计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(11):100−110 [doi: 10.15888/j.cnki.csa.008798]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于事件抽取的学科建设知识图谱构建与应用①

李家瑞, 李华昱, 闫　阳, 付亚凤

(中国石油大学 (华东) 计算机科学与技术学院, 青岛 266580)
通信作者: 李华昱, E-mail: lhyzj@upc.edu.cn

摘　要: 学科建设是高校发展的核心, 随着高校学科建设的不断深入与强化, 学科建设信息持续增加, 且以离散的文件组织形式难以对学科建设成果进行高效的管理, 不利于后续分析与评估工作的开展. 针对此问题, 对学科建设知识图谱的构建及相关应用进行了研究. 首先通过 BERT-BiLSTM-CRF 模型对学科建设文本进行事件抽取, 并使用爬虫进行相关知识的补充. 然后选择属性图模型存储知识, 完成学科建设知识图谱的初步构建. 基于构建好的知识图谱, 搭建了学科建设可视化系统, 并引入最小斯坦纳树算法实现智能问答应用. 最后, 通过对学科建设事件抽取与智能问答方法进行实验分析, 验证了本文所提出方法的有效性.

关键词: 知识图谱; 学科建设; BERT-BiLSTM-CRF; 斯坦纳树; 智能问答; 事件抽取

引用格式: 李家瑞,李华昱,闫阳,付亚凤.基于事件抽取的学科建设知识图谱构建与应用.计算机系统应用,2022,31(11):100–110. http://www.c-s-a.org.cn/1003-3254/8798.html

Construction and Application of Discipline Construction-oriented Knowledge Graph Based on Event Extraction

LI Jia-Rui, LI Hua-Yu, YAN Yang, FU Ya-Feng

(College of Computer Science and Technology, China University of Petroleum, Qingdao 266580, China)

Abstract: Discipline construction is the core of the development of colleges and universities. With the deepening and strengthening of discipline construction in colleges and universities, the information on discipline construction increases continuously. Nevertheless, the results of discipline construction can not be effectively managed in the manner of discrete document organization, which is not conducive to subsequent analysis and evaluation. To solve this problem, this study focuses on the construction and further application of discipline construction-oriented knowledge graphs. For this purpose, events are extracted from discipline construction texts by the BERT-BiLSTM-CRF model, and related knowledge is supplemented by the crawler. Then, the property graph model is selected to store knowledge, and a preliminary discipline construction-oriented knowledge graph is thereby built. Subsequently, this knowledge graph is availed to build a visualization system for discipline construction, and the minimum Steiner tree algorithm is adopted for the application of intelligent question answering. Finally, the validity of the proposed method is verified by experimental analysis of the methods of discipline construction-oriented event extraction and intelligent question answering.

Key words: knowledge graph (KG); discipline construction; BERT-BiLSTM-CRF; Steiner tree; intelligent question answering; event extraction

学科是高等学校事业发展的基础, 其建设水平代表着高校的办学质量和竞争优势[1]. 近年来, 随着我国建设世界一流大学和一流学科方案的提出, 高等教育的水平和质量在突飞猛进, 高校发展也处于激烈的竞

① 基金项目: 山东省自然科学基金面上项目 (ZR2020MF140); 中国石油大学 (华东) 研究生创新工程 (YCX2021128)
收稿时间: 2022-02-22; 修改时间: 2022-03-30; 采用时间: 2022-04-13; csa 在线出版时间: 2022-07-15

争之中. 作为高校工作的核心, 学科建设对于高校的发展愈发重要. 因此, 加强对高校学科建设现状的全面掌控, 分析实际存在的问题并寻找有效、合理的解决途径, 对于促进学科建设、推动高校发展有着极大的促进作用. 由于学科建设具有综合性强、覆盖范围广等特点[2], 通过传统的文件整理方法难以对学科建设的成果进行高效的管理, 而且后期对其考核的工作难度大, 容易消耗大量的人力和时间. 因此研究如何将分散、无序的高校学科建设成果, 从科研水平、人才培养、基地建设等维度进行全面整合, 并实现学科建设信息的高效查询和直观显示, 对于了解高校学科建设水平具有很强的现实意义.

2012 年, Google 首先提出知识图谱 (knowledge graph, KG) 的概念, 旨在融入现有的搜索引擎以提高搜索结果的质量. 知识图谱以图的形式对有关联的实体和概念进行融合, 可以对现实世界的事物和它们之间的关系进行形式化的描述, 其基本单位是 (实体, 关系, 实体) 三元组和“实体-属性”值对[3]. 作为大数据时代下一种新型高效的知识组织方式, 知识图谱已经取得了极大的发展, 比如规模较大的通用知识图谱 Freebase、DBpedia、Wikidata 等, 其中覆盖了现实世界中大量的常识性知识. 同时知识图谱技术在诸多领域中都有相关的研究与实现, 如搜索引擎、法律法规[4,5]、医疗诊断[6,7] 等, 为目标领域的建设发挥了积极的促进作用, 被认为是推动互联网和人工智能发展的核心驱动力之一.

本文面向高校计算机学科领域, 结合知识图谱方法构建学科建设知识图谱, 对高校的学科建设成果进行全面的信息整合, 同时研究知识问答和可视化展示等应用技术, 使得相关人员能够更加直观地了解高校的学科建设水平. 本文首先通过深度学习方法, 以事件抽取的方式从文本中抽取出学科建设事件触发词和事件元素, 并通过爬虫爬取网络资源进行领域知识的补充. 之后基于属性图模型, 对获取到的知识进行存储, 实现学科建设知识图谱的构建. 然后, 针对自然语言形式的用户提问, 利用知识图谱的图型结构, 研究了一种基于斯坦纳树的智能问答方法. 最后搭建了学科建设可视化系统, 整合相关信息查询、多关键词搜索、智能问答等多种功能, 并将查询结果以力导向图等方式进行呈现. 本文的主要技术路线如图 1 所示.

本文的组织结构如下: 第 1 节介绍基于 BERT-BiLSTM-CRF 和爬虫的学科建设知识图谱的构建过程; 第 2 节介绍基于斯坦纳树的智能问答方法; 第 3 节对本文采用的事件抽取模型和智能问答方法进行实验分析; 第 4 节介绍学科建设可视化系统的搭建与评估; 第 5 节为结论与展望.

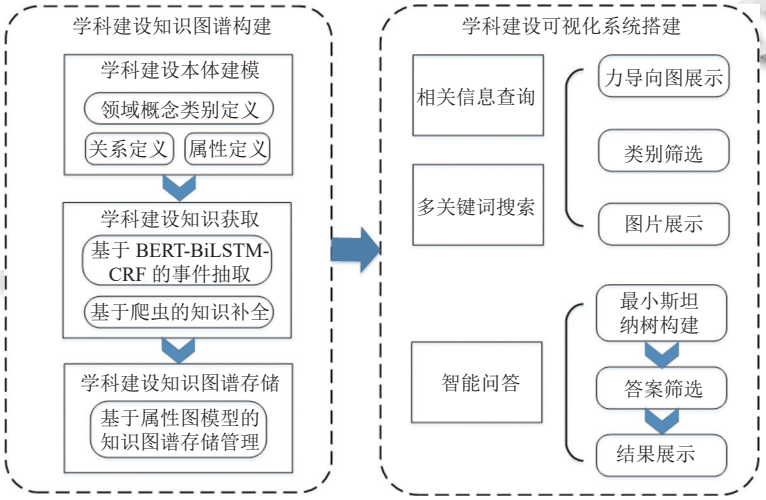

图 1 技术路线图

1 学科建设知识图谱构建

1.1 学科建设本体模型

在构建知识图谱时, 通常有两种构建思路: 自顶向下和自底向上. 其中, 自顶向下的构建方式是指先为知识图谱定义好本体与数据模式, 再将实体关系等加入到知识库中; 而自底向上指的是先从一些开放的数据中提取出知识, 选择其中置信度较高的加入到知识库中, 然后再构建上层的本体模型. 由于领域知识图谱是

面向具体的领域构建, 只有包含高准确度的知识才能为上层应用提供研究基础, 因此通常采用自顶向下的构建流程. 在自顶向下构建知识图谱时, 需要首先定义好本体模型与数据模式. 本体作为某一具体领域内知识的规范化描述, 对领域知识图谱的类集、关系集、属性集等进行了形象化定义, 是对知识图谱模式层的管理[8]. 通过构建本体模型, 可以形式化地表达出特定领域中各类概念及其间关系, 为用户提供对该领域知识的共同理解, 并对实体、关系以及实体属性等进行约束规范, 作为后续知识抽取与组织的指导.

在经过查找资料、咨询专家等形式的研究后, 使用 Protégé工具构建了学科建设知识图谱的本体模型[9], 本体模型通过 OntoGraf 功能进行展示的效果如图 2 所示.

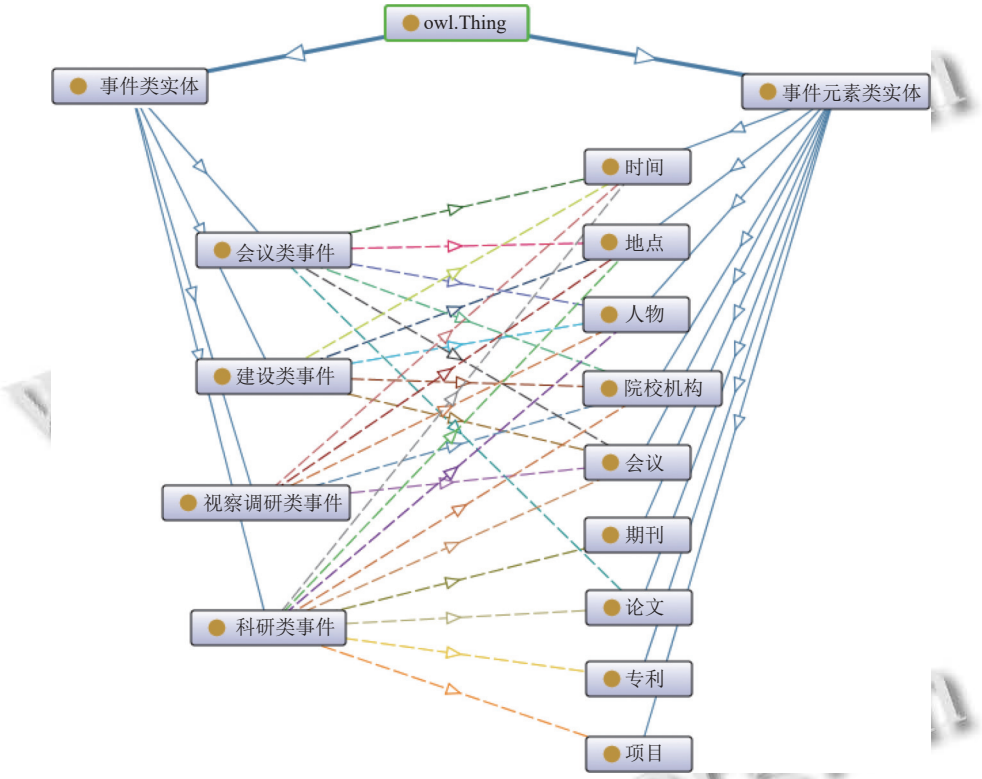

图 2 学科建设本体模型

本体中主要包含“科研类事件”“会议类事件”“建设类事件”“视察调研类事件”共 4 个事件类实体类别, 以及“时间”“地点”“人物”“院校机构”“会议”“期刊”“论文”“专利”“项目”共 9 个事件元素类实体类别. 概念之间通过多种语义关系相互关联, 关联关系可以使用 owl 中的 ObjectProperty 进行形式化描述. 例如, 会议类事件相关的语义关系通过 owl 表示如下.

```
<owl:ObjectProperty rdf:about="#会议时间">
<rdfs:domain rdf:resource="#会议类事件"/>
<rdfs:range rdf:resource="#时间"/>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:about="#会议地点">
<rdfs:domain rdf:resource="#会议类事件"/>
<rdfs:range rdf:resource="#地点"/>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:about="#参会人员">
<rdfs:domain rdf:resource="#会议类事件"/>
<rdfs:range rdf:resource="#人物"/>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:about="#会议名称">
<rdfs:domain rdf:resource="#会议类事件"/>
<rdfs:range rdf:resource="#会议"/>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:about="#参会院校机构">
<rdfs:domain rdf:resource="#会议类事件"/>
<rdfs:range rdf:resource="#院校机构"/>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:about="#会议相关论文">
<rdfs:domain rdf:resource="#会议类事件"/>
<rdfs:range rdf:resource="#论文"/>
</owl:ObjectProperty>
```

构建好的学科建设本体可以作为后续学科建设事

件抽取和学科建设知识图谱存储的指导. 在抽取学科建设事件时, 按照事件类概念进行触发词抽取, 进而判断事件的类型并抽取事件相关元素; 将知识图谱中的数据存储至图数据库时, 根据本体中的概念与语义关系, 映射生成数据库中的节点和边的标签分类, 保证了知识结构的完整性.

1.2 学科建设知识获取

知识图谱由一条条知识组成, 知识是否正确及其覆盖范围是知识图谱能否成功实现的关键, 因此如何正确获取所需知识是知识图谱构建的基础. 现实世界的数据主要分为 3 类: 结构化数据、半结构化数据和非结构化数据, 而学科建设领域知识主要包含于文本类非结构化数据以及网页类半结构化数据中. 例如, 对于文本“2020 年 8 月 29 日, 中国石油大学 (华东) 与东软教育科技集团有限公司签署合作协议, 共建石大东软青岛软件学院”, 其中包含的学科建设领域知识可以用三元组的形式表示为: (石大东软青岛软件学院, 建设时间, 2020 年 8 月 29 日)(石大东软青岛软件学院, 建设单位, 中国石油大学 (华东))(石大东软青岛软件学院, 建设单位, 东软教育科技集团有限公司).

经实际分析及效果评估后, 本文以互联网中的网络资源作为主要数据源, 从高校网站中爬取包含知识较为丰富的文本数据进行学科建设事件的抽取, 然后以“百度百科”作为实体属性的数据来源, 爬取其中的半结构化数据完成知识的补全.

1.2.1 事件抽取

在文本抽取方面, 抽取任务主要包括实体抽取、关系抽取、属性抽取及事件抽取等. 其中, 事件抽取是自然语言处理中一种经典的信息抽取任务, 旨在抽取出非结构化文本中的事件信息并以结构化的形式展现, 在舆情监测、文本摘要、军事情报等领域中有着广泛的应用[10]. 由于事件通常由代表事件发生的事件触发词以及描述事件结构的元素所构成, 因此事件抽取主要包括事件触发词抽取、事件元素抽取和事件属性抽取等. 为了更好地抽取高校学科建设文本中的知识, 本文对其结构特征进行研究后, 将文本抽取任务定义为事件抽取任务.

学科建设事件抽取主要包括以下步骤: (1) 学科建设文本数据获取, 从权威的高校官方网站中获取相关的文本类非结构化数据. (2) 文本数据预处理, 结合学科建设领域的文本特征, 去除原始爬取数据中的无用分句. (3) 通过 BERT-BiLSTM-CRF 模型进行事件的触发词和事件元素抽取, 抽取时采用 BIO 标记方法. 其中事件触发词用于判定事件类型, 本文定义的学科建设事件类型及事件元素分类如表 1 所示. (4) 将抽取结果整理后, 以知识形式进行保存. 本文在进行事件抽取时, 依据学科建设领域的事件特征, 去除了原始文本中与客观事件无关的表述, 并选择合适的粒度划分事件类型和事件元素种类. 对于其他领域的事件抽取任务, 可以参照该领域的事件特征, 调整相关数据的预处理及标注过程, 从而将以上事件抽取流程应用至目标领域, 实现方法的有效迁移.

表 1 学科建设事件类型及事件元素分类

概念	具体分类
事件类型	科研类事件、会议类事件、建设类事件、视察调研类事件
事件元素	时间、地点、人物、院校机构、会议、期刊、论文、专利、项目

使用 BERT-BiLSTM-CRF 模型进行序列标注, 主要包括基于 BERT 的词向量表示、基于 BiLSTM 的上下文特征学习以及基于 CRF 的最大标签序列输出 3 个部分[11]. 基于 BERT-BiLSTM-CRF 模型的学科建设事件触发词和事件元素抽取的过程为: 首先利用 BERT[12] 模型作为特征表示层, 获取输入文本中的字符所对应的向量化表示; 然后将得到的向量输入到 BiLSTM[13] 中, 以学习输入文本中的上下文特征信息; 最后通过 CRF[14] 对 BiLSTM 模型的输出序列进行处理, 结合前后文序列标签的相关性, 输出最终的标记序列. 图 3 为本文所使用的 BERT-BiLSTM-CRF 模型的结构.

每个 LSTM 单元的结构如图 4 所示, 通过遗忘门、输入门、输出门 3 种结构来控制单元的状态[15]. 其中, 遗忘门决定需要遗忘前一个单元中的哪些信息; 输入门控制当前单元需要加入哪些信息以更新单元状态; 输出门决定输出哪些信息. 在时刻 t 单元的状态更新可由式 (1)–式 (5) 得到:

$$f_t = \sigma\left(W_f \times [H_{t-1}, X_t] + b_f\right) \tag{1}$$

$$i_t = \sigma(W_i \times [H_{t-1}, X_t] + b_i) \tag{2}$$

$$C_t = f_t \times C_{t-1} + i_t \times \tanh(W_C \times [H_{t-1}, X_t] + b_C) \tag{3}$$

$$o_t = \sigma(W_o \times [H_{t-1}, X_t] + b_o) \tag{4}$$

$$H_t = o_t \times \tanh(C_t) \tag{5}$$

其中, f、i、o分别代表遗忘门、输入门和输出门; σ为 Sigmoid 激励函数; W和b分别是权重矩阵和偏置项; H代表隐藏层状态; C代表单元状态.

1.2.2 知识补全

通过模型抽取出的事件元素, 其缺少相关的实体属性, 例如对于实体“中国石油大学 (华东)”, 还应具有“创办时间”“简称”以及图片等类型的属性. 因此, 本文以百度百科为数据来源, 使用 Requests 库从百科页面中爬取已抽取实体的相关属性信息, 从而对知识进行补充[16]. 基于百度百科的知识补全流程如图 5 所示.

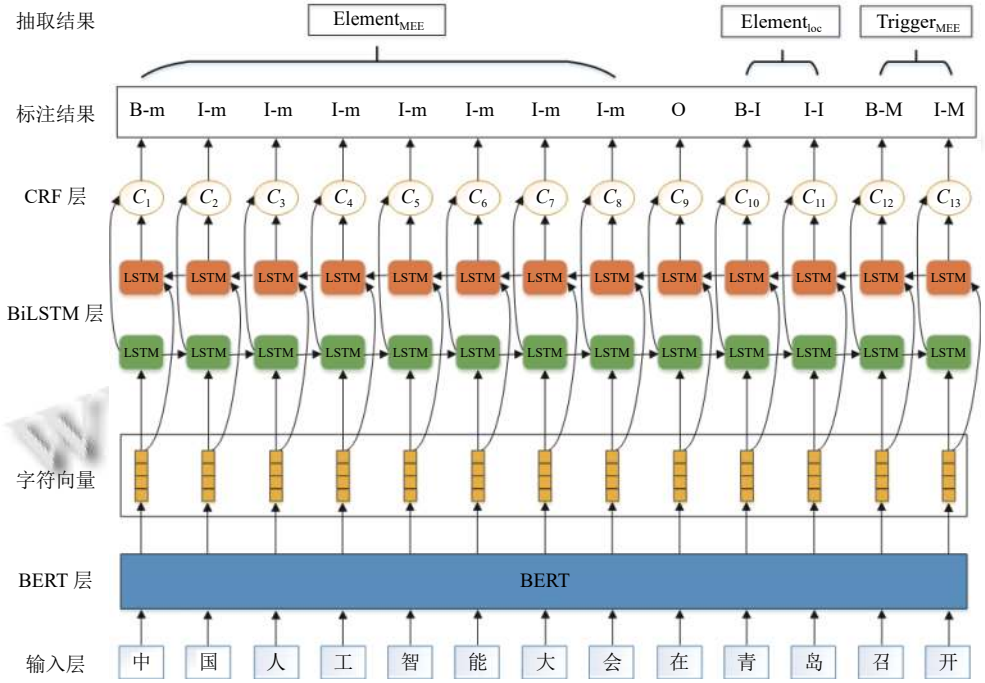

图 3 BERT-BiLSTM-CRF 模型结构

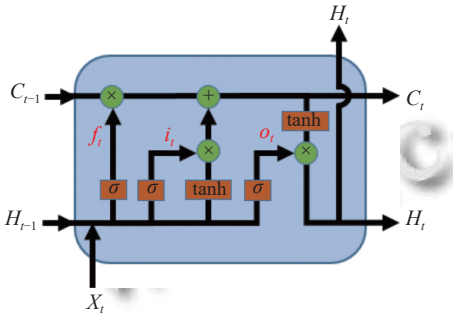

图 4 LSTM 单元结构

(1) 定位目标页面 URL. 通过观察发现百度百科网页的 URL 具有一定的规律性, 因此可以将“https://baike.baidu.com/item/”与目标实体的名称进行拼接, 形成待爬取的目标页面 URL.

(2) 发起请求并获取页面数据. 得到目标页面的 URL 后, 使用 Requests 库的 get 方法发起请求, 并获取响应回来的页面数据.

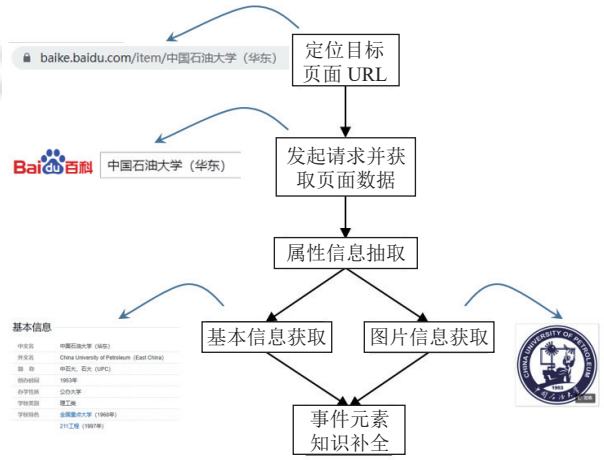

图 5 基于百度百科的知识补全流程

(3) 属性信息抽取. 从获取到的页面数据中解析出基本信息数据以及说明图集链接, 对于基本信息数据, 通过去除空白字符等预处理操作后, 形成实体的各类

属性信息; 如果页面中包含说明图集链接, 则按照新的 URL 继续爬取, 获取实体的图片信息. 通过以上步骤, 最终完成目标实体的属性信息抽取过程.

1.3 学科建设知识图谱存储

建立本体并获取到有效的知识后, 需要选择合适的方式存储知识图谱. 本文使用有向标签属性图作为数学模型, 按照表 2 所示策略对学科建设知识图谱数据进行管理[17]. 在属性图模型中, 将实体对象存储为节点, 将实体间的语义关系存储为边, 每个节点和边都具有唯一 ID, 节点和边可以具有多个标签, 然后使用抽取的实体属性信息对节点属性进行相应的填充.

表 2 学科建设知识图谱存储策略

数据类型	数据结构	形式	作用
节点	标签	字符串	区分不同类型的节点
	属性	键-值对	可存储多个节点相关的属性
边	标签	字符串	区分不同类型的关系

2 基于斯坦纳树的学科建设智能问答

智能问答是指通过与用户进行交互, 精确定位和理解用户所提出的问题, 最终以一问一答的形式给出问题的准确回答[18]. 由于用户问题的灵活性, 以及自然语言语法的复杂性, 因此要做到为复杂问题提供精确回答的难度较大. 早期的问答主要通过构造问答模板的方法来实现, 但这种方法需要生成大量的模板, 人工处理成本高昂, 而且在一个领域中已经构建的问答模板不易直接迁移到另一个新的领域中, 导致其可复用性差[19]. 目前大部分的研究集中在基于深度学习的方法, 以此实现自动化的智能问答系统. 但是这种方法需要大量的训练数据集来保证结果的质量, 前期需要投入较多的精力和时间.

对于用户所提出的问题, 其期望的答案通常与问题中的关键词具有一定的联系, 而知识图谱中包含了大量的实体关系, 所以在包含问题所有关键词的关系子图内, 很有可能含有此问题的对应答案. 根据上述思路, 本文提出了一种基于最小斯坦纳树的学科建设智能问答方法. 其中, 最小斯坦纳树算法是指为了将指定点集合中的所有点连通, 允许在给定点集外增加额外的点, 使得生成的最短网络开销最小的算法[20]. 通过将最小斯坦纳树算法引入知识图谱中, 能够快速匹配到包含问题所有关键词的关系子图, 从而降低问题的处理成本并保证结果的准确性.

2.1 学科建设智能问答流程

学科建设智能问答的处理流程如图 6 所示, 主要包括以下步骤.

(1) 对用户输入的原始问题进行分词处理, 为了保证分词的准确度, 本文使用了 jieba 分词工具中的自定义词典功能, 防止领域实体被分割为不完整的词语.

(2) 将分词序列与已经构建好的学科建设知识图谱中的数据进行匹配, 提取其中学科建设相关的实体, 形成基准节点集.

(3) 结合知识图谱与基准节点集, 运行最小斯坦纳树算法, 得到能够包含所有基础节点的最小斯坦纳树.

(4) 对最小斯坦纳树进行扩充, 形成一个包含问题答案的中间子图, 去除子图中的基准节点后得到候选节点集.

(5) 进一步分析用户输入问题, 去除问题分词序列中的助词、连词等无效词, 形成问题关键词集合.

(6) 使用 BERT 模型对问题关键词进行分类, 基于分类结果筛选候选节点集, 并展示最终的预测答案.

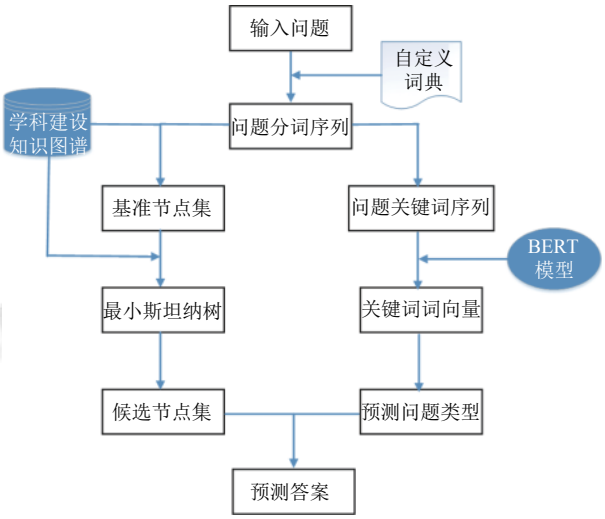

图 6 学科建设智能问答处理流程

2.2 基于知识图谱的最小斯坦纳树构建

结合用户问题与学科建设知识图谱构建出的最小斯坦纳树, 是后续生成预测答案的基础. 在解析用户问题时, 首要的工作就是对问题进行准确的分词. 对于具体领域, 仅使用现有的分词词典会无法识别领域内的专业词汇, 降低分词的准确度, 进而导致问题解析不够准确. 因此本文根据学科建设领域具有的特点, 构建面向领域的分词词典 Dic_T. 该分词词典主要包括通用分

词词典和领域分词词典两部分. 其中, 通用分词词典包括了中文语言的所有常用词; 而领域分词词典指的是结合学科建设领域的背景知识, 构建出的具有针对性的分词词典, 包括领域内的各种专业词汇, 如人名、地名、院校机构名称等. 根据构建出的分词词典对用户提出的问题进行文本分词, 得到分词结果$\{W_{T1},W_{T2},\cdots,W_{Tn},W_{S1},W_{S2},\cdots,W_{Sm}\}$.

分词完成后, 考虑到分词结果中可能会存在大量的无意义词汇, 对问题解析的价值不大, 而且会影响后续问答处理过程的速度和准度, 因此本文构建了具有领域特点的停用词词典Dic_S, 用以过滤用户问题中的无效信息. 此停用词词典包括常用停用词词典和专用停用词词典两部分. 其中, 常用停用词词典包括中文语言的日常停用词和标点符号; 专用停用词词典指的是结合学科建设领域的背景知识, 构建出的针对该领域的停用词词典, 包括价值低、不属于常见停用词且与问答任务无关的专业字词. 根据构建出的停用词词典对分词结果中的无意义信息进行过滤, 最终得到了问题的有效分词序列$\{W_{T1},W_{T2},\cdots,W_{Tn}\}$.

在构建最小斯坦纳树时, 首先将问题关键词映射到领域知识图谱中形成基准节点$\{BN_1,BN_2,\cdots,BN_l\}$, 即将$\{W_{T1},W_{T2},\cdots,W_{Tn}\}$中的每一个分词, 与学科建设知识图谱中的所有节点名称进行遍历比较, 通过一对一映射得到与其唯一对应的节点. 然后使用最小斯坦纳树算法[21] 构建出能够连接所有基准节点的最小斯坦纳树. 为了保证问答结果的准确度, 本文又利用知识图谱中的丰富信息对该最小斯坦纳树进行了扩充, 即找出树中所有路径节点的一阶直接相连邻居节点, 并添加至该最小斯坦纳树中得到问答相关的语义子图, 以此增大找出问题答案的可能性.

去除问答相关的语义子图中所有的基准节点后, 余下节点共同构成了候选答案节点集合$\{CN_1,CN_2,\cdots,CN_k\}$, 为后续候选节点筛选与答案生成做准备.

2.3 候选节点筛选与答案生成

在筛选候选答案时, 主要通过对问题的关键词进行分类从而得到答案类型. 其中问题的关键词是指去除问题分词序列$\{W_{T1},W_{T2},\cdots,W_{Tn}\}$中的所有基准节点$\{BN_1,BN_2,\cdots,BN_l\}$对应的分词后, 所形成的关键词集合$\{W_{K1},W_{K2},\cdots,W_{Kp}\}$.

将关键词集合中的所有关键词依次输入使用通用语料集训练好的 BERT-Base-Chinese 模型后, 得到所有关键词的词向量集合$\{V_{K1},V_{K2},\cdots,V_{Kp}\}$, 将所有关键词的词向量进行平均加权求和后, 计算与自定义的答案类型关键词之间的相似度, 然后取相似度最高的类型作为问题的答案类型. 本文自定义的答案类型对应关键词信息如表 3 所示.

表 3 答案类型及其对应的关键词信息

答案类型	对应关键词
时间	时间、时候、何时
地点	地点、地方、何地、何处、哪里
人物	人、人物、何人、谁、哪位
院校机构	学校、学院、院校、机构、部门
会议	会议、大会、参加
期刊	期刊、刊物、杂志
论文	论文、文章
专利	专利、发明
项目	项目、决议

最后, 通过预测出的答案类型对候选答案节点集合$\{CN_1,CN_2,\cdots,CN_k\}$进行筛选, 选取对应类型的候选答案节点所代表的实体, 得到最终问题的预测答案.

3 实验与结果分析

本文主要使用基于文本事件抽取的方法来构建学科建设知识图谱, 并结合斯坦纳树算法对学科建设问答应用进行了研究. 为了验证本文提出方法在学科建设领域的有效性, 分别对这两部分工作进行实验和结果分析, 并展示了最终基于知识图谱搭建的学科建设可视化系统.

3.1 知识抽取

3.1.1 实验数据集与评价指标

本文所使用的数据集为实验室构建, 通过爬取国内 8 所高校相关的学科建设文本, 并结合人工标注, 完成学科建设数据集的构建. 本文构建的领域数据集共有 5 732 条样本, 采用 BIO 标记方式, 并按照 8:1:1 的比例来划分训练集、验证集和测试集. 其中需要抽取 4 类触发词与 9 类事件元素, 各类标注对象的具体标注说明如表 4 所示.

在文本事件抽取环节, 实验选取精准率 (P)、召回率 (R) 以及 $F1$ 值作为模型性能的评价标准, 其定义如式 (6)–式 (8):

$$P=\frac{N_{\text{correct}}}{N_{\text{correct}}+N_{\text{incorrect}}}\times 100\% \tag{6}$$

$$R=\frac{N_{\text{correct}}}{N_{\text{correct}}+N_{\text{unlabeled}}}\times 100\% \tag{7}$$

$$F1 = \frac{2PR}{P+R} \times 100\% \tag{8}$$

其中, N_{correct}为标注正确的结果数量, $N_{\text{incorrect}}$为标注错误的结果数量, $N_{\text{unlabeled}}$为未标注出的结果数量.

表 4 学科建设事件数据集标注说明

类别	标注对象	标注名称	标注说明
触发词	科研类事件	RESEARCH	发表、发明、获批、承接、发布、研究
	会议类事件	MEETING	举行、举办、召开、参加、出席
	建设类事件	CONSTRUCT	建设、共建、合建
	视察调研类事件	INSPECT	视察、调研
事件元素	时间	Time	事件中涉及的时间
	地点	Site	事件中涉及的地点
	人物	Person	参与事件的人物
	院校机构	College	参与事件的院校机构
	会议	Meeting	召开的会议名称
	期刊	Journal	事件中涉及的期刊名称
	论文	Paper	发表的论文名称
	专利	Patent	发明的专利名称
	项目	Project	事件相关的项目名称

3.1.2 实验参数配置与结果分析

实验采用 Windows 10 操作系统, 使用 RTX2080Ti 的 GPU 进行训练, Python 版本为 3.6, PyTorch 版本为 1.3.1, 实验参数配置如表 5 所示.

表 5 实验参数配置

参数项	参数设置
批处理尺寸	16
嵌入向量维数	256
隐藏层节点数	768
初始学习率	0.001
Dropout	0.1
优化器	Adam
Epoch数	50

模型抽取结果随 Epoch 的变化情况如图 7 所示, 可以看出, 在初始状态下, 随着 Epoch 的增加, 模型的 $F1$ 值开始出现大幅度上升, 当 Epoch 达到 25 次左右时, 模型的 $F1$ 值开始趋于稳定.

为了验证 BERT-BiLSTM-CRF 模型对学科建设数据集的适用性, 实验对触发词和事件元素的抽取效果进行了对比, 各类数据的抽取结果如图 8 所示.

对于部分事件元素类别, 如会议和项目, 其抽取结果的 $F1$ 值相比于该模型的整体效果较低, 其原因是学科建设中会议与项目实体出现的次数较少, 且不同会议或项目的表述也存在一定差异. 从实验结果可以看出, 模型对于事件类型的分类效果较好, 且能够自动识别出未标记的实体名称, 表明其对于文本的上下文理解有着明显的优势.

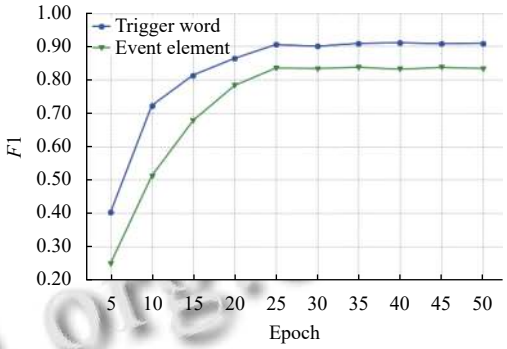

图 7 不同 Epoch 下事件抽取 $F1$ 值分析

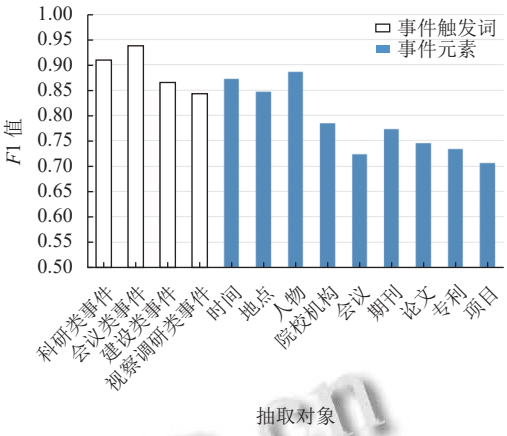

图 8 学科建设事件抽取结果

3.2 智能问答

本小节以问题“中国石油大学 (华东) 的肖军弼老师与哪位老师在计算机教育上共同发表过论文?”为例, 对本文所提出的基于最小斯坦纳树的智能问答方法进行详细说明.

首先是问题分词步骤, 先使用分词词典对问题进行文本分词, 得到初始分词序列{“中国石油大学 (华东)”, “的”, “肖军弼”, “老师”, “与”, “哪位”, “老师”, “在”, “计算机教育”, “上”, “共同”, “发表”, “过”, “论文”, “?”}; 再根据停用词词典, 去除分词结果中的无意义停用词, 得到问题的最终分词序列{“中国石油大学 (华东)”, “肖军弼”, “老师”, “哪位”, “老师”, “计算机教育”, “共同”, “发表”, “论文”}.

然后构建问答语义子图, 将问题最终分词序列中的每个分词, 逐一与知识图谱中的实体节点进行遍历

比较, 匹配得到“中国石油大学 (华东)”“肖军弼”和“计算机教育” 3 个基准节点. 通过最小斯坦纳树算法构建能够连接所有基准节点的最小斯坦纳树, 产生了新增的路径节点“Event_673”“Event_229”和“Event_801”. 在知识图谱中寻找以上 3 个路径节点的直接相连邻居节点对该最小斯坦纳树进行扩充, 能够得到问答相关的语义子图. 去除该问答相关的语义子图中所有的基准节点后, 余下节点共同构成了候选答案节点集合{“Event_673”, “Event_229”, “Event_801”, “张千”, “刘素芹”, “曹绍华”, “2017 年 8 月 10 日”, “2009 年 9 月 25 日”, “2009 年 5 月 25 日”, “网络与课堂生态交融的‘两化一体’课程考核体系”, “‘计算机网络’课程教学方法体系的研究与实践”, “‘计算机网络’课程双语教学模式的探讨”}.

之后再进行答案的筛选与生成. 去除问题分词序列中所有基准节点对应的分词, 形成关键词集合{“老师”, “哪位”, “老师”, “共同”, “发表”, “论文”}. 将所有关键词依次输入 BERT-Base-Chinese 模型中, 得到其对应的词向量, 对所有关键词的词向量进行平均加权求和后, 通过词向量的余弦值计算其与自定义的答案类型关键词之间的相似度, 判断出问题的答案应当为“人物”类型. 最后, 在候选答案节点集合中筛选实体类型为“人物”的节点, 得到问题的最终预测答案为“张千”“刘素芹”和“曹绍华” 3 位老师.

对于此问题的问答语义子图的构建以及预测答案生成的结果显示如图 9 所示.

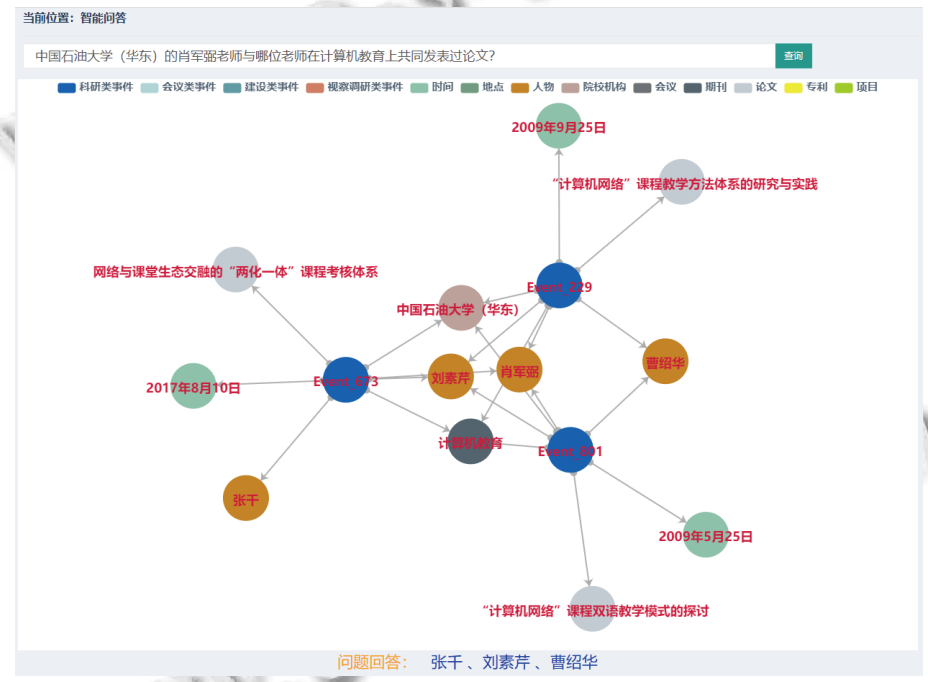

图 9 智能问答结果

4 学科建设可视化系统

4.1 数据存储与管理

作为目前最流行的属性图数据库, Neo4j 实现了专业数据库级别的图数据模型的存储, 具有存储容量大、查询性能高等优点. 因此本文选择 Neo4j 图数据库存储学科建设知识图谱中的数据, 并使用 Cypher 图查询语言实现底层数据库的高效操作[22]. 例如, 查询“在青岛以外地点所举办的会议中相关的院校机构有哪些”, 可以使用如下语句:

```
match (s:Site)-[r1]-(m:MEETING)-[r2]-(c:College)
where not s.name='青岛' return*
```

最终 Neo4j 数据库中的数据规模统计如表 6 所示.

表 6 学科建设知识图谱数据规模

元素	种类	总数量
实体	13	18574
关系	26	30931
属性	279	54962

4.2 可视化系统构建

本文基于 Python 的 Flask 框架搭建了学科建设可

视化系统. 系统采用 B/S (browser/server) 前后端分离的结构模式实现, 前端使用 ECharts 工具实现数据的图形显示效果[23], 通过文本、力导向图和图片等多种形式对学科建设知识进行可视化显示.

可视化系统的功能主要包括相关信息查询、关系显示、多关键词检索等, 可以从实体、属性、关系等多个维度完成知识的搜索与展示. 图 10 为在可视化系统中“一种准确而高效的领域知识图谱构建方法”的查询结果, 界面中通过图形绘制直观展示了实体相连关系与图片属性, 实现了学科建设知识的整合与展示的目标.

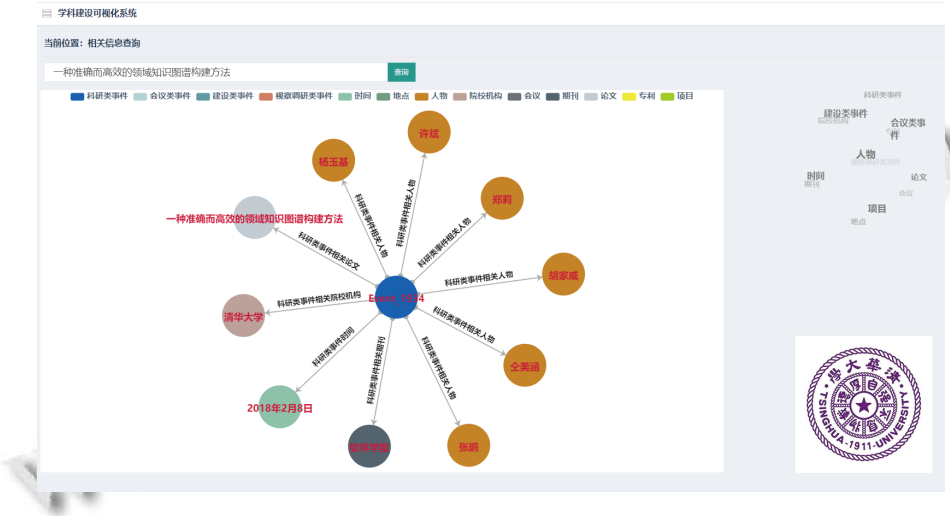

图 10 可视化系统查询结果

4.3 可视化系统评估

为了验证本学科建设可视化系统能否解决实际场景中的问题, 达到整合高校学科建设成果的目的, 在系统构建完成后, 邀请了数十位参与过学科评估工作的专家进行测试. 参与测试人员需从数据完整性、功能效果、运行效率以及人机交互性 4 个维度对系统进行打分评价, 最后针对每个评价维度, 取其所有分数的平均值作为最终的用户满意度. 测试结果显示, 以上 4 个维度的满意度分别达到了 80.65%、89.03%、84.51%、91.61%, 说明此系统的功能较为完善, 基本满足准确查询学科建设信息的要求, 同时系统界面友好, 易于用户上手和操作. 但是系统的数据覆盖范围不够全面, 且部分功能在处理复杂问题时需要消耗较长的时间, 影响用户操作体验, 后续需要进行改进. 总体而言, 此学科建设可视化系统能够减少用户查询学科建设相关信息所需要的时间, 实现了高校建设成果的整合与展示, 在学科评估工作中有良好的应用前景.

5 结论与展望

本文对学科建设知识图谱的构建及基于领域知识图谱的应用进行了研究. 以自顶向下方式构建好领域本体后, 通过 BERT-BiLSTM-CRF 模型对学科建设文本进行事件抽取, 并使用爬虫对抽取到的实体进行相关知识的补充. 然后基于属性图模型实现了知识的存储, 完成学科建设知识图谱的构建. 最后, 结合构建好的知识图谱, 引入最小斯坦纳树算法实现学科建设问题的智能问答应用.

目前对于学科建设领域的知识图谱研究较少, 本文也是为了将大量离散的学科建设信息进行整合, 同时将知识以可视化的形式直观展示, 因此对图谱的构建及问答应用进行了初步的探索. 但目前本文构建的学科建设知识图谱中包含的概念类别较为有限, 对学科建设中复杂的知识关系没有很好的概括. 未来我们将继续深入对此问题的研究, 改善知识图谱的结构体系, 设计更有效的应用方法, 争取为以后的学科建设与评估等工作做出更多的贡献.

参考文献

1 何兆华, 童广运. 学科建设与高等学校的发展. 陕西教育学院学报, 2007, 23(2): 1–4. [doi: 10.3969/j.issn.1008-598X.2007.02.001]
2 刘献君. 论高校学科建设. 高等教育研究, 2000, (5): 16–20.
3 杨玉基, 许斌, 胡家威, 等. 一种准确而高效的领域知识图谱构建方法. 软件学报, 2018, 29(10): 2931–2947. [doi: 10.13328/j.cnki.jos.005552]

4 Tang MW, Su C, Chen HH, *et al*. SALKG: A semantic annotation system for building a high-quality legal knowledge graph. Proceedings of 2020 IEEE International Conference on Big Data (Big Data). Atlanta: IEEE, 2020. 2153–2159. [doi: 10.1109/BigData50022.2020.9378107]

5 Schneider JM, Rehm G. Curation technologies for the construction and utilisation of legal knowledge graphs. Proceedings of the LREC 2018 Workshop on Language Resources and Technologies for the Legal Knowledge Graph. Miyazaki, 2018. 23–29.

6 Sun HX, Xiao J, Zhu W, *et al*. Medical knowledge graph to enhance fraud, waste, and abuse detection on claim data: Model development and performance evaluation. JMIR Medical Informatics, 2020, 8(7): e17653. [doi: 10.2196/17653]

7 Goodwin T, Harabagiu SM. Automatic generation of a qualified medical knowledge graph and its usage for retrieving patient cohorts from electronic medical records. Proceedings of the 2013 IEEE 7th International Conference on Semantic Computing. Irvine: IEEE, 2013. 363–370. [doi: 10.1109/ICSC.2013.68]

8 任飞亮, 沈继坤, 孙宾宾, 等. 从文本中构建领域本体技术综述. 计算机学报, 2019, 42(3): 654–676. [doi: 10.11897/SP.J.1016.2019.00654]

9 Musen MA, Team P. The Protégé project: A look back and a look forward. AI Matters, 2015, 1(4): 4–12. [doi: 10.1145/2757001.2757003]

10 Xiang W, Wang B. A survey of event extraction from text. IEEE Access, 2019, 7: 173111–173137. [doi: 10.1109/ACCESS.2019.2956831]

11 张秋颖, 傅洛伊, 王新兵. 基于 BERT-BiLSTM-CRF 的学者主页信息抽取. 计算机应用研究, 2020, 37(S1): 47–49.

12 Devlin J, Chang MW, Lee K, *et al*. BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: ACL, 2018. 4171–4186.

13 Graves A, Mohamed AR, Hinton G. Speech recognition with deep recurrent neural networks. Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver: IEEE, 2013. 6645–6649. [doi: 10.1109/ICASSP.2013.6638947]

14 Lafferty JD, McCallum A, Pereira FCN. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Proceedings of the 18th International Conference on Machine Learning. Williams College: Morgan Kaufmann Publishers Inc., 2001. 282–289.

15 Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735–1780. [doi: 10.1162/neco.1997.9.8.1735]

16 于娟, 刘强. 主题网络爬虫研究综述. 计算机工程与科学, 2015, 37(2): 231–237. [doi: 10.3969/j.issn.1007-130X.2015.02.007]

17 王鑫, 邹磊, 王朝坤, 等. 知识图谱数据管理研究综述. 软件学报, 2019, 30(7): 2139–2174. [doi: 10.13328/j.cnki.jos.005841]

18 Pundge AM, Khillare SA, Mahender CN. Question answering system, approaches and techniques: A review. International Journal of Computer Applications, 2016, 141(3): 34–39. [doi: 10.5120/ijca2016909587]

19 Andrenucci A, Sneiders E. Automated question answering: Review of the main approaches. Proceedings of the 3rd International Conference on Information Technology and Applications (ICITA '05). Sydney: IEEE, 2005. 514–519. [doi: 10.1109/ICITA.2005.78]

20 Smith WD, Shor PW. Steiner tree problems. Algorithmica, 1992, 7(1): 329–332.

21 Lu XL, Pramanik S, Saha Roy R, *et al*. Answering complex questions by joining multi-document evidence with quasi knowledge graphs. Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. Paris: ACM, 2019. 105–114. [doi: 10.1145/3331184.3331252]

22 王鑫, 傅强, 王林, 等. 知识图谱可视化查询技术综述. 计算机工程, 2020, 46(6): 1–11. [doi: 10.19678/j.issn.1000-3428.0057669]

23 王子毅, 张春海. 基于 ECharts 的数据可视化分析组件设计实现. 微型机与应用, 2016, 35(14): 46–48, 51. [doi: 10.19358/j.issn.1674-7720.2016.14.015]

(校对责编: 牛欣悦)