合肥工業大學

大学生创新训练计划项目开题报告

项 目 级 别： □国家级 ☑ 省级 □校级

项 目 编 号： S201910359345

项 目 名 称： 基于 Spark 平台的人工智能知识的知识图谱构建

项目负责人： 文华

联 系 电 话： 18856302551

项目组成员： 尧铖/2017217987,刘宏鑫/2017217989,周余/2017218005

起 止 年 月： 2019 年 5 月至 2020 年 4 月

指 导 教 师： 罗月童

所 在 学 院： 计算机与信息系

创新创业教育中心

2019 年

填表须知

1、请将项目级别对应选项的“□”打“√”。

2、要按顺序逐项填写，内容要实事求是，表达要明确、严谨。空缺项要填“无”。可自行复印或加页，但格式、内容、大小均须与原件一致。要求一律用A4纸正反页打印，于左侧装订成册。

3、《项目开题报告》中栏目“1至8”由学生填写，栏目“9”由教师填写，栏目“10”由学院负责人填写。

4、《项目开题报告》由所在学院审查、签署意见后，一式三份（均为原件），报送创新创业教育中心。

项目名称	基于 Spark 平台的人工智能知识的知识图谱构建

1、项目来源及研究目的和意义

随着计算机大数据的快速发展,可以借助于互联网平台的各种工具找到有价值内容,但海量数据給筛选、组织与评价带来极大困难。项目拟利用分布式爬虫实现 PathFinder 算法获得学科信息并借助 Spark 平台过滤无效信息,再借助 Spark 优化基于词向量的文本特征选择算法完成学科专有名词分类,同时采用 Bayes 统计推断完成关键词与相关信息的联想,完成内容的可视化。项目意义在于通过 Spark 完成人工智能知识的重整,实现了一个学习者尤其是本科生适用的知识图谱工具。

2、国内外研究概况及发展趋势

随着 Web 技术飞跃式发展,互联网先后经历了三个时代,它们分别具有不同的特征:文档互联的“Web1.0”时代,数据互联为特征的“Web2.0”时代以及当下正在发展的知识互联的崭新“Web3.0”时代[1][2]。知识互联为人们的学习与交流提供了极大便利,人类的知识交互达到了历史的新高峰。然而,互联网上的知识来源复杂、良莠不一,零散混乱、体系松散,尤其是在大数据的时代背景下,这給内容的筛选、组织与评价带来了极大挑战。知识图谱是人工智能领域一项重要的技术分支,具有强大的语义处理能力与开放互联能力[3]。值得注意的是,目前国内尚无针对人工智能这一领域的知识图谱工具。人工智能正处于快速发展阶段,了解、学习、掌握有关知识与技术是学生、工程师、科研人员所面临的一大挑战,优秀的知识架构可以帮助学习者达到事半功倍的效果。

本项目拟构建的知识图谱本质上是一个学科术语索引、解释与联想的集成工具,而术语的获取基于既定文本集。原始文本集的数据存在一定程度的混乱,不利于数据抽取等后续操作,因此,需对文本进行分类。文本分类是指按照文本的含义与特征将待处理的文本划分为不同类别的文本进行其类别判定的全过程,而文本分类是在类别判定前就已经确定好了的[4]。在中文文本分类中,特征值的权值计算方法决定了文本特征的提取。因此,需要选取恰当的模型对文本进行特征提取。词向量空间模型(Vector Space Model,VSM)在文本分类中被广泛地应用,在向量模型中文档被分割为一个个特征项,再分别以特征向量表示之。VSM 是目前文本处理最常用的模型[5]。当前较为流行的高效获取词向量的工具是由 Google 的 Tomas Mikolov 团队提出并实现的 Word2vec,该算法具有在较短时间内从大规模语料库中构建高质量的词向量的能力。

传统基于 VSM 的文本分类方法进行文本预处理、特征选择、文本向量化与文本分类器的生成,这带来可观的时间开销与空间消耗,尤其在处理大规模数据文本时,处理时间过长成为无法回避的问题[6]。目前对 VSM 的优化与改进虽然取得了一定的效果,但都无法从根本上跨越效率低下的瓶颈。马力等在《基于词向量的文本分类研究》中提出了一种基于词向量的文本特征选择改进算法,较传统的特征选择算法分类准确率得到改进[7]。张群等在《词向量与 LDA 相融合的短文本分类方法》中设计了一种基于词向量与 LDA 主题模型相融合的短文本分类方法,较单一的基于向量空间模型、基于词向量、基于 LDA 主题模型的分类方法,该方法的准确率、召回率、F1 值均有提升,但仅应

用于最近邻分类器[8]。张静宜等在《基于词向量特征的文本分类模型研究》中针对具体场景,采用 word2vec 中的词向量与 IF-IDF 中的词频重要性相结合构建特征工程,然后选择XGBoost集成学习分类算法构建模型,较传统VSM提高了文本分类效率与分辨准确率[9]。VSM效率低下的原因主要是没有实现并行化处理,这个问题得不到解决,那么当在全局文本集较大时,则无法兼顾效率与准确率。诸如为了分布式系统设计的Cascade算法,其在最后阶段仅能运行在一台机子上,限制了算法效率。为了解决上述问题,一种有益的尝试是借助于 Spark 等大数据平台[10]。Spark 是基于内存计算的大数据并行计算框架,因为它基于内存计算,所以提高了在大数据环境下数据处理的实时性,同时保证了高容错性和高可伸缩性,允许用户将 Spark 部署在大量廉价硬件之上,形成集群。

“可视化”作为专业术语最初是由 1987 年 2 月美国国家自然科学基金会(National Science Foundation,NSF)主办的一个专题研讨会给出,当时称作“科学计算可视化”(Visualization Scientific Computing)[11]。可视化分为:数据可视化(Data Visualization)、信息可视化(Information Visualization)与知识可视化(Knowledge Visualization),三者既有区别,也有联系。其中,知识可视化在知识的传播与创新过程中得到了广泛应用与发展。知识可视化在应用于具体学科知识时,还有结合应用学科的不同需要,分门别类,根据学科特点采用不同的可视化工具[12]。对于一些重要的基础课程和专业课程,借助知识可视化工具建构学科知识的整体脉络与框架显得尤为重要。

知识图谱(Knowledge Map)在2005年由陈悦等率先在中国引入并命名[13],是人工智能领域一项重要的技术分支,它通过结合自然语言处理、机器学习、数据统计分析和大数据处理等技术来构建一个高效可靠的知识仓库,并以图的形式展示各个知识模块节点的关联性,从而帮助使用者更好的获取知识并从知识的联系中得到有价值的[3]信息。

数据抽取或信息抽取是一个将非结构化或半结构化的文本数据转化为机器可以理解的结构化数据的过程[14]，针对其研究旨在于开发更有力的信息获取工具，以应对大数据时代海量数据的挑战[15]。数据抽取有两大关键技术：命名实体识别。命名实体是文本中基本的信息元素,是正确理解文本的基础。实体消歧。命名实体的歧义是指一个实体的指称项可以对应多个实体概念[15]。数据抽取是构建知识图谱的必要前提，抽取效果直接影响后续知识图谱的存储。显然，数据抽取来源的选择也需要仔细的考量与抉择。目前，针对学术类知识图谱，大多数有稳定、成熟的数据库作为数据源。而在工业界知识图谱的应用场景中，往往存在数据杂源异质、结构松散的特点，网络爬虫成为其数据抽取的主要手段。网络爬虫（Web Crawler），是按照一定规则，自动地抓取万维网信息的程序或脚本，为后续的数据操作提供数据集。

综上分析,本项目拟利用分布式爬虫并结合 PathFinder 算法或从现成学科数据库获取具体学科的知识内容。再借助于 Spark 框架优秀的并行化处理能力，过滤所获取内容中的无意义数据，并在其上优化基于词向量的文本特征选择算法等相关文本分类算法，完成内容分类。其次，采用垂直搜索引擎工具完成关键词与相关信息的联想。最终，通过数据可视化技术将取得的不同模块的内容，即知识图谱进行可视化展示。

值得一提的是，当前知识图谱在某些领域的应用取得了一定的成果，有的甚至直接推动了该领域的发展与变革[16]。但是，知识图谱在一些领域的应用则乏善可陈，有的领域甚至还处在初步探索中。具体有以下几个方面的内容：

（1）通用知识图谱的应用

特点是不面向特定领域,可将其类比为“结构化的百科知识”。这类知识图谱包含了大量常识性知识,强调知识的广度[3]。具有代表性的大规模通用知识图谱有 YAGO、DBpedia，等等。中文通用知识图谱有 Zhishi.me、SSCO，等等[3]。对于大学本科生群

体，正处于学习打稳专业知识根基的阶段，通用知识图谱可以帮助其拓展视野、增长见识，加强对与本专业交叉的领域认知与探索。但通用知识图谱的工具多样，杂源异质，系统性差，不利于本科生对知识的提取与组织。

（2）垂直知识图谱的应用

又称为行业知识图谱，特点是则面向特定领域,基于行业数据构建,强调知识的深度。垂直知识图谱可以看作基于语义技术的行业知识库,其潜在使用者是行业的专业人员[3]。CiteSpaceⅡ是垂直知识图谱工具的杰出代表，其凭借多元、分时、动态的引文分析可视化技术所绘制的 CiteSpace 知识图谱,能够将一个知识领域来龙去脉的演进历程集中展现在一幅引文网络图谱上,并把图谱上作为知识基础的引文节点文献和共引聚类所表征的研究前沿自动标识出来[17]。而在本科学习过程中有许多课程可以在课下更深入地了解，垂直知识图谱取得的效果可预见是彻底与显著的：将学科知识完全以图形化的方式形象展示，便于学习者系统学习，对学科的来龙去脉、历史沿革与发展走势一览无余。令人遗憾的是，目前没有合适的针对大学本科生的此类工具。

综上可知，当前的知识图谱工具普遍存在下列问题：

（1）通用知识图谱工具涉面较广，但知识冗余混乱、组织零散、系统性差，不利于用户的专业学习；

（2）垂直知识图谱工具种类少，成熟的应用仅限于某些领域，在一些具有较大应用需求的领域未获重视，前景广阔。

基于以上观点，本项目拟采用以下方法解决上述问题：

（1）网络爬虫（Web Crawler）可以按照既定规则爬取万维网（World Wide Web，WWW）上的内容，但是爬取的内容冗杂，加大了后续操作如数据过滤的难度。在爬取数据时，借助 PathFinder 算法筛选高于阈值的目标内容。PathFinder 算法天然地可以过滤掉节点间的非重要关系[18]。相关度高的爬取内容有助于达到知识图谱构建的预期目的：知识面向给定领域、高度组织、系统化。

（2）大量数据有效处理的问题，是限制垂直知识图谱工具发展的原因之一。对于爬取的数据，借助 Spark 的一些常用转换操作可以初步过滤无价值的内容。再借助 Spark 平台对文本分类的各个过程进行并行化，将数据导入弹性分布式数据集（Resilient Distributed Datasets，RDD），实现基于内存的数据计算，这种并行结

构允许在内存中保存工作集并重复利用。此外，还可以管理分区来优化数据放置，并使用大量透明基元操作数据。所有这些功能都允许用户轻松设计新的数据处理管线[19]。

参考文献

[1]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(4):589-606.

[2]SHETHA,THIRUNARAYANK.SemanticsempoweredWeb3.0:managingenterprise,social,sensor,andcloud-baseddataandserviceforadvancedapplications[M].SanRafael,CA:MorganandClaypool,2013.

[3]万倩,欧阳峰,赵明.知识图谱在广电网络运营大数据分析中的应用[J].广播与电视技术,2018,Vol.45(12).

[4]赵政.文本向量化方法对文本分类效果影响的改进研究[D].首都经济贸易大学,2018.

[5]杨开平.基于语义相似度的中文文本聚类算法研究[D].电子科技大学,2018.

[6]光顺利.基于 Spark 的文本分类的研究[D].长春工业大学,2013.

[7]马力,李沙沙.基于词向量的文本分类研究[J].计算机与数字工程,2019,Vol.47(2).

[8]张群,王红军,王伦文.词向量与 LDA 相融合的短文本分类方法[J].2016,Vol.277(12).

[9]张敬谊,张亚红,李静.基于词向量特征的文本分类模型研究[J].行业应用,2017(5).

[10]赛金辰.基于 Spark 的 SVM 算法优化及其应用[D].北京邮电大学,2017.

[11]赵国庆,黄荣怀,陆质坚.知识可视化的理论与方法[J].开放教育研究,2005,Vol.11(1).

[12]赵慧臣.知识可视化的视觉表征研究综述[J].远程教育杂志,2010.

[13]陈悦,刘则渊.悄然兴起的科学知识图谱[J].科学学研究,2005,Vol.23(2).

[14]BankoM,CafarellaMJ,SoderlandS,etal.OpenInformationExtractionfromtheWeb[C]//Proceedingsofthe20thInternationalJointConferenceonArtificialIntelligence.Hyderabad,India:[s.n.],2007:2670-2676.

[15]李畅.信息抽取和实体消歧[J].福建电脑,2014.

[16]孙文沣,邱艳娟,高岩.基于 CiteSpace 的大数据文献可视化分析[J].信息通信技术与政策,2018.

[17]陈悦,陈超美,刘则渊,胡志刚,王贤文.CiteSpace 知识图谱的方法论功能[J].科学学研究,2015,Vol.33(2).

[18]汤天波.Pathfinder 算法优化研究[J].计算机应用与软件,2015,Vol.32(11).

[19]臧艳辉,赵雪章,席运江.Spark 框架下利用分布式 NBC 的大数据文本分类方法[J].计算机应用研究,2018,Vol.36(12).

3、项目主要研究内容

项目流程大概分为以下几个步骤，首先借助分布式爬虫实现 PathFinder 算法获取学科知识数据，经 Spark 平台对数据进行初步过滤，再利用 Word 分词组件完成分词，之后采用 x^2 统计完成文本特征选择并用贝叶斯（Bayes）方法进行文本分类，以上工作结束后，借助基于贝叶斯统计推断的研究方法实现关键词联想。以上工作都运行于 Spark 平台，结合分布式框架提高工作效率。上述步骤必须结合图 1 所示知识图谱的体系架构综合考量（如下所述）。最终，实现内容的可视化，完成知识图谱的创建工作。

（1）基于网络爬虫的数据获取

本项目拟构建人工智能知识的知识图谱，但目前并不存在有关内容的开源数据库

或信息源，因此，利用分布式爬虫获取内容是唯一有效的方法。然而，传统的分布式爬虫虽然可以有选择地访问网页与相关链接并获取所需信息，但获取内容仍含有一定的无价值数据。在大数据环境下，分布式架构的分布式爬虫比单机多核的串行爬虫具有更高的效率与更新速度。爬取相关度更高的内容也是一个值得考虑的问题，为了解决这个问题,我们拟借助分布式爬虫实现 PathFinder 算法,根据相关度阈值获取内容。

另一个值得关注的问题是数据爬取源，恰当的数据源不仅可以更快速地得到所需内容，而且获取内容更“干净”、更接近直接在工程上应用。本项目拟实现所构建知识图谱的相关信息的联想，对信息热度、就业热度等进行统计分析，为学生的深入学习乃至就业择业提供参考。因此，数据源对最终结果的准确度、完整度至关重要。譬如：构建“编程语言”的知识图谱时，可选择“TIOBE 编程语言排行榜”作为信息热度的数据源；构建“机器学习”的知识图谱时，可选择“CSDN 博客”、“牛客网”、“Leet Code 中文官网”作为行业形势与就业热度的数据源。

（2）数据处理与信息联想

文本预处理是将文本表示成一组特征项。将每个词作为文本的特征项是目前常用的处理方法，针对本项目的文本特征项主要是专有名词与术语，本项目拟在 Spark 平台下利用 Word 分词，实现分布式工作。Word 分词是用 Java 实现的，实现了多种分词算法，并利用 ngram 模型消除歧义，能有效对数量词、专有名词与人名进行识别。

特征降维方法拟采用特征选择。预处理得到的词作为特征项，一个词就是一个维度，对高维度的文本特征空间进行降维能在一定程度上提高分类的精度。针对本项目的实际需要，我们拟在 Spark 平台优化 x^2 统计（CHI-Square）作为特征选择方法。x^2 统计是用来度量两个变量的相关性的，值越大，表示两个变量的相关关系越紧密。

文本分类是对已经表示成便于机器学习的文本模型，利用分类算法进行学习，构造出文本分类器。在已经完成 Spark 采用 x^2 统计完成文本特征选择的基础上，利用贝叶斯(Bayes)方法进行文本分类。贝叶斯算法较一般的 SVM(Support Vector Machine，SVM）算法具有计算开销小、处理速度快的优点。

本项目所谓信息联想主要是关键词联想。关键词联想广泛应用于搜索引擎中，以 Bing、Google 为代表的搜索引擎在用户键入信息时，通过同义词、近义词或相关算法等列出备用信息供用户选择。在本项目拟构建的人工智能的知识系统中，用户希望通过某个关键词能够了解该领域的行业形势与前沿应用。为达到上述效果，我们拟采用

通过 Spark 平台优化的基于贝叶斯统计推断的研究方法,对关键词进行知识发现和结果推荐。

（3）知识图谱与数据可视化

知识图谱（Knowledge Map）构建的主要目的是获取大量的、使计算机可读的知识。通过阅读本项目拟构建的知识图谱，读者可清晰地了解某个术语在整个人工智能领域中的位置，认知某门课程的知识脉络，某个课本知识在具体行业中的应用范围与热度。在知识图谱中，两个节点的距离越近或之间连线越粗说明关系越密切，离心距（某个节点到对应图谱中心的距离）越短说明对应节点对该图谱核心的贡献越大。

知识图谱构建流程大致如下：

数据获取：借助分布式爬虫实现 PathFinder 算法获取学科知识数据，经 Spark 平台对数据进行初步过滤；

知识抽取：

本项目面临的知识抽取工作主要是术语抽取。

a.实体抽取：基于 VSM 描述文本进行实体抽取，最重要的是选择与分类相关的特征构造实例特征向量。因此，利用 Word 分词组件完成分词，将所得数据各个术语提取出来；

b.关系抽取：将关系抽取看作是一个分类问题，采用 x^2统计完成文本特征选择并用贝叶斯（Bayes）方法对 a.所得分词，在训练语料的基础上构造分类器，进行文本分类；

c.属性抽取：将判别属性视作分类问题，在文本分类过程中完成此项工作。

知识融合：

a.实体对齐：本项目将实体对齐定义为用户输入与知识库中的实例匹配；

b.知识推理：为了保持项目拟构建工具的稳定性，本次项目不对知识库中的内容进行知识推理，而借助基于贝叶斯统计推断的研究方法实现关键词联想；

c.本体构建：本体是用于描述一个领域的术语集合，本项目的本体构建暨术语的提取与分类在知识抽取阶段已完成；

d.质量评估：对给定输入所生成的知识图谱与现有知识体系进行对比、评价。

创建完成：

至此，知识图谱的创建工作基本结束。

图 1 知识图谱构建技术流程

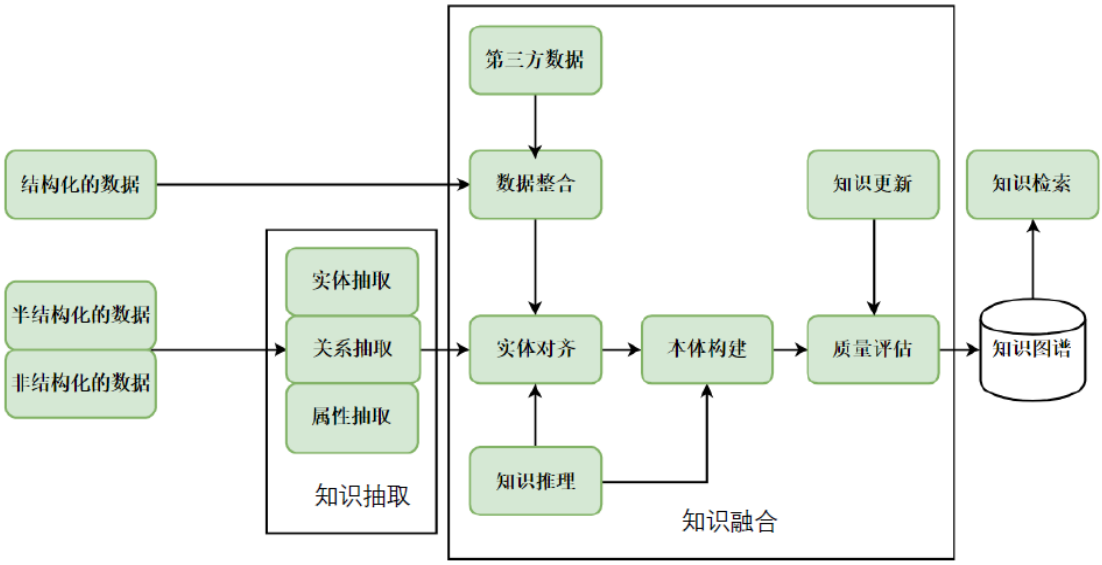
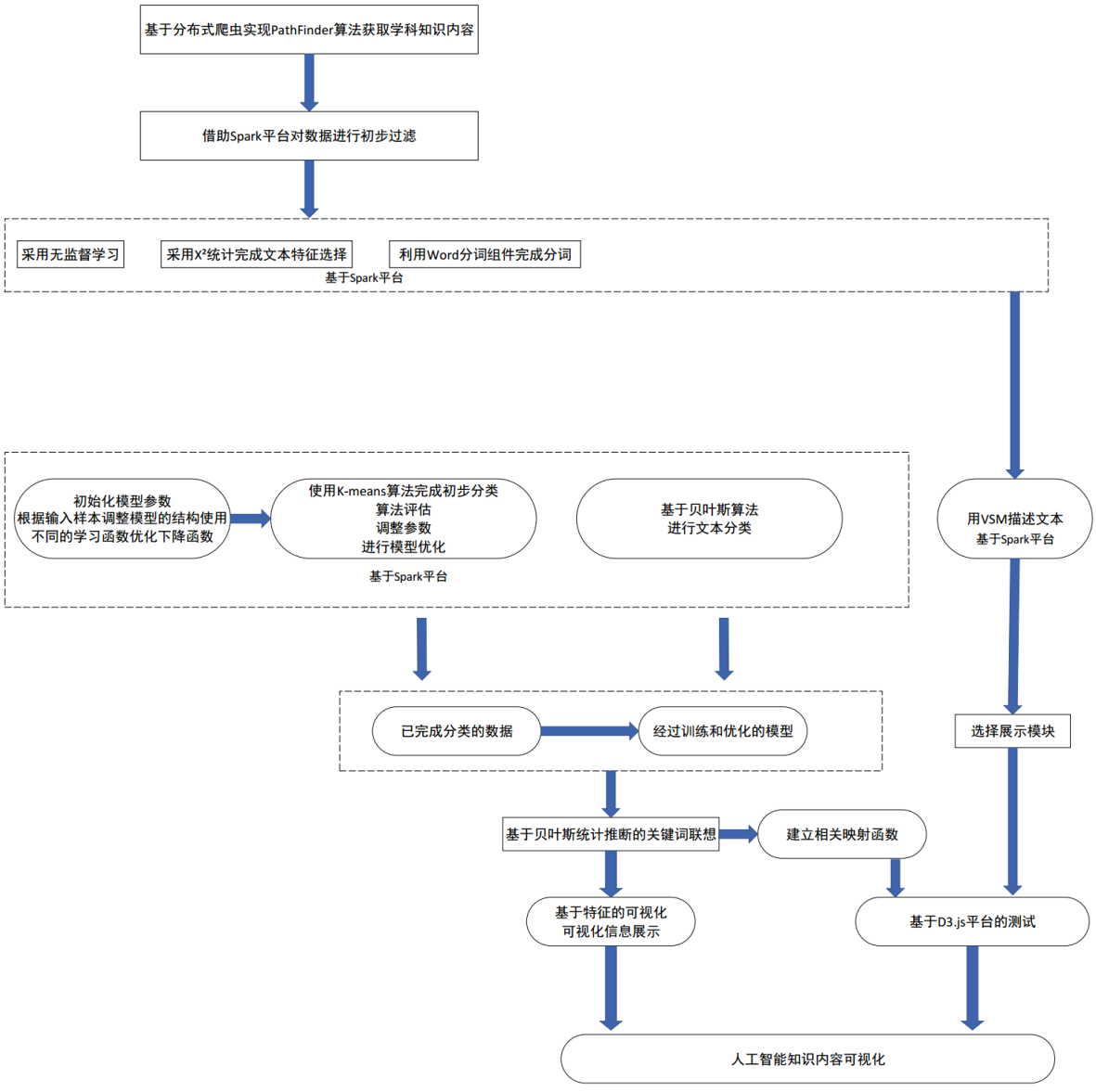


图 2 项目工作流程


基于分布式爬虫实现PathFinder算法获取学科知识内容
借助Spark平台对数据进行初步过滤
采用无监督学习
采用X²统计完成文本特征选择
利用Word分词组件完成分词
基于Spark平台
初始化模型参数
根据输入样本调整模型的结构使用
不同的学习函数优化下降函数
使用K-means算法完成初步分类
算法评估
调整参数
进行模型优化
基于Spark平台
基于贝叶斯算法
进行文本分类
用VSM描述文本
基于Spark平台
已完成分类的数据
经过训练和优化的模型
选择展示模块
基于贝叶斯统计推断的关键词联想
建立相关映射函数
基于特征的可视化
可视化信息展示
基于D3.js平台的测试
人工智能知识内容可视化

4、已完成的前期研究工作及成果

针对选题，团队成员已经做了大量前期工作，譬如查阅了大量选题有关的资料，掌握了项目所需主体平台的搭建与一些算法。

5、拟采用的研究方案和要解决的关键技术问题

(1) **拟解决关键问题**

① 如何准确、高效地进行文本分类是本项目中的一个难题，文本分类直接影响到后续的关键词联想与文本可视化，对最终知识图谱的呈现效果意义重大。文本分类过度，可能会导致词汇残缺、引发歧义；分类不足，则可能引起 VSM 维度较低，知识图谱间的节点之间的关系混乱。而且，文本分类必需快速完成，否则会引起用户体验不良。

② 如何美观、简洁地呈现节点之间的关系以及与之相关内容是本项目的另一个难题。目前，知识图谱构建缺乏开源的工具，很多研究工作都不具备实用性，而且很少有工具发布。这说明本项目无现成的先例可循，一切实现工作需视具体环境自行设计算法。这为本项目加大了实现难度，但也说明了本项目的提出具有较为广阔的应用前景。

(1) **解决途径**

① 算法与相关模型的选择

本项目拟借助 Spark 平台优化文本分类阶段所用的各个算法：利用 k-means 算法对初步过滤的数据进行初步分类，将各个节点按照既定目标分类，同时采用贝叶斯方法完成文本的最终分类，将每个词汇提取为 VSM 的一个维度，用于后续 VSM 描述文本。

② 知识图谱的存储

知识图谱主要有两种存储方式：一种是基于 RDF 的存储，另一种是基于图数据库的存储。它们之间的区别如表 1 所示。RDF 一个重要的设计原则是数据的易发布以及共享，图数据库则把重点放在了高效的图和搜索上。其次，RDF 以三元组的方式来存储数据而且不包含属性信息，但图数据库一般以属性图为基本的表示形式，所以实体和关系可以包含属性。

由表 1 可知，RDF 存储方式主要用于学术界场景。本项目拟实现的是人工智能领域的知识图谱，偏向于学术性质，因此我们选择 RDF 作为本项目知识图谱的存储。

表 1 RDF 与图数据库对比

RDF	图数据库
存储三元组（Triple）	节点和关系可以带有属性
标准的推理引擎	没有标准的物理引擎
W3C 标准	图的遍历效率高
易于发布数据	事物管理
多数为学术界场景	多数为工业界场景

6、研究工作进度安排

表 2 进度安排

2019.4	调研与准备阶段
2019.5-2019.7	收集数据,使用相关算法模型进行数据预处理
2019.8-2019.10	对已处理的数据文本分类，优化模型
2019.11-2020.12	项目平台的初步搭建,测试工具效果
2020.1-2020.2	准备中间检查,对项目的研究过程进行整理
2020.3	结题报告书和相关材料的撰写
2020.4	结题答辩

7、项目预期成果及成果形式

1. 完成本项目的各项预期目标；

2. 证明本项目所提出知识图谱构建方案的基本可行性；

3. 撰写基于本项目的研究论文 1 篇。

7、项目经费预算（请认真、详细填写附表 1）

表 3 经费预算

序号	支出项目	金额（元）	主要用途	备注
1	高性能内存条	2800	搭建、运行大数据平台，实现相关算法	
2	相关资源购买	400	各种书籍纸质资源	
3	纸质材料打印	300	打印材料	
4	高性能显卡	4000	可视化需要高等级	
			显卡支持	
5	高性能硬盘	500	爬取的大量数据需介质存储	
总计			8000	

9、导师意见：

签名：

年月日

10、学院意见：

单位（盖章）：负责人签字：

年　　月　　日

11、学校专家组评审意见： 专家组组长签字： 年　　月　　日
12、学校审批意见： 主管部门（盖章）：负责人签字： 年　　月　　日

附表 1：

大学生创新创业训练计划项目经费预算表

项目名称	基于 Spark 平台的人工智能知识的知识图谱构建	项目负责人	文华	联系方式	18856302551
所在学院	计算机与信息系	指导教师		罗月童	

科目	预算（元）	备注
1、设备费	7300	
2、材料费	0	
3、差旅费	0	
4、印刷费/出版费	0	
5、资料费	400	

6、办公用品费	300	
7、外协费（委托业务费）	0	
8、会议费/培训费	0	
9、其他	0	必须说明内容
合计	8000	
项目负责人签字：　　年　　月　　日		
指导教师签字：　　年　　月　　日		

注：此表将作为报销时的参考依据，请严格按照审批后的总金额预算。