

文章编号: 2095-2163(2022)02-0137-07

中图分类号: TP393.09

文献标志码: A

基于知识关联树的知识协同模型研究及应用

白伟华, 朱嘉贤, 蔡文伟

(肇庆学院 计算机科学与软件学院, 广东 肇庆 526061)

摘要:融合精准教育、多领域知识和多种应用技术的个性化教育是大数据时代的新型工程培养模式。面对这种知识协同的需求,本文应用知识图谱和知识挖掘技术相结合,构建一种基于知识关联树的知识服务体系;融合了“知识”、“资源”、“用户”和“服务”4个核心要素,构建知识结构与资源相互联系的关联树,同时将用户的学习需求映射成核心知识关联树,实现为用户提供推荐精准学习路径(关联的核心知识集)和学习资源的知识协同服务。

关键词:知识协同;知识关联树;知识挖掘;知识服务;个性化教育

Research and application on knowledge collaborative model based on knowledge association tree

BAI Weihua, ZHU Jiaxian, CAI Wenwei

(Department of Computer Science, Zhaoqing University, Zhaoqing Guangdong 526061, China)

[Abstract] Integrating precision education, multi-domain knowledge and various application technologies for personalized education is a new engineering cultivation model in the era of big data. For these demands of knowledge collaboration, a knowledge service system based on a knowledge association tree is presented. With the combination of four core elements of knowledge, resources, users and services, it constructs an association tree which links the knowledge structure and resources, and maps the learning needs of users into the core knowledge association tree to provide users with recommend precise learning paths (associated core knowledge sets) and knowledge collaboration services with learning resources.

[Key words] knowledge association; knowledge association tree; knowledge mining; knowledge service; personalized education

0 引言

大数据支持下的智慧教育和个性化精准服务相结合,形成了个性化精准教育的新模式,面对多用户、多角色和复杂知识域的相互关联环境,将用户与领域的知识和相应的应用技术相融合的服务模式——知识协同,实现对知识的管理和个体学习指引,既满足了个性化学习需求又要符合培养工科团队协作的需要。利用知识图谱技术对多领域知识实现关联融合是一种新的有效的解决方案。应用知识图谱和大数据资源管理技术相结合的知识服务体系,能实现多领域知识的协同及资源管理的优化配置,符合培养复合型技术团队的需求^[1]。“知识”、“资源”、“用户”和“服务”是大数据时代背景下智慧教育的4个核心组成要素;知识是资源的索引,资源是知识的承载体,以服务满足用户的需求,服务实现的方式就是为用户提供个性化、精准化和有效的数据资源(知识)。

1 相关研究

在知识协同的应用中,主要有:

(1)基于知识超网络模型。是一种由几个不同类型的知识网络所组成的超网络;

(2)知识协调机制。基于成员之间互依的知识进行协调的机制;这些机制和模型着重强调了用户知识学习、学习小组中各成员之间知识分享和协助性地解决问题,主要通过超网络这一模型构建基于以某领域知识或任务为目标的用户关系网络、载体知识文本网络或知识进化网络^[2]。

学习路径的智能推荐是知识协同的重要研究内容,研究人员提出通过学习者的多维度且动态的个体多参数模型来反映学习者多方面状态的“学习者建模”方法,以实现精准个性化学习的学习路径推荐^[3]。在构建基于多维个性化参数的学习者模型中,为实现学习路径推荐,学者提出了5个核心个性化参数:“学习目标”即需掌握的一个或多个的知识

基金项目:广东省教育科学规划项目(2019KTSCX199);肇庆市科技专项资金项目(2020G1004);肇庆学院校级项目(2019012612, ZLGC201933);广东高校公共计算机课程教学教改项目(2021-GGJSJ-012)。

作者简介:白伟华(1976-),男,博士,副教授,主要研究方向:云资源调度、大数据分析、知识协同。

通讯作者:朱嘉贤 Email:1030013@qq.com

收稿日期:2021-12-01

理论,或专业知识体系;“技能学习”即知识应用于实践的需求;“知识背景”即个体知识体系特征,或当前已经具备的知识理论;“时间限制”和“学习风格”^[4]。研究人员根据个性化参数所描述的方面,将19项个性化参数归并为以下三个维度:

(1)描述个体需求方面的学习动机和实现目标对应的参数项——“为何学”;

(2)描述知识体系结构及要求方面的知识点及其体系构成和实践技能对应的参数项——“学什么”;

(3)描述学习者个体特征方面的已掌握的知识体系结构、学习偏好、学习能力及风格等对应的参数项——“如何学”。

结合学习路径推荐的需求,可以将“学习目标”、“技能学习”、“知识背景”、“时间限制”和“学习风格”对应的个性化参数项按上述3个维度构建为树结构:学习者模型(为何学(学习目标,学习动机)、学什么(技能学习)、如何学(知识背景,时间限制,学习风格))^[5-6]。

在网络通信技术的支持下,面对海量增长的学习资源,有效地标识对应数据资源的特征是其能被智慧推荐、个性化精准推荐和智慧教育所用的重要手段。学习资源是一种数据资源,是一种涵盖知识内容(学习者需要掌握的知识)逻辑结构,承载着知识点内容信息的实体^[7]。为解决高效地利用并共享海量的学习资源这一问题,一方面要有有效的资源标注模型,另一方面是要有安全快捷的资源共享模型。结合信息体及其逻辑结构特征,在标注技术上,目前采用的主要方法有:本体(Ontology)、语义Web、XML、元数据(Metadata)、概念图(Concept map)、知识图谱(Knowledge Graph)等^[8]。学习资源是广泛存储在不同的服务节点上,该模式本就是一个去中心化的存储,应用区块链去中心化的模式,研究人员提出了基于区块链技术的在线学习资源管理模式^[9]。

当前所提出的学习资源管理、共享和应用模型,以及相关应用的研究中,可以发现这些方法存在以下问题:

(1)无法将海量的数据资源、个性化精准化的知识学习任务、知识学习路径规划和用户角色四者的关联与学习过程中复杂的知识路径进行映射和合理性规划判定;

(2)针对智慧学习,这些方式无法满足用户个性化、特征化和精准化的需求,无法对学习过程和进度进行量化,以及学习效果的评价和预判;

(3)目的性不明确,无法针对个体和群体分配合理的学习任务和路径规划,无法评估相关数据资

源的有效性和必要性。

本文提出了一种针对知识逻辑结构、资源内容结构和用户知识背景结构的知识关联树映射模型;提出了一种基于知识关联树的学习资源搜索、推荐和应用的模式;为支持针对个性化特征的学习路径推荐,提出了一种基于知识背景和知识结构相似度计算的推荐方法。

2 知识关联树模型设计

2.1 概念结构建模

在网络上,知识通过资源呈现出来,用户通过资源的学习获取知识,知识是抽象体,而数据资源是知识的承载体。在组织和应用上,要呈现出以下的相互关系或关联:

(1)同系列的知识之间的层次关系。例如:通过广义表描述“数据结构”这门课程涵盖的知识点之间的层次关系,数据结构(线性表(栈(表达式求值,迷宫求解,数制转换……)),队列(树遍历,图遍历……)),串(……),数组(……),……),树(二叉树(遍历,哈夫曼编码,……)),B树/B+树(……),……),图(……),……)。

(2)用户(学生)具备或学习的专业领域的核心知识的层次关系。例如:用户A(计算机(数据处理(数据结构(搜索(算法(……)),排序(……)),……)),……),数据库(……),……),……),电子信息(……),……)。

(3)知识之间的聚合关联。不同系列知识或不同课程之间,其知识点有部分存在聚合关联的,例如:“数据库”和“数据结构”这两门课程中针对“排序”、“搜索”、“B树/B+树”等,又如“操作系统”和“数据结构”中的“栈”、“队列”等,这些知识点都存在聚合关联,而这些关联是同层次上的,但又属于不同知识系列,即无向无序的。

(4)用户(学生)之间基于领域知识的耦合关联。在知识协同应用中,用户需要共同完成一个任务,每位团队成员有各自的职责,完成相应的子任务,成员在该任务中所需具备的核心领域知识是互补的、个体的,但整体上基于任务又是完整的,所以用户之间基于核心领域知识映射在不同的结点上,是一种耦合关联。

在应用过程中,存在用户与知识结点之间、资源与知识结点之间以及用户与资源之间3类有向的量化关联,其量化关联可被描述为两者之间的权重(系数)。

(1)用户与知识结点之间的权重描述了用户在

其核心知识领域或学生对各知识掌握的需求程度;

(2)资源与知识结点之间的权重描述了数据资源涵盖或对相应知识点的支撑程度,也可以表示资源对学习相应知识点的有效程度;

(3)用户与资源之间的权重描述了用户(学生)对该资源在相应知识点的有效性评价,也可以表示对该资源的偏好程度。

2.2 逻辑结构及形式化定义

为描述各结点之间的层次关系、耦合关联、聚合关联以及不同类别结点之间的量化关联,整个模型知识结构树和关联由两个基本元素组成。

定义 1 课程知识结构树是一棵深度为 4 的多分支树,由分支结点(第 1~3 层结点)和叶子结点(第 4 层结点)组成。其中,树根结点(即第 1 层)到第 3 层上的分支结点统一被映射为:课程——基础内容分支——知识点分布,叶子结点(第 4 层结点)为资源结点。 T_c - 课程知识结构树:

$$T_c = \langle t_{sub}, t_{source}, R_t \rangle$$

其中:分支结点 t_{sub} 为一个四元组: $t_{sub} = (t_{id}, f, key, d)$; t_{id} 为唯一标识树结点的编号 id ; f 为学科领域集; key 为知识点关键词集; d 为知识点定义或文本描述。

叶子结点(资源) t_{source} 为一个五元组:

$$t_{source} = (S_{id}, S_k, S_d, S_{ly}, S_{URL});$$

其中, S_{id} 为资源标号唯一标识的 id 编码; S_k 为知识点关键词集; S_d 为资源描述; S_{ly} 为资源类别描述; S_{URL} 为资源存储链接地址。

树内结点关联 R_t 描述了 T_c 内两个结点之间的关联属性,用一个四元组表示:

$$R_t = (r_m, r_s, w, t)$$

其中, r_m 为主结点(双亲结点)的编号 id ; r_s 为从结点(孩子结点)的编号 id ; w 为两者关联程度权重; t 为关联的类别(分支结点关联或分支结点与叶子结点关联两种情况)。

定义 2 用户知识结构树

$$U_i = (u_{ifo}, \langle u_f, T_c \rangle, R_{u_i})$$

其中: u_{ifo} 为用户信息;

$u_{ifo} = \{u_i | (u_{id}, u_{age}, u_{deg}, u_{dept}, \dots)\}, i = 1, 2, \dots, n$;
 u_{id} 成员 id ; u_{age} 年龄; u_{deg} 学历; u_{dept} 所在单位及部门等;
 $\langle u_f, T_c \rangle$ 为描述用户的专业领域,其掌握的核心知识及其对应的树; R_{u_i} 是用户与技能关联,描述用户在每个核心知识上的能力以及对应的权重系数。

定义 3 用户资源关联 $R_{us} = (u_{id}, S_{id}, C, D)$, 表示了用户与资源之间的关联信息。

其中, u_{id} 为用户的 id 编码; S_{id} 为资源结点编码; C 是用户对资源的评价、评分等; D 则是由用户对资源的评价、评分等核算的偏好程度权重。

定义 4 知识学习协同树 $KT = (T_p, M_F, R_{PF})$ 。

其中, T_p 为领域核心知识树,描述各专业领域的主要核心知识结构; M_F 是有限集合, $M_F = \{T_{c_1}, T_{c_2}, \dots, T_{c_m}\}$ 是被拆分后子对应领域涵盖的核心知识树集; m 为拆分后的课程数; R_{PF} 为领域内核心课程关联,描述对应领域内核心知识的组成及对应的权重系数(即对应知识点的重要程度描述)。

定义 5 知识点聚合关联 $R_{kc} = (K_{id}, t_{id}, C_c, F)$, 表示了课程间相近、类似或相同的知识点,能有效地描述课程群重复的知识点,能协同各课程学习中对知识点的学习时间的分配或掌握程度。

其中, K_{id} 为聚合关联 id ; t_{id} 为可聚合知识点 id 集合; C_c 是描述该聚合关联的文本或知识点公认定义(有效定义); F 为该知识点对聚合的关联程度或贡献度权重系数。

“用户—课程知识结构树—资源—关联”例子,如图 1 所示,描述了用户(U20210123, 梁家栋, ……),专业领域(计算机, 自动化),课程(数据结构 $T_{c_{DS}}$, 数据库 $T_{c_{DB}}$, 操作系统 $T_{c_{OS}}$, 信号处理 $T_{c_{SP}}$), 课程中知识点聚合关联 R_{kc} (队列, B 树/B+树)以及相互关联。

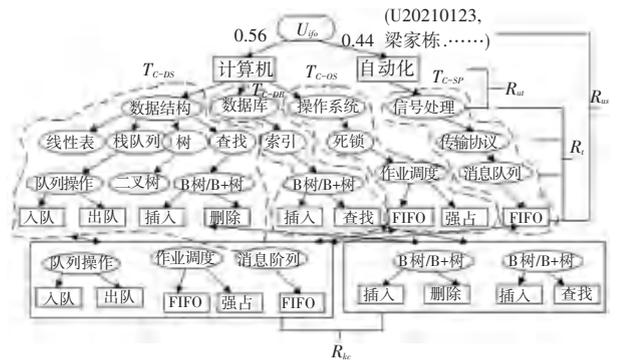


图 1 “用户—课程知识结构树—资源—关联”实例
Fig. 1 An example of user-knowledge structure tree-resources-relation

3 知识关联树的运算及应用

3.1 构建 T_p -领域核心知识树

领域核心知识树是整个知识协同模型的核心元素,是各类聚类计算的基本单元。结合学科领域知识的标注,构建领域核心知识树的主要步骤如下:

(1)以《中华人民共和国国家标准学科分类与代码》(GB/T 13745—2009)以及知网上的关键词条作为结点核心关键字,新建领域核心知识树中第 1~2 层结点,同时以 GB/T 13745—2009 对结点 id 进行编码,并完成“树内结点关联”实现“双亲结点与孩

子结点”之间的一对多映射关联。

(2)通过网络爬虫,完成两类基础数据的爬取。

①以(1)中第1~2层结点中的关键字为“核”,获取当前知名高校的专业培养方案提纲及课程参考,构建专业领域内的核心课程;

②根据课程安排,以课程名为“核”,爬取相关课程的参考教程及其目录,并以目录为结点,构建课程

核心知识点结点。

“计算机科学与技术”专业领域核心知识树实例,如图2所示,参照《中华人民共和国国家标准学科分类与代码》(GB/T 13745-2009)对树根节点和分支节点进行编码,一方面对领域进行标准化分类,另一方面也方便并提高搜索树中节点或路径的效率。

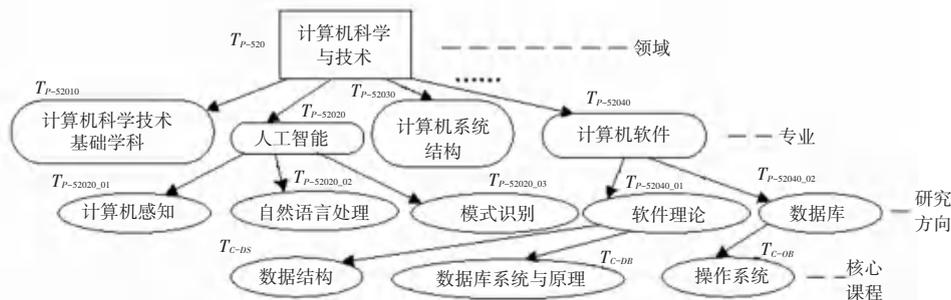


图2 “计算机科学与技术”专业领域核心知识树 T_p (部分)

Fig. 2 The part of the T_p of computer science and technology

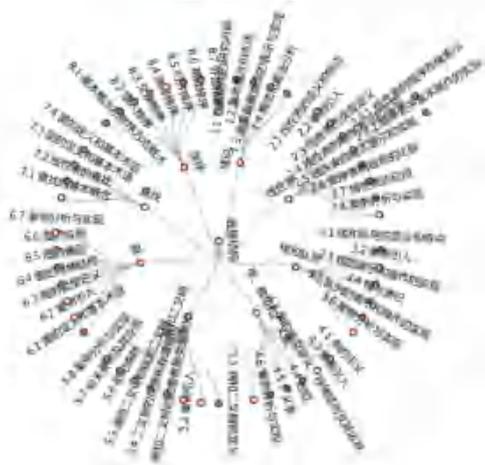
3.2 构建 T_c -课程知识结构树

T_c -课程知识结构树是整个模型的核心结构树,是模型的交互层,实现上下层各实体之间的媒介,其上层实体是用户、领域核心知识树以及知识学习协同树,下层实体主要有资源以及知识点聚合关联 R_{kc} ,对应的一个实例如图1所示。构建 T_c 时,选取目前在相应课程上最为经典的教程为模板,构建基于该教程为模板的课程知识结构树,其主要过程为:

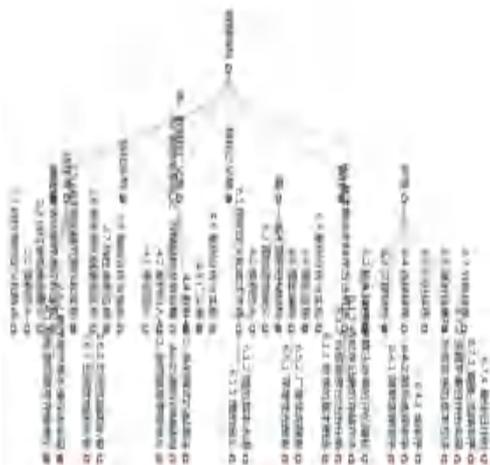
(1) 获取教程的目录结构,将目录按其大纲级别映射到结构树的第2~3层的结点上,同时将其上下层关联插入到树内结点关联 R_i 中;

(2) 标注第2~3层分支结点中的数据域:“key”和“d”,同时为每个第3层结点至少生成一个孩子结点 t_{source} ,并初始化该孩子结点。

图3为可视化生成的“数据结构”知识结构雷达树图和一般树。



(a) 雷达树图



(b) 一般树图

图3 “数据结构”知识结构雷达树图和一般树图

Fig. 3 The knowledge structure tree of the course of data structure

3.3 发现知识点聚合关联 R_{kc}

根据开源分词器—Jieba 实现以《中华人民共和国国家标准学科分类与代码》(GB/T13745-2009)和具有公认性的词条:“知网词条”为核心,定义核心关键词的自定义词典,并以“单词_词性_词频”的格式加载到知识点聚合关联运算中。

本文提出了基于标注关键词相似度的知识点聚合关联 R_{kc} 。首先,计算出知识结构树中结点在核心关键词的相似度;其次,以相似度值进行聚类;最后,以阈值进行过滤完成知识点聚合关联。

3.3.1 计算核心词的相似度

利用早期的研究成果对知识结构树中第三层结

点中的数据域:“key”中“研究领域核心词”计算两关键词 w_1, w_2 的相似度。关键词相似度函数公式

$$sim(w_1, w_2) = \frac{\alpha \cdot p(a) \cdot (s(w_1) + s(w_2))}{\beta \cdot (p(a) \cdot (s(w_1) + s(w_2)) + p(a \cdot w_1) \cdot s(w_1) + p(a \cdot w_2) \cdot s(w_2))} \cdot \frac{\log(10 \cdot \max_{-} p(a \cdot w))}{\delta} \quad (1)$$

其中, $p(a)$ 为两关键词 w_1, w_2 共同祖先节点所在的位置; $s(w_1)$ 和 $s(w_2)$ 表示两关键词 w_1, w_2 在树中的位置; $p(a \cdot w_1)$ 和 $p(a \cdot w_2)$ 表示与两关键词 w_1, w_2 共同祖先节点的位置差; $\max_{-} p(a \cdot w) = \max(p(a \cdot w_1), p(a \cdot w_2))$, 即 $p(a \cdot w_1)$ 和 $p(a \cdot w_2)$ 中的最大值。

3.3.2 以相似度值进行聚类

在词相似度基础上实现知识点聚合关联度计算。将聚合关联的知识点结构及关键描述匹配的相似度转化为对两棵子树的相似度计算,进而再转化为结构中各结点的数据域“key”的相似度计算。沿用前期研究结果,其知识点聚合关联度计算为公式(2)^[10]:

设 t_{sub_i} 和 t_{sub_j} 是分别待验证聚合关联的两知识点结点 (T_c 第 3 层上的分支结点), k_i 和 k_j 为结点数据域“key”中关键词的数量, F_T_n 表示 t_{sub_i} 和 t_{sub_j} 子树包含的分支数。

$$sim(t_{sub_i}, t_{sub_j}) = \frac{1}{F_T_n} \sum_{K=1}^{F_T_n} \max_{1 \leq p \leq k_i, 1 \leq q \leq k_j} sim(w_{p_i}, w_{q_j})$$

$sim(w_1, w_2)$ 定义为公式(1)^[10]:

其中, w_{p_i} 和 w_{q_j} 为结点数据域“key”描述词集合中第 p, q 个关键词, 计算子式 $\max_{1 \leq q \leq k_j, 1 \leq p \leq k_i} sim(w_{p_i}, w_{q_j})$ 表示相似度取在描述关键词集内词相似度 $sim(w_1, w_2)$ 的最大值。

3.3.3 知识点聚合关联

根据经验设定阈值 θ_{min} , 通常以课程与专业需求联系紧密度作为权重值参考, 检索不同课程 T_c 第 3 层上的分支结点的关键字, 并按公式(2)计算不同课程间相应分支结点的相似度, 当 $sim(t_{sub_i}, t_{sub_j}) \geq \theta_{min}$ 时, 则产生对应结点的知识点聚合关联, 并生成 R_{kc} 记录。

依照上述过程对不同科目中相关相似或重复的知识点进行聚类分析, 并形成知识点聚合关联 R_{kc} 。计算机专业课程群知识点聚合关联 R_{kc} (部分) 实例如图 4 所示。

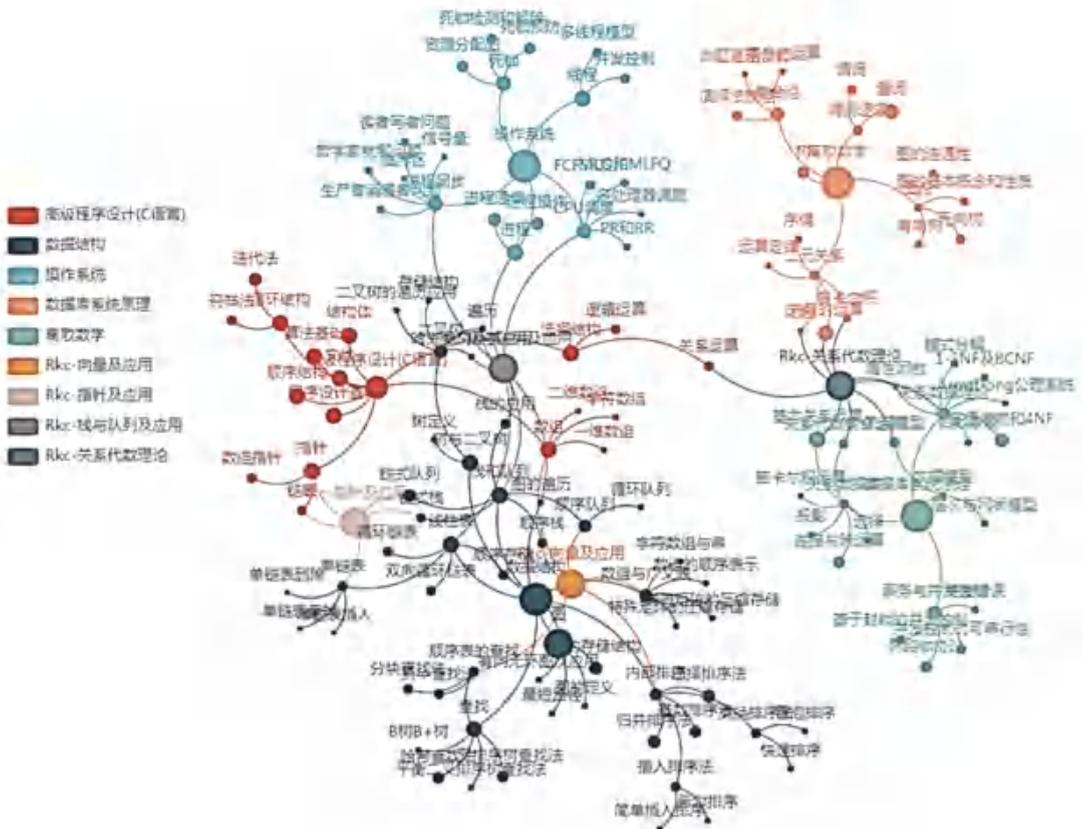


图 4 计算机专业课程群知识点聚合关联 (部分) 例子

Fig. 4 The aggregation and association of knowledge points R_{kc} in the computer professional curriculum group

通过分析各知识点的核心关键词以及主要内容文本描述,实现对课程群知识点聚合关联的数据挖掘。如图4所示,通过对《高级语言程序设计(C语言)》、《数据结构》、《操作系统》、《数据库系统原理》和《离散数学》的知识点分析,挖掘出4个知识点聚合关联分别为: R_{kc} —向量及应用、 R_{kc} —指针及应用、 R_{kc} —栈与队列及应用和 R_{kc} —关系代数理论,并以可视化方式呈现所聚合关联的知识点集合。

3.4 学习路径映射应用

依据发现知识点聚合关联过程,在学习路径映射上,首先按照学习目标或职位需求,基于知识关联树逻辑结构,构建对应的专业知识/技能图谱,通过图谱上的核心关键词,利用公式(1)和公式(2)进行

知识点聚合关联运算,完成库内知识关联树的检索,并生成基于该图谱的学习路径映射集。

如图5所示,以“大数据专业工程师”专业知识/技能图谱为索引,通过知识点聚合关联运算,可以找到对应的学习路径映射集:“专业理论基础知识”(知识点1(链表,栈,队列,栈与队列,线性表,单链表,循环链表),知识点2(快速排序,堆排序,插入排序,内部排序,快速排序算法,插入排序法)……)、“技能应用”(技能1(高级程序设计(Java),Java/Scala,Spark,Hadoop)……)。

通过得到的学习路径映射集合,构建以知识点和技能为结点的有向无环图,完成基于知识关联树的学习路径规划,并推荐给学习者。

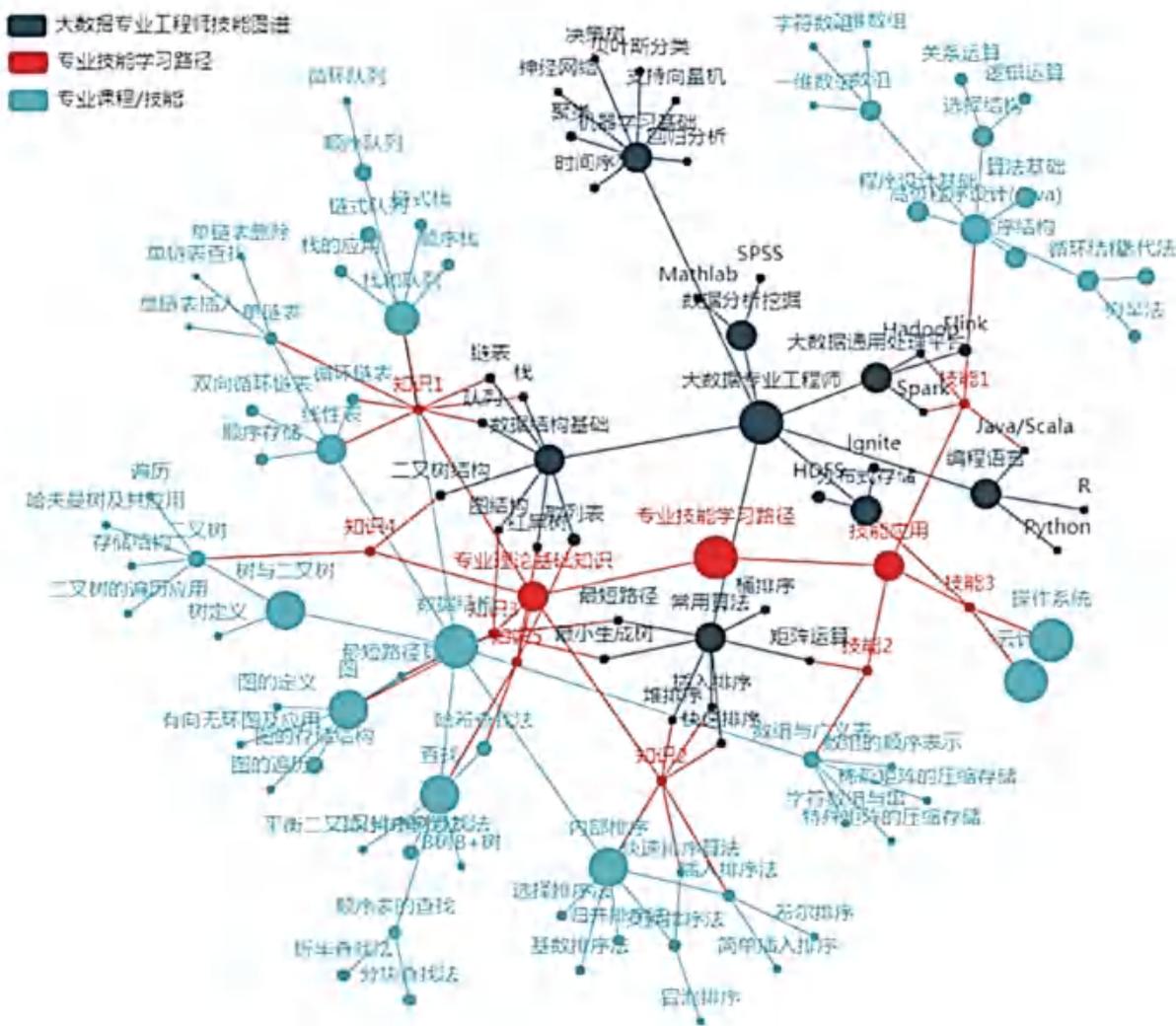


图5 学习路径映射应用——大数据专业工程师(部分)实例
Fig. 5 A part of the learning path mapping for big data professional engineers

4 结束语

多领域知识与技术融合的知识协同培养模式。本文将知识图谱和大数据资源管理技术相结合,构建一个个性化精准教育是大数据技术支持下一种基于 (下转第147页)

个性化精准教育是大数据技术支持下一种基于