(i,j) = \frac{sim_u_{i,j}}{Nu_i}, Sv(i,j) = \frac{sim_v_{i,j}}{Nv_i}$$

以通信原理课程为例，计算该课程相关度的结果如表 II 所示。电子信息领域其他课程与该课程最相关的课程及其对应的相关度如表 III 所示。我们可以看到与通信原理高度相关的课程。在实际教学应用场景中，这些高度相关的课程可以辅助通信原理的学习，帮助教师安排合适的教学方案，并使学生能够从宏观上整体把握需要学习的课程间关系。我们提出的方法在比较课程间相似度方面具有很大优势，因为我们不仅考虑了两个课程名称之间的相关性，还考虑了课程各层级知识概念的重叠率，从而确定课程间关系。

表 III 每门课程最相关的课程及其对应的相关度

课程	最相关课程	相关度 tion 度
通信原理	信号与系统	0.127
信号与系统	通信电子电路	0.119
数字电路与逻辑设计	电子与电路基础	0.201
概率论基础理论与数学	线性代数和几何学	0.084
电子与电路基础	通信电子电路	0.345
电磁场和电磁波	大学物理	0.029
线性代数和几何 y	信息基础人人爽理论 y	0.015
数字信号处理	信号与系统	0.064
大学物理	电磁场与电磁波	0.101
C/C++ 编程	线性代数与几何学	0.046
通信电子电路	信号与系统	0.035
信息学基础理论	通信原理	0.149

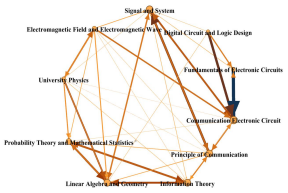
B. 探索多门课程之间的关系

根据计算两门课程相关度的方法，我们可以按照相关度对每门课程与其他课程进行排序，从而分析多门课程之间的关系并对其进行分类。

我们采用谱聚类方法 [14] 并输入关联度矩阵，所得结果与上述人工分类结果高度相似，可用作课程分类的聚类方法。首先，我们选择数据资源相对完整的十个学科，数据支持的课程集合为 C, $C = \{c_1, c_2, ...c_i, ...c_n\}$($n = 10$)。相关度 $S(i,j)$表示 c_j 与 c_i之间的相关程度。因此，对于每门 c_i 课程，都有$S_w(i) = \{S(i,1)\ S(i,2)$ ....., $S(i,j)$ ...}，表示 c_i与其他课程的相关度集合。当需要比较多门课程之间的相关性时，我们提出 $W(i,j)$来表示 c_j 与 c_i之间的关联权重。因此，每门课程 c_i 与其他课程的相关性都具有权重集合 $W(i) = \{W(i,1)\ W(i,2)$ ..., $W(i,j)$ ...}。在

同一课程中，我们对相关度进行归一化，得到该课程的 c_i 与其他课程每个 c_j 的归一化相关度，从而更好地表示同一维度上的相关程度。

为了更直观地呈现多门课程之间的关系，我们使用相关边的粗细来反映相关度的权重。为了便于观察课程间相关度，我们移除了权重小于 0.25 的相关边，如图 2 所示。本部分得出的多门课程之间的关系结果，对教师制定教学计划具有一定的参考价值。



#	关键词热度	
1	傅里叶变换	120
2	频道	102
3	AM	99
4	卷积	90
5	筛选	87
6	高斯白噪声	86
7	Z 变换	80
8	FM	71

图 2. 多门课程相关度权重图　图 3. 我们的知识图谱可视化系统界面中知识概念频率排名。

C. 探索最有价值且最完整的知识概念

由于我们的知识图谱包含大量知识概念，学生难以从中聚焦于最有价值、最重要的知识概念。因此，我们分别统计数据库中所有知识概念的出现次数，并计算其交集。首先，我们统计所有知识概念的出现频率。由此可得图 3 中知识概念的频率排名，从而探索电子信息领域中最重要的知识概念。

随后，我们统计图融合后同一门课程在其 $KG_{textbook}$、KG_{slide} 和 $KG_{syllabus}$ 中的知识概念交集。这样，我们就能找出一门课程中最有价值的知识概念，因为它们同时出现在 $KG_{textbook}$、 KG_{slide} 和 $KG_{syllabus}$中。我们发现，通信原理中的16个知识概念出现在 $KG_{textbook}$、 KG_{slide} 和 $KG_{syllabus}$ 中：“内积、正交性、相关系数、能量功率谱密度、希尔伯特变换、解析信号、带通信号、匹配滤波器、频分复用、时分复用、奈奎斯特准则、升余弦滚降、相位模糊、最优接收、星座图、频谱”。通过这种方式，我们可以找出通信原理中最有价值的知识概念。因此，教师和学生在教学与学习过程中会重点关注这些知识概念。

IV. 讨论与结论

我们提出了一种新颖的知识图谱构建与图融合方法，借助自然语言处理和机器学习工具实现，是一种自动化框架，适用于异构学习资源，也适用于其他学科。通过我们的图融合方法，将单个知识图谱融合为来自多个数据源的知识图谱。同时，通过数据挖掘与分析，我们推断电子信息课程间的关系，并建立学科间关系，以梳理学科知识框架。

我们的方法不仅符合学习者的认知习惯，还能帮助学生理解不同课程之间的知识概念，并全面掌握某一学科的整体知识概念，从而提高学习效率。此外，该方法还能为教师实现相应的教学目标提供有力的数据支持，并突出核心知识概念，进而提高教学质量。

致谢

本研究受到2022年北京市高等教育“本科教学改革与创新项目”以及2022年北京邮电大学教育教学改革项目（2022JXYJ-F01）的资助。

参考文献

[1] A. Siren 和 V. Tzerpos，“利用开放教育资源（OER）自动创建学习路径：系统性文献映射”，*IEEE*学习技术汇刊，第15卷，第4期，第493–507页，2022年。[2] A. Kulkarni、S. L. Chu、N. Sharma 和 S. Sathe，“通过情境化聊天机器人，在视频在线学习平台上开展基于兴趣的学习”，载于*2022*年国际高级学习技术会议（*ICALT*），2022年，第66–70页。[3] H. Li，“基于神经网络的中学数学学科知识映射构建与应用研究”，硕士论文，云南师范大学，2022年。[4] S. Ji、S. Pan、E. Cambria、P. Marttinen 和 S. Y. Philip，“知识图谱综述：表示、获取与应用”，*IEEE*神经网络与学习系统汇刊，第33卷，第2期，第494–514页，2021年。[5] Z. Wang、J. Zhang、J. Feng 和 Z. Chen，“通过在超平面上平移进行知识图谱嵌入”，载于*AAAI*人工智能会议论文集，第28卷，第1期，2014年。[6] Y. Shen、Z. Chen、G. Cheng 和 Y. Qu，“CKGG：面向高中地理教育及更广领域的中文知识图谱”，载于国际语义网会议。施普林格，2021年，第429–445页。[7] C. Wang 和 S. Xu，“基于科学知识图谱的高校体育教学评价指标体系构建”，移动信息系统，第2022卷，2022年。[8] A. Vaswani、N. Shazeer、N. Parmar、J. Uszkoreit、L. Jones、A. N. Gomez、Ł. Kaiser 和 I. Polosukhin，“注意力就是一切”，神经信息处理系统进展，第30卷，2017年。[9] H. L. Nguyen、D. T. Vu 和 J. J. Jung，“智能系统的知识图谱融合：综述”，信息融合，第61卷，第56–70页，2020年。[10] Z. Kang、G. Shi、S. Huang、W. Chen、X. Pu、J. T. Zhou 和 Z. Xu，“面向多视图谱聚类的多图融合”，基于知识的系统，第189卷，文章编号105102，2020年。[11] A. L. F. Han、D. F. Wong 和 L. S. Chao，“结合中文特点的条件随机场中文命名实体识别”，载于语言处理与智能信息系统。施普林格·柏林·海德堡，2013年，第57–68页。[12] Y. Li、Q. Zhou、Y. Li、Z. Li、R. Liu、R. Sun、Z. Wang、C. Li、Y. Cao 和 H. T. Zheng，“前车之覆，后车之鉴：用于中文拼写检查的错误驱动的对比概率优化”，2022年。[13] N. Guarino、D. Oberle 和 S. Staab，“什么是本体？”，本体手册，第1–17页，2009年。[14] U. Von Luxburg，“谱聚类教程”，统计与计算，第17卷，第395–416页，2007年。