基于知识图谱技术的计算机教育学情内涵探索*

甘书灵 史东辉** 王园园

安徽建筑大学电子与信息工程学院，合肥 230601

摘 要 教育是中国古往今来关注的重点。伴随着新时代教育教学改革中计算机人才培养的需要，在计算机教育领域中学生学习情况的内涵探索可以帮助把握学生适应性学习方式和提供个性化教学服务。本研究整理教育研究者们在计算机教育学情领域研究的重点，利用文本挖掘技术从非结构化文本中获取计算机教育及学习过程的有用信息，结合自然语言处理中的 TF-IDF 算法和依存句法分析算法，将碎片化计算机教育和学生学习过程的领域知识进行抽取、整理和融合，构建可共享的可视化知识图谱，从而帮助更多的教育研究者把握计算机教育的工作内涵，以期为更多的高校教师提供数字化分析工具。

关键字 计算机教育，TF-IDF，依存句法分析，知识图谱

Exploring the Connotation of Computer Education Learning Based on Knowledge Graph

Shuling Gan Donghui Shi** Yuanyuan Wang

School of Electronics and Information Engineering
Anhui Jianzhu University
Hefei, China
donghui_shi@163.com

Abstract—Education is the focus of China's attention throughout the ages. Along with the need for computer talent training in the new era of education and teaching reform, the connotation exploration of students' learning in the field of computer education can help grasp students' adaptive learning methods and provide personalized teaching services. This study sorts out the key points of education researchers in the field of computer pedagogy, uses text mining technology to obtain useful information about computer education and learning process from unstructured texts, combined with TF-IDF algorithm and dependency parsing algorithm in natural language processing. The fragmented domain knowledge of computer education and students' learning process is extracted, sorted and integrated to build a shareable visual knowledge graph, so as to help more education researchers grasp the work connotation of computer education and provide more college teachers with digital analysis tools.

Keywords—Computer Education, TF-IDF, Dependency Syntactic Parsing, Knowledge Graph

1 引 言

目前，中国教育现代化取得重要进展，为进一步推动我国成为人才强国，《中国教育现代化 2035》强调了因材施教等八大理念，提出包括加快信息化时代教育变革等十大战略任务[1]。这意味着以学生为中心的精准教学和学情监测需要得到更进一步的重视，而计算机手段能够为教育内涵的分析提供智能化、可视化帮助。随着我国计算机技术的持续发展，高校计算机教育教学的方式和“线上线下混合式教学模式”的探索等成为研究的热点[2]。目前许多研究者在寻找高校计算机教育教学方法的创新手段，提出教育新需求新工具新发展。另外，还有研究者提出将学情分析与教学过程整合[3]，并强调计算机教育中自主学习能力的培养是教学过程的重点[4]等。学情分析是教学设计的重点工作，建立学情分析内容框架十分迫切[5]，这有利于实现教学认识和开展有效教学[6]。然而目前在计算机教育方面的学习过程内涵分析存在内涵空洞模糊等问题，处于经验式、印象式判断阶段[7-8]。虽然有许多学者从各方各面总结学情分析要素，但目前尚未有较为完整的学情分析框架。宋丹等人[9]提出基于学情数据构建智慧教学模式，学情个性识别技术和知识图谱技术能够为适应性学习提供决策支持。张喜征等人[10]提出可以通过构建知识图谱将碎片化知识整合成群体知识。赵宇博等人[11]提出学科知识图谱能够将学科知识资源进行有序组织，为学习者提供个性化服务。

基于以上经验，为了将计算机教育领域学情分析过程的碎片化知识进行整合，本文提出通过收集计算

*基金资助：本文得到安徽省教育厅高等学校省级质量工程项目（项目编号：No.2021cyxy022，2022xskc004）和安徽建筑大学教研项目（项目编号：No.zxjjxm092）资助。

** 通訊作者：史东辉。

机教育领域的相关文献构建知识图谱，然后利用自然语言处理技术将文献资源进行语义分析，将计算机教育学情的内涵、知识及影响因素等进行整合，提取出与学情有关的信息，例如教师教学方法、学生学习能力、课程知识难度等，从而为计算机教育创新发展导航、为可视化研究和查询推理提供方法支持，有利于学情知识库的传播共享。

2 数据来源与研究方法

2．1 样本来源与数据选择

本研究以 CNKI 中国学术期刊（网络版）中文数据库作为数据来源。通过高级检索选择主题“计算机 * (教育 + 教学 + 课程) * (学情 + 学习情况 + 学习模式 + 学习方式) (精确)”进行检索，检索时间跨度为 2013 年至 2022 年，来源类型是学术期刊，除去“编者按”、“开栏语”、重复文献及其他非相关文献，共获得 1721 篇有效文献。将以上获得的数据导出，每条题录数据包含标题、摘要和关键词等信息。

2．2 数据处理工具

本研究主要使用Python3.8[12]和图数据库neo4j[13]。Python编程语言在数据分析领域应用前景广阔，其工具箱涵盖了数据爬虫、数据分析和机器学习等常用库[14]。首先使用Python语言中jieba库，它是中文分词第三方库，通过调用jieba库完成中文分词工作，实现词频计算工作。然后采用HanLP中的parseDependency()方法实现基于神经网络的依存句法分析。最后将处理好的关系数据保存在图数据库Neo4j，它与传统的SQL数据库相比，能够利用图形结构数据存储功能实现复杂数据结构的关联关系和实时海量数据的存储和管理[15-16]，从而有效减少检索和分析数据的时间和精力。

2．3 数据分析方法

首先，采用词频-逆文本频率(Term Frequency-Inverse Document Frequency, TF-IDF)算法对数据文本进行高频词汇统计，获得关键词。TF-IDF算法可以对获得的文献进行词频权重分析，从而获得权重较高的特征词，并且过滤掉没有实际意义的常用词汇[17]，常常用于新闻文本、社会交流文本和医学文本等多个领域进行文本信息的检索与挖掘[18]。

其次，利用基于神经网络的依存句法分析（Dependency Syntactic Parser based on Neural Networks）对包含多个关键词的句子集合进行实体关系抽取，形成三元组数据表。依存句法分析注重句子内部组成成分的依存关系，其生成的依存句法树是一种图数据，句子中的词作为图的节点，而词与词的关系作为边[19]。而图神经网络在自然语言处理的应用是能够使每个节点循环地从其邻节点收集信息，即节点的隐含状态能够沿着图结构循环进行信息交换[20]，模型设计主要参考文献[21]。

最后，通过关联数据转化，构建学情分析领域的知识图谱。知识图谱是结构化的语义知识库，其基本组成单位是“实体-关系-实体”三元组以及“实体-属性”值对[22]，通过图结构的数据模型将零散抽象的知识进行可视化展示，为个性化推荐、智能问答等自然语言处理处理任务提供有力支持[23]。

3 有关计算机教育学情的知识抽取

3．1 基于 TF-IDF 算法的关键词抽取

文本挖掘工作是指利用智能算法从大量非结构化文本中抽取或标记文本中词语关系，进而实现结构分析、语义理解等功能。本研究对学情领域中有效文献的摘要文本数据进行知识挖掘，抽取关键词并进行词频分析。TF-IDF词频统计算法，词频TF表示关键词在文本中出现的频率,并且为了防止其偏向长文本，对其进行归一化处理：

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{1}$$

式（1）中：i用于给关键词t标记序号，j用于给文本d标记序号，$n_{i,j}$表示关键词t_i在摘要文本d_j中出现的次数，$\sum_k n_{k,j}$表示在d_j中所有关键词出现的次数总和。

逆文档频率IDF用于度量关键词的普遍重要性，即包含关键词t的文本d越少，IDF越大。

$$IDF_i = log\frac{|D|}{1+|\{j:t_i \in d_j|} \tag{2}$$

式（2）中：$|D|$是文本总数，$|\{j: t_i \in d_j|$表示包含关键词t_i的文本数量，为了避免分母为零的情况，即关键词t_i不存在文本d_j中，用$1 + |\{j: t_i \in d_j|$表示包含关键词t_i的文本数量。

词频-逆文档频率$TF - IDF$用于评估关键词对于文本库中特定文本的重要程度，可有效区别于其它文本语料库，其计算公式见式（3）。

$$TF - IDF_{i,j} = TF_{i,j} * IDF_i \tag{3}$$

根据以上公式进行实验计算，得到关键词及其$TF - IDF$值列表，表 1 列举了前 16 项关键词。

另外，提取关键词的算法还有许多，可以通过多次筛选提高关键词质量，有必要时，可以结合专业领域的知识语料以及人工筛选手段，选择合适的关键词。本研究结合人工筛选方法，删去例如“基于（TF-IDF值为0.03382）”、“本文（TF-IDF值为0.01428）”、“我国（TF-IDF 值为 0.00913）”、“一种（TF-IDF 值为 0.006816）”等与学情领域知识无关的词汇。

表 1 基于 TF-IDF 的关键词（前 16）

序号	关键词	TF-IDF	序号	关键词	TF-IDF
1	学习	0.27557	9	实践	0.03807
2	教学	0.16896	10	模式	0.03435
3	学生	0.11246	11	研究	0.03418
4	课程	0.08921	12	设计	0.03370
5	教师	0.06052	13	知识	0.03327
6	教育	0.05962	14	教学模式	0.03292
7	学习者	0.04531	15	课堂教学	0.03166
8	课堂	0.04026	16	评价	0.03043

3．2 基于神经网络的依存句法分析

在自然语言处理中，依存语法是指用词语之间的依存关系来描述语言结构，可以理解为核心词（也可以称为支配词，head）和依存词（也可以称为从属词，dependency）之间有着依存关系（Relation）。一般来说利用语法关系构建依存关系，形成树结构。常见语法关系包括主谓关系、并列关系、动宾关系等，更多依存关系见文献[24]，表 2 列举了学情文献中关键词的重要依存关系。

例如通过对“数字化教学是推动教学改革、适应信息技术快速发展的必经之路。”这一句话进行句法结构分析，获得如图 1 所示的依存关系示例，由此可以得到关键词（words）及其词性标签（Part-of-speech tags，简写为 POS tags）和依存关系标签（Dependency labels，也可以称为弧标签，arc labels）。例如序号为 1 的词语“数字化”是动词，序号为 2 的词语“教学”是名词，这两者的关系是定中关系。

1	数字化	数字化	v	v	_	2	定中关系	_	_
2	教学	教学	n	n	_	3	主谓关系	_	_
3	是	是	v	v	_	0	核心关系	_	_
4	推动	推动	v	v	_	3	动宾关系	_	_
5	教学改革	教学改革	i	l	_	13	定中关系	_	_
6	、	、	wp	w	_	7	标点符号	_	_
7	适应	适应	v	v	_	5	并列关系	_	_
8	信息	信息	n	n	_	9	定中关系	_	_
9	技术	技术	n	n	_	11	主谓关系	_	_
10	快速	快速	d	d	_	11	状中结构	_	_
11	发展	发展	v	v	_	7	动宾关系	_	_
12	的	的	u	u	_	11	右附加关系	_	_
13	必经之路	必经之路	nz	nz	_	4	动宾关系	_	_
14	。	。	wp	w	_	3	标点符号	_	_

图 1 依存关系示例

借助 DependencyViewer 可视化软件获得依存句法分析的可视化关系，如图 2 所示。词语之间的非对称关系用一个有向弧表示，由依存词指向核心词，从分析结果可以知道，句子的核心谓语是“推动”,主语是“数字化教学”，宾语是“教学改革、适应信息技术快速发展的必经之路”。

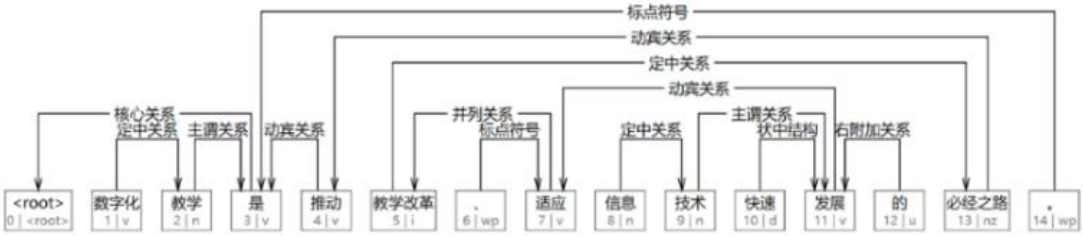

图 2 依存关系可视化示例

然而简单的依存句法分析可能存在多种句法结构导致的语义分歧，即相同句子的不同依存句法结构。为了减小句法结构中的语义分歧，利用神经网络和依存句法分析组合的联合模型，生成带有概率关系的依存结构图，并将依存关系概率作为神经网络信息传递的权值，利用神经网络多层训练的优势，对关键词的所有语义进行训练和预测。神经网络模型如图 3 所示。

由图3可知输入层包括words、POS tags和arc labels，分别对应向量x^w，x^t和x^l，即输入层:

$$X = [x^w, x^t, x^l] \tag{4}$$

同时使用矩阵W来表示传播参数，隐藏层的传播参数是:

$$W_1 = [w_1^w, w_1^t, w_1^l] \tag{5}$$

使用cube函数作为隐藏层激活函数可以捕获来自关键词、词性标签和依存关系三个维度特征的不同组

合形式，即：

$$h = (w_1^w x^w + w_1^t x^t + w_1^l x^l + b_1)^3 \quad (6)$$

最后Softmax层使用激活函数得到输出：

$$Z = softmax(w_2 h) \quad (7)$$

本研究首先通过TF-IDF算法结合人工筛选得到200个领域关键词，获得包含三个及其以上关键词的句子集合Ω，然后利用融合神经网络的依存句法分析获得三元组关系，主要保留并列关系、主谓关系、动宾关系、定中关系和状中关系，删去没有实际意义的附加关系和非关键词之间的关系，表2展示了依存关系三元组示例。

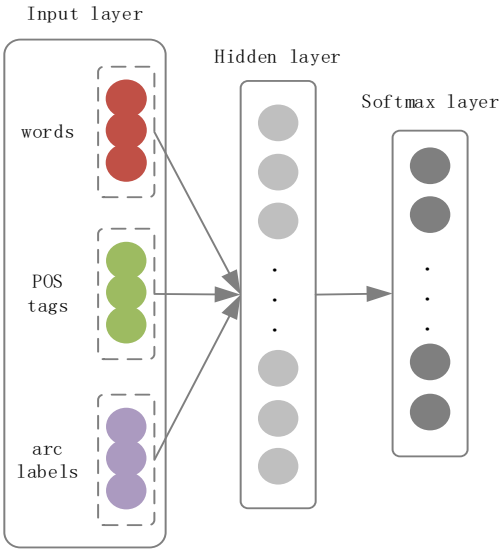

图 3 神经网络模型

表 2 三元组对应关系

序号	依存词	依存关系	核心词
1	自主	状中关系	学习
2	合作	状中关系	学习
3	学习	定中关系	目标
4	学习	定中关系	进度
5	学习	定中关系	平台
6	策略	并列关系	指导
7	指导	并列关系	评价
8	教学	定中关系	模式
9	教学	定中关系	内容
10	激发	动宾关系	兴趣
11	设计	动宾关系	活动
12	组织	动宾关系	资源

4 计算机教育学情知识图谱的构建及内涵探索

4．1 知识图谱的构建

计算机教育领域学情知识图谱的构建是通过知网平台获得研究者对计算机教育及学习过程等相关的研究文献，利用关键词提取方法建立学情碎片信息的知识集合，通过融合神经网络的句法依存分析模型对知识联系进行构建，获得三元组数据，最后进行人工筛选整理，将数据进行转化，从而构建学情领域知识图谱。知识图谱构建过程如图 4 所示。

知识图谱能够以图谱形式描述学情教育领域关注实体、概念及其关系，将其与句法依存分析得到的三元组关系进行知识对齐，从而搭建完整的知识网络，表 3 展示依存语法与知识图谱的三元组对应关系。而可视化分析技术能够帮助用户直观了解和分析知识信息，从而提高知识图谱的表达。

本研究利用 neo4j 图数据库管理系统设计构建知识图谱，将三元组数据进行存储和关联关系的可视化展示，图 5 展示了学情领域知识图谱部分内容。由此可见，与计算机教育教学过程相关的因素十分丰富，与之密切相关的学习过程也存在许多要把握的重点，而计算机属于新型的教育分支，具有其特色课程。

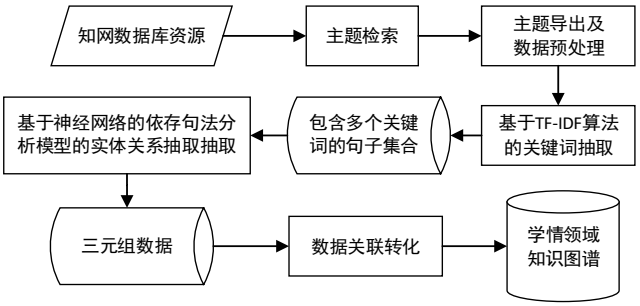

图 4 知识图谱构建过程

4．2 计算机教育学情内涵探索

通过知识图谱节点之间关系的呈现，可以分析计算机教育的学情内涵，计算机教育的学情内涵不仅包括计算机科学基础知识和编程能力，还应注重培养学生的创新思维和实践能力、信息素养等方面的能力。

表 3 三元组对应关系

方法名称	三元组形式		
依存语法	依存词	依存关系	核心词
知识图谱	实体	关系	实体
	概念	属性	属性值

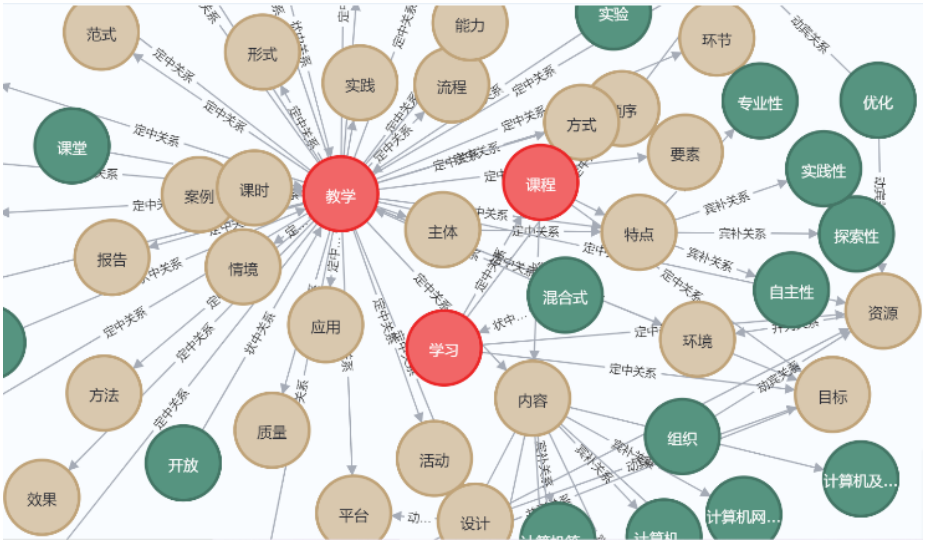

图 5 学情知识图谱（部分知识网络）

这些能力的培养可以通过理论教学、实践操作和项目实践等方式。下面将从课程、学习和教学三个主要方面进行讨论分析：

（1）课程方面：以课程为节点的知识图谱的内容如图 6 所示。通过对课程内容进行深入分析，发现针对计算机教育课程方面需要把握内容、知识点、目标等，同时要重点关注知识点的组织结构、关联性和难易程度等方面内容。另外，计算机教育课程具有自主性、探索性、实践性和专业性等特点，需要专注培养自身算法优化能力、用计算思维解决科学问题的能力、知识和技能的转化能力、信息检索能力、合作探究能力和外语综合运用能力等内容。将课程与学习、教学两大方面建立联系，通过关系挖掘，把握学生在学习过程中对不同知识点上的掌握程度，为教学过程提供科学指导，从而为优化课程内容提供依据。

（2）学习方面：以学习为节点的知识图谱的内容如图 7 所示。通过对学生学习进行深入分析，可以发现针对计算机教育的学习过程要重点把握方式、策略、目标、过程、动力、经验、课程、模式、活动、体验、竞争、机会等多方面内容。利用知识图谱技术，可以对学生的学习行为进行分析，包括学习时间、学习进度、学习方式、学习评价和学习策略等方面，从而实现对学生的学习情况和学习效果的评估。通过对更多的学生学习过程产生的数据和关联路径进行分析，可以发现学生的学习习惯和学习特点，为制定个性化的学习计划提供参考。同时，学习与教学建立联系，为教师提供改进教学策略的依据，进而指导教师制定更加个性化的教学方案。

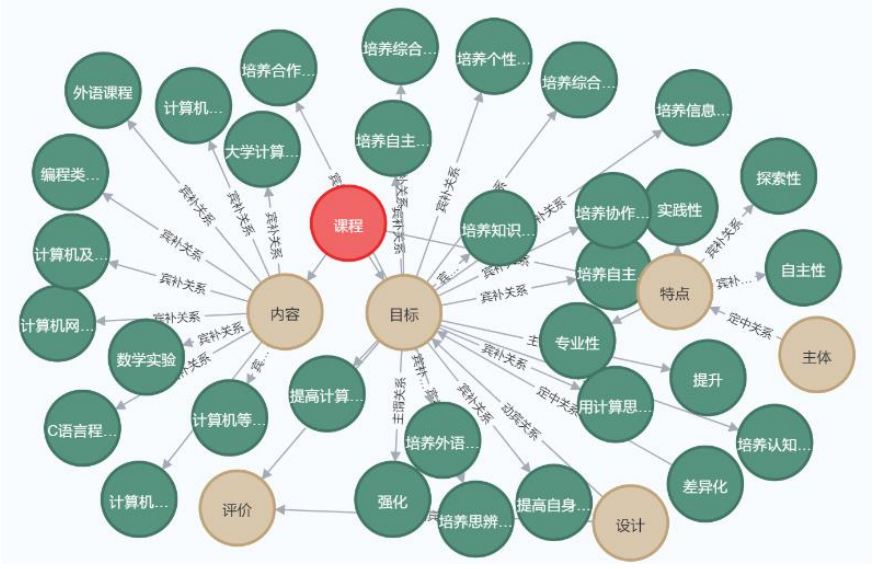

图 6 以课程为节点的知识图谱的部分内容

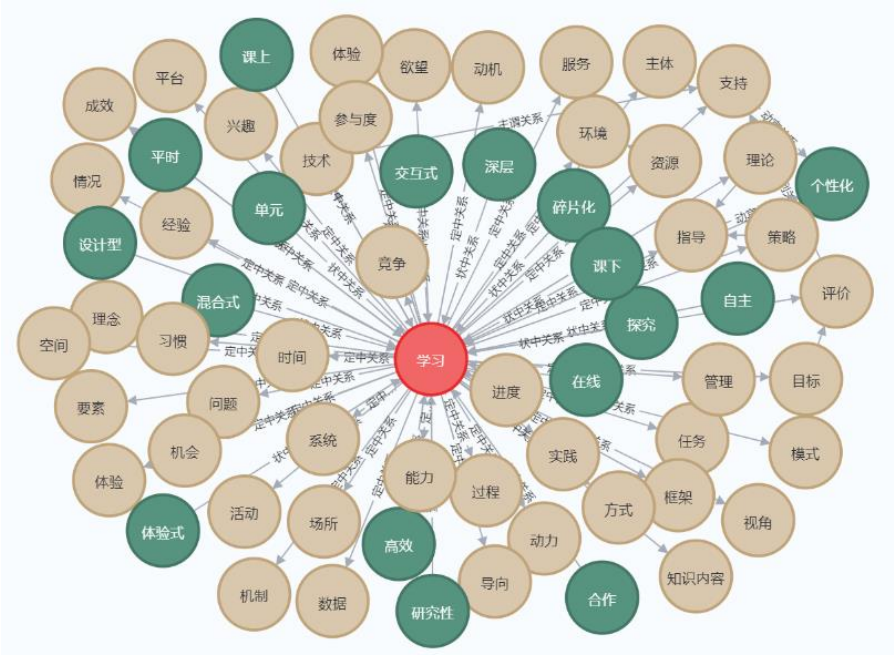

图 7　以学习为节点的知识图谱的部分内容

（3）教学方面：以教学为节点的知识图谱的内容如图 8 所示。通过对教师教学进行深入分析，可以发现计算机教学要把握设计、案例、技术、质量、形式、实践、环节、平台、资源、情境等多方面内容，教学改革要为激发学习兴趣、变革学习方式、强化学习能力、提升综合素质等多方面服务，在自主学习、合作学习、探究学习、混合式学习、设计型学习、交互式学习、体验式学习等多特征学习方式方法进行创新，从而提高学生积极性、灵活性和动态获取知识内容等能力。同时要重点关注知识点的组织结构、关联性和难易程度等方面内容。根据分析结果，优化计算机教育的教学方法，例如针对学生的不同学习能力和兴趣，采用不同的教学策略；针对学生容易出错的知识点，加强教学重点等。

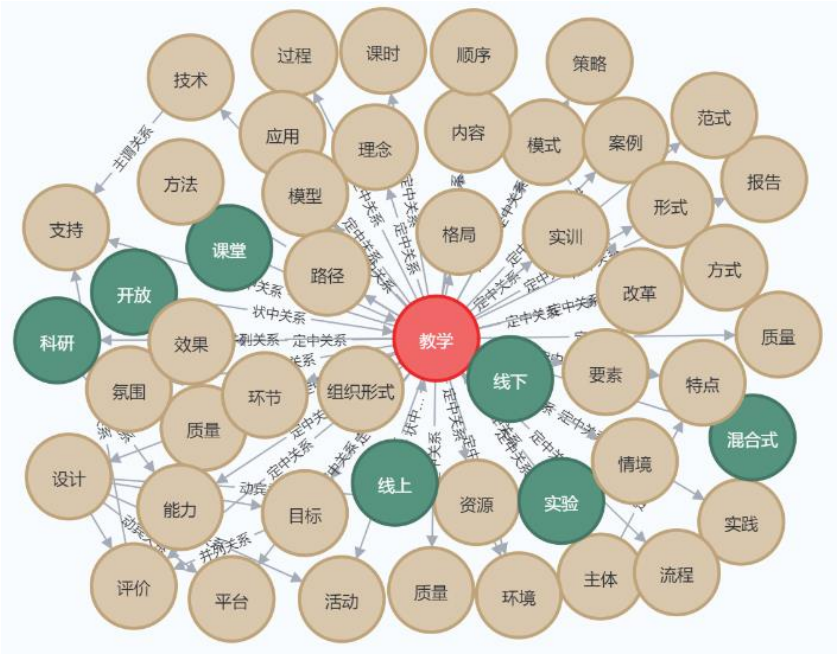

图 8　以教学为节点的知识图谱的部分内容

利用知识图谱技术可以更加详细地分析计算机教育的学情内涵，为教学提供更加科学、有效的指导和支持。为了把握计算机教育和学情分析的各项内容，教育研究者们可以参考计算机教育领域学情知识网络，

帮助了解学生的学习特点和需求，制定更加完善的教育计划和教育改革目标，优化教学内容和方法，落实教育政策、完善学情分析机制，提高教学效果，促进现代化教育的文档蓬勃发展。

5 结束语

本文基于知识图谱技术对计算机教育的学情内涵进行了探索。通过收集计算机教育领域的相关研究文献，利用自然语言处理技术将文献资源进行语义分析，结合 TF-IDF 算法统计领域关键词，使用融合神经网络的依存句法分析获得三元组关系，从而构建知识图谱，辅助计算机教育学情的内涵和影响因素的探索和知识整合。根据知识图谱中所呈现的关系和问题，教师可以从课程、学习和教学三方面对计算机教育进行学习情况的把控。

为了获得以上方法实施效果，将以上研究结果用于国家级双语教学示范课程《C++面向对象程序设计》的教学实践中。首先构建该课程本体库和知识图谱。通过知识图谱可以知道学生在课堂、作业和实验等教学实施过程具有较强的学习差异性，并且学生的学习方式、学习风格等对其成绩也有差异性影响。针对学习态度较差的学生以及学习方式模糊的学生，教师可以更进一步关注他们的学习动态，重视教育学中差异性教学的重要性，达到学情预警的作用。另外，学生在双语教学过程中的表现受英语水平影响较大。教师在制定教学目标时应当根据知识点难度合理分配教学资源和教学任务，在课堂、作业和实验过程中强化重、难点知识的教授，并且可以将英语作为一个教学指标，在双语教学中适当加入专业英语知识的学习。由于学生学习情况与许多因素相关，知识图谱能够提供更加全面的学情分析途径，在计算机教育领域，从更多的角度去考量学情因素及其影响，能够辅助教师把握计算机教育学情内涵、提高教育教学效果、促进学生差异性能力培养。

未来将利用学情领域知识图谱实现计算机教育领域知识的智能问答和智能推荐等功能，为计算机教育创新发展导航提供方法支持，利用大数据手段实现学情知识库和教育知识图谱的传播共享。

参 考 文 献

[1] 中华人民共和国教育部.中共中央、国务院印发《中国教育现代化 2035》[EB/OL].[2023-02-16]. http://www.gov.cn/zhengce/2019-02/23/content_5367987.htm

[2] 柏琪,许睿婧,余星星.高校“线上线下混合式教学模式”的探索与实践[J].计算机技术与教育学报,2022, 10(02):75-78.

[3] 安桂清.论学情分析与教学过程的整合[J].当代教育科学,2013, 373(22):40-42.

[4] 赵钦.大学计算机教育中自主学习理念的渗透[J].教育理论与实践,2011,31(36):61-62.

[5] 邵燕楠,黄燕宁.学情分析:教学研究的重要生长点[J].中国教育学刊,2013, 238(02):60-63.

[6] 邵朝友,朱伟强.以课例研究为载体开展学情分析[J].中国教育学刊,2015, 262(02):73-76.

[7] 马文杰,鲍建生.“学情分析”:功能、内容和方法[J].教育科学研究,2013, 222(09):52-57.

[8] 王琪,靳莹.中等教育学段学情分析研究述评[J].教育理论与实践,2023, 43(02):54-57.

[9] 宋丹,胡瑛,方正军等.基于学情数据的智慧教学模式研究与实践[J].高等工程教育研究,2022. 197(06):116-120.

[10] 张喜征,罗文,蔡月月.基于知识图谱的用户生成内容平台中碎片化知识整合研究[J].科技管理研究,2019, 39(05):159-165.

[11] 赵宇博,张丽萍,闫盛等.个性化学习中学科知识图谱构建与应用综述[J/OL].计算机工程与应用:1-24[2023-02-20].

[12] Python[EB/OL].[2023-02-16]. https://www.python.org/

[13] Neo4j[EB/OL].[2023-02-16]. http://neo4j.org/.

[14] 姜秋香,郭伟鹏,王子龙等.Python 语言在水文水资源领域中的应用与展望[J/OL].计算机工程与应用:1-16[2023-02-20].

[15] 赵雪芹,杨一凡,于文静.基于 Neo4j 图数据库的工程档案知识图谱构建及应用[J].档案与建设,2022, 401(05):48-51.

[16] Saad Mohamed,Zhang Yingzhong,Tian Jinghai,Jia Jia. A graph database for life cycle inventory using Neo4j[J]. Journal of Cleaner Production,2023,393.

[17] 董伟,董思遥,王聪,陶金虎.基于 TF-IDF 算法和 DTM 模型的网络学习社区主题分析[J].现代教育技术,2022, 32(02):90-98.

[18] WANG Zhuohao,WANG Dong,LI Qing.Keyword Extraction from Scientific Research Projects Based on SRP-TF-IDF[J].Chinese Journal of Electronics,2021, 30(04):652-657.

[19] 范涛,王昊,吴鹏.基于图卷积神经网络和依存句法分析的网民负面情感分析研究[J].数据分析与知识发现,2021, 5(09):97-106.

[20] 陈雨龙,付乾坤,张岳.图神经网络在自然语言处理中的应用[J].中文信息学报,2021, 35(03):1-23.

[21] Danqi Chen and Christopher Manning. 2014. A Fast and Accurate Dependency Parser using Neural Networks. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 740 - 750, Doha, Qatar. Association for Computational Linguistics.

[22] 刘峤,李杨,段宏等.知识图谱构建技术综述[J].计算机研究与发展,2016, 53(03):582-600.

[23] 侯中妮,靳小龙,陈剑赟等.知识图谱可解释推理研究综述[J].软件学报,2022, 33(12):4644-4667.

[24] Stanford Typed Dependencies Manual[EB/OL].[2023-02-16]. https://downloads.cs.stanford.edu/nlp/software/dependencies_manual.pdf