|  |  |  |
| --- | --- | --- |
| SRT计划项目申请书 | | |
|  | 项目名称: | 先秦人物知识图谱的构建 |
|  | 申请者: | 季劼旻 |
|  | 学 院: | 信息科学技术学院 |
|  | 专 业: | 信息管理与信息系统 |
|  | 指导教师: | 何琳 职称: 教授 |
| 2018年 3 月 15 日 | | |
| 南京农业大学教务处制 | | |

**一、简表**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 申请人姓名 | | 季劼旻 | 学号 | 31316122 | 班级 | 信息162 | |
| 年级 | | 2016级 | 电话 | 17761705096 | E-mail | 740846055@qq.com | |
| 项目名称 | | 先秦人物知识图谱的构建 | | | | | |
| 项目来源 | | 1. 自立项目√ B、教师科研课题的子项目 C、其它 | | | | | |
| 项目类型 | | A、实验研究√ B、调查研究 C、软件制作 | | | | | |
| 经费来源 | | A、学校资助√ B、导师课题资助 C、企业资助 | | | | | |
| 经费额度 | | 1000元 | | | | | |
| 指导教师姓名 | | 何琳 | | 指导教师职称 | | | 教授 |
| 合作者  姓名、学院、班级 | | 信息科学技术学院 信息162 陈嘉诚 宗睿 宋子安 | | | | | |
| 申请时间 | | 2018年3 月15 日 | | 完成时间 | 2019年 5 月 日 | | |
| 项目研究  内容  摘  要 | **摘要：**知识图谱作为人工智能领域的热点方兴未艾，然而将其应用于历史研究这一领域还存在着很多空白。本课题主要目的是构建先秦历史人物的知识图谱，分为知识模型构建，数据获取，知识抽取与融合，知识验证四个步骤；涉及自然语言处理，数据库原理，知识图谱构建等领域，使用户的搜索结果层次化，结构化。旨在梳理先秦人物之间的关系，提高检索系统的效率，广度与深度。  **关键词：**知识图谱；先秦人物；自然语言处理 | | | | | | |

二、立论依据（包括项目的研究意义、现状分析，并附主要参考文献及出处）

|  |
| --- |
| 1、研究意义  自党的十九大召开以来，弘扬传统文化成为各领域的热点。中华文化源远流长，先秦时期诸子百家创造了灿烂的文化艺术，具有鲜明的特色，对先秦人物的研究和整理对弘扬和继承中华优秀传统文化有着非常积极的作用。由于时间久远，先秦人物文献资料复杂且分散；绝大多数人对这一时期的历史所知不多；而知识图谱能构建一个与搜索结果相关的完整的知识体系，具备知识推理的逻辑能力，提高了搜索的深度与广度。通过知识图谱可以将其系统的整合起来，为检索和相关研究提供便利。  2.现状分析：  （1）知识图谱的概念与意义  知识图谱最早起源于Google Knowledge Graph，是一种结构化的语义知识库，用于以符  号形式描述现实世界中的概念及其相互关系。其基本组成单位是“实体-关系-实体”三元组，以及实体及其相关属性值对，实体间通过关系相互联结，构成网状的知识结构[1]而随着互联网信息总量的爆炸性增长，传统搜索引擎所采用的字符串匹配方式已经很难满足人们快速准确获取知识，全面掌控信息资源的需求。知识图谱为解决之一问题提供了思路。它的研究价值在于在当前web基础上构建了一层覆盖网络（overlay network），将互联网中累积的信息组织起来，成为可以被利用的知识，使用户的搜索结果具有层次化，结构化等重要特征。[2]目前，知识图谱已被广泛应用于智能搜索、智能问答、个性化推荐等领域，在更深层次上揭示人类认知的整体性与关联性，为万维网上的知识互联奠定扎实的基础[3]。  （2）知识图谱的总体应用  知识图谱研究已经取得了很多成果，形成了一些开放的知识图谱[4]，应用范围广泛，包括：科学政策与管理、战略管理、科学经济学、科学合作; 信息检索; 信息科学; 科学知识图谱可视化; 科学行动理论(科学知识社会学)和国家创新系统; 网络计量学、知识管理、科学知识图谱和信息计量学理论。总体来说，国内知识图谱应用研究主要基于学术数据库，而利用专利数据，对技术领域的研究热点与前沿进行可视化图谱分析的研究比较少。而我们本次项目就是在知识图谱的基础上，实现对先秦时期人物关系简单梳理、整合，便于用户查找和使用[5]。  （3）知识图谱在历史研究方面的应用  知识图谱自提出以来一直在发展，但目前为止在历史方面应用较少。只有哈佛大学的CBDB有过涉足，该数据库共收录约370,000人的传记资料，这些人物主要出自七世纪至十九世纪[6]。但对于先秦时期的历史，仍是知识图谱从未涉足的处女地。本项目知识图谱的构建是在该空白领域的一次尝试。  参考文献及出处：  [1][2]刘峤,李杨,段宏,刘瑶,秦志光.知识图谱构建技术综述[J].计算机研究与发展,2016,53(03):582-600.  [3]徐增林,盛泳潘,贺丽荣,王雅芳.知识图谱技术综述[J].电子科技大学学报,2016,45(04):589-606.  [4]漆桂林,高桓,吴天星.知识图谱研究进展[J].情报工程,2017,3(01):4-25.  [5]《情报工程》2017 年第 1 期，知识图谱专题稿件。  [6]Harvard University, Academia Sinica, and Peking University, China Biographical Database (January 1, 2018) |

三、研究方案

|  |
| --- |
| **1. 项目研究目标、内容和拟解决的关键问题**  一.研究目标  构建先秦历史人物事件的文化知识图谱，梳理先秦历史人物之间的关系；提高检索的效率  与质量，为用户搜索和科研提供便利；旨在对我国早期历史追根溯源，弘扬我国优秀的传统历史文化。  二．研究内容  （1）建立合适的数据模型：  高质量标准化的数据模型（schema）能有效降低不同来源数据对接融合的成本，  是项目成功的一半。我们拟从两方面着手：一是查阅已有的先秦人物分类文献，采用其中出现频率较高的，典型的关系；二是参考同研究领域的中国历代人物传记数据库（CDBD）知识模型，综合两者来建立高质量的schema。  （2）获取先秦著名人物列表  从指定范围的先秦文献中，运用实体识别技术获得出现频率较高的人物名。考虑到实体识别的准确性有待提升，难度较高，也可以从现有的先秦人物文献、知识库获取先秦著名人物列表。  （3）数据爬取  为了确保数据质量，优先考虑权威的、稳定的、数据格式规整且前后一致、数据完整的网页；为了降低爬取成本，优先考虑免登录，免验证码，无访问限制的页面。爬下来的数据保存好爬取时间、爬取来源、或网页地址以便后续处理。  （4）先秦人物关系处理  根据所构建的知识图谱进行知识计算与应用，按照时代，地域，国家，思想流派的分类梳  理先秦人物之间的关系，对知识图谱进入深入解读。  三．拟解决的关键问题  1.如何从源数据中抽取信息：   鉴于本项目将百科词条作为主要的信息源，因此本部分讨论如何从半结构化的web页面中提取三元组。经过查阅相关资料，可以使用正则表达式匹配从百科页面信息盒中提取所需数据。  2.如何进行知识融合：  由于知识主要来源于结构不同的几种百科知识库，故要建立它们之间的关联。此部分主要  分为实体对齐，属性对齐两方面。  （1）实体对齐：  抽取实体上下文信息，对其进行中文分词后，利用主题模型对上下文建模并通过吉布斯采  样法求解模型参数，计算出主题单词概率矩阵，提取特征词集合及对应特征矩阵; 然后，利用最长公共子序列( LCS) 算法判定实体属性相似度，当相似度位于下界与上界之间时，进一步结合百科类实体上下文主题特征进行判定。  （2）属性对齐：  首先利用统计理论确定属性的数据类型，并给出类型感知的属性相似度计算方法。在此基  础上，将属性关系识别建模为多分类问题，抽取有效特征对不同关系进行描述并用于随机森林  模型的构建，识别出属性之间的关系。 |

|  |
| --- |
| **2. 拟采取的研究方法、技术路线、实验方案及可行性分析**  一．研究方法与实验方案  （1）构建合适的数据模型  （2）编写Python脚本爬取百度百科，互动百科，维基百科上先秦时代人物的词条。  （3）使用包装器对获取的半结构化数据进行标注抽取得到大量的三元组。  （4）对获得的三元组进行实体对齐属性对其冲突消解规范化的操作进行融合。并进行一致性验证  （5）将融合而成的知识图谱存入关系型数据库中。通过知识计算等手段梳理先秦人物之间的关系。  拟先采取小规模数据实验的方法，验证预先准备的技术流程并发展出一套成熟的构  建机制，再大规模构建知识图谱。  二．技术路线  （1）技术储备：  查阅相关文献及技术资料，了解知识图谱的具体构建流程与技术细节。  （2）数据模式设计：  了解先秦时期的背景知识如官名，爵位等，并参考中国历代人物传记数据库（CDBD）的设  计，构建知识图谱的数据模式，并依此选择高质量的数据源。  （3）数据获取及处理：  爬取各类百科知识中的先秦人物词条，并定义包装器从这些半结构化数据获取知识，获得大量的RDF三元组，多元组事件及时序信息。  （4）知识融合：  对从不同百科知识库获取的异构知识进行实体对齐，属性对齐等操作，规范化数据，消  解冲突，融合而成的知识库存入关系型数据库中。  （5）知识验证：  知识图谱的构建是一个动态的过程。通过一致性验证，判断新引入的知识与原有的知识是  否存在矛盾冲突，进行相关实体与关系的更新，确保知识图谱的一致性与准确性。  （6）知识计算与应用：  这一部分主要是根据图谱提供的信息获得更多的隐含知识，像是通过本体或者规则推理技  术可以获取数据中存在的隐含知识；通过链接预测预测实体间隐含的关系；通过社区计算在知识网络上计算获取知识图谱上存在的社区，提供知识间关联的路径等。  三．可行性分析  1.团队可行性：小组成员参加过用Python进行自然语言处理的技术兴趣小组，积累了一定的自然语言处理技术与经验，导师能够及时答疑解惑。  2.技术可行性：Python语言在爬取网页与自然语言处理方面相较于其他语言有优势。  3.资源可行性：能够通过CNKI等数据库查询所需文献  4.方法可行性：研究思路明确，课题新颖，将知识图谱应用于历史研究的工作比较少。 |

|  |
| --- |
| 3.本项目的创新之处  将知识运用于历史人物研究方面的工作甚少。目前知名的是北京大学中国古代史中心与哈佛大学费正清研究中心、台湾“中央研究院”历史语言研究所联合主持构建的中国历史人物数据库(CDBD)，而其中收录个人物理唐宋及以后居多；本项目是将知识图谱应用于先秦人物关系的一次尝试。  4.项目研究计划及预期进展  （1）2017.06-2017-08：自学Python语言及NLP技术；查阅技术文献，熟悉基本工具。  （2）2017.09-2017.10：小规模数据试验，完成技术验证工作，构建成熟的工作流程  （3）2017.10-2018.02：正式开始知识图谱的构建工作，到二月底基本完成构建。  （4）2018.03-2018.05：整理材料，汇总撰写研究报告。  本小组计划每月进行两到三次的工作进展会议，确保各个成员之间默契配合，相互监督，确保研究计划按时顺利完成。 |
| 5.预期研究成果  研究先秦人物知识图谱的构建方法，并撰写结题报告,争取发表1到2篇论文。 |

四、研究基础

|  |
| --- |
| 1. 与本项目有关的研究工作积累和已有的研究工作成绩  （1）我们小组目前已经阅读了许多关于知识图谱构建过程和发展现状相关方面的资料文献，了解到当前社会对知识图谱的需求越来越广泛。并且有很多学者在这方面进行了相关的研究，取得了很不错的成绩。我们小组已经掌握了一定的相关知识基础对我们的研究有很大帮助。  （2）我们小组的成员已经学习了信息资源管理导论等与本研究相关的基础课程，同时C语言，数据结构，数据库等课程对本项目研究也有技术支撑。此外我们能够初步熟悉python等程序设计语言。  2. 已具备的条件、尚缺少的条件和拟解决的途径（包括利用教学实验室、科研实验室和实习基地等的计划与落实情况）  I.已具备的条件：  （1）研究课题针对的是先秦人物，可以通过网络搜集到大量资料和样本。  （2）网络资源丰富，我们可以随时访问多个数据库，图书馆也有大量的文献期刊可供我们查阅，能够顺利完成相关文献资料的收集，使我们的研究完成的更顺利。同时，老师能在我们困惑时对我们进行指导。  （3）我们小组成员都是信息专业，具备一定的程序编写能力，而且正在学习科技文献检索和信息组织的相关专业知识，我们拥有相对熟练的计算机知识，能通过运用各种软件对数据进行整理和分析；小组成员都团结互助，具备吃苦耐劳的精神品质；且小组内人手一台笔记本，操作方便。Ⅱ.尚缺少的条件：相关专业知识，如统计学和对古代事件的梳理。  Ⅲ.拟解决途径：通过大量查阅文献并向指导老师和专业课老师学习补足专业知识。 |

五、经费预算

|  |  |  |
| --- | --- | --- |
| **支 出 科 目** | **金 额**  **（元）** | **计 算 根 据 及 理 由** |
| 图书资料费 | 400 | 用于购买资料、图书 |
| 调研费 | 200 | 网上文献收集，资料查阅等 |
| 资料文印费 | 200 | 用于文献、资料的打印 |
| 耗材费 | 200 | 用于购买电子器件等耗材 |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
| 总计 | 1000 |  |

六、**审查意见**

|  |  |
| --- | --- |
| 指导  教师  意见 | 选题具有较高的现实意义，利用知识图谱的方法研究先秦人物关系构建，研究前期文件调研充分，研究方法合理，技术路线翔实，同意立项  指导教师签名 何琳 2018年 3 月 22日 |
| 学院  意见 | 负责人签名、公章： 年 月 日 |
| 学校  意见 | 负责人签名、公章： 年 月 日 |
| 备注 |  |

七、申请者承诺

我保证上述填报内容的真实性。如果获得资助，我与本项目组成员将严格遵守学校的有关规定，在不影响课程学习的同时，保证项目研究工作的时间，并按计划认真开展研究工作，在项目研究过程中或结束时，接受学校对本项目的中期检查和结题验收，并按时提交工作总结和结题报告。

申请者（签名）：

年 月 日

推 荐 书

|  |
| --- |
| 推荐人签名： 年 月 日 |

注：此表由具有副高以上职称的教师（非指导教师）填写